Face Matching Under Time Pressure and Task Demands

Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA, 92697-5100

Robyn L. Vast (robyn.vast@psychology.adelaide.edu.au)
School of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Marcus A. Butavicius (marcus.butavicius@dsto.defence.gov.au)
Intelligence Surveillance and Reconnaissance Division, Defence Science and Technology Organisation
PO Box 1500, Edinburgh SA 5111 AUSTRALIA

Abstract

Understanding how people recognize and match faces is important in many real-world situations, including policing, military, security and retail environments. We investigate the effects of time pressure and additional task demands on face matching performance. In a 2x2 factorial design—varying whether there was high or low time pressure, and whether or not an additional task had to be completed—participants were asked to judge whether each of a series of face image pairs were of the same person. Large individual differences were observed. Recall was higher than precision, and performance worsened under high time pressure with the additional task. Learning effects within conditions were observed, and response times were generally independent of the decisions made. Some implications of these findings for applied environments are discussed.

Introduction

Face matching is the process of verfiying the equivalence of two or more people on the basis of their facial characteristics. It involves visually perceiving a face, and matching this face to a known individual or identification image (Lewis & Edmonds 2003). Understanding how people match faces is an important issue in cognitive science for both theoretical and practical reasons. On the theoretical front, it involves studying core cognitive representational and decision-making processes with complicated natural stimuli. On the practical front, understanding how people match faces has application in areas such as policing, the military, security, eyewitness testimony, social interactions and biometric identification.

Previous research has considered the processes used, and conditions under which, humans can remember, match and recognize both familiar and unfamiliar faces (e.g., Baudouin, Gilibert, Sansone & Tiberghein 2000; Nagayama 1999; Shapiro & Penrod 1986). Through focusing primarily on the effects of disguises, and changes in appearance, race, and gender, past research has highlighted that humans are often not very skilled at recognizing unfamiliar faces (Burton, Miller, Bruce, Hancock & Henderson 2001; Luckman, Allinson, Ellis & Flude 1995; Shapiro & Penrod 1986). Other research suggests that image quality (Bruce et al. 1999; Henderson, Bruce & Burton 2001; Inui & Miyamoto 1984; Liu, Seetzen, Buton & Chaurdhuri 2003), illumination and exposure time (DiNardo & Rainey 1991; Laughery, Alexander & Lane 1971), image colour (Bruce et al. 1999; Yip & Sinha 2002), moving as opposed to static facial images (Henderson et al. 2001; Kemp, Towell & Pike 1997; Knight & Johnston 1997; Lander & Bruce 2003) and distance and perspective (Hager & Ekman 1979; Harrigan & Taing 1997) can also influence face recognition.

In operational environments, face matching tasks are often performed repetitively over long time periods, under strict time limits and processing guidelines, and in conjunction with other related tasks. For example, face matching tasks are often performed in banks, airports, retail outlets, for access to buildings or licensed venues, and when obtaining official documents or producing proof of age or identity. It is an interesting empirical question, therefore, how time pressure and the need to perform additional tasks might affect face matching performance. The present research examines these issues.

Experiment

Method

Participants Thirty-one first year psychology students from the University of Adelaide and 17 members of the general community were involved in the study. The students received course credit for their participation. Of the 48 participants, there were 33 females and 15 males, ranging in age from 17 to 58 years (M=23.27, SD=9.63).

Design A within-subjects design was used. Participants completed 400 trials, 100 under each condition (high time pressure with additional task, low time pressure with additional task, high time pressure with no additional task, low time pressure with no additional task). Each condition included 90 'same face' pairs showing two images of the same person and ten (randomly distributed) 'different faces' pairs showing different people. Every participant viewed the same pairs

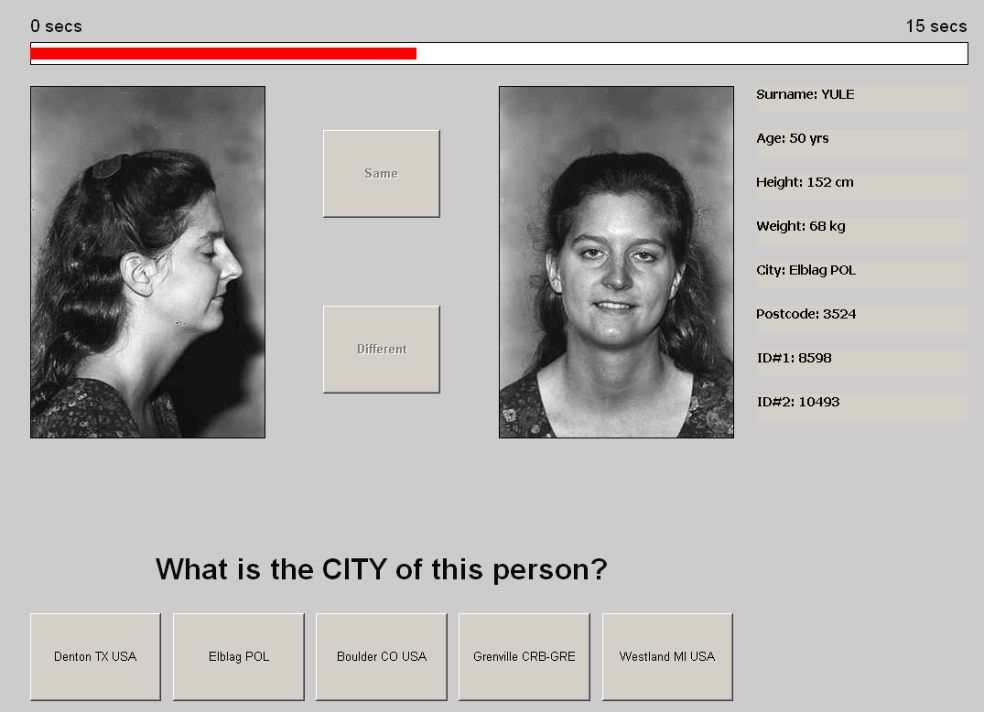

Figure 1: The experimental interface.

of images in the same order, with each face and each face pair presented only once during the experiment. The order in which participants completed the four conditions was counterbalanced using a pseudo-Latin Squares design. Thus, the order and presentation of the images remained exactly the same for every participant: only the experimental condition in which the participant viewed each face pair was varied.

Face Stimuli The face images were selected from the FERET database (Phillips, Wechsler, Huang, & Rauss 1998). The facial imagery in this database was collected in the US between 1993 and 1996 and includes 500 greyscale images of human faces, most of which are photographed several times with variations of angle, expression, lighting or perspective, and some with changes in hairstyle, glasses or clothing.

For the experimental stimuli, the first image selected was a frontal pose with a regular facial expression (i.e., smiling or neutral). For 'same face' pairs, the comparison image was selected from the range of remaining images of the same person, which usually involved a slight alteration such as an alternative facial expression, quarter left or right pose, half left or right pose, or profile left or right pose, or changes to lighting, hairstyle, glasses or clothing. For 'different faces' pairs, the comparison image was selected from those of different, but similar looking, people.

Procedure Figure 1 shows the experimental interface. The first face is shown on the right hand side, with the comparison face on the left. 'Same' and 'different' decision buttons are located between the faces. An identification card, giving information on the surname, age, height, weight, city, postcode and two different identification numbers of the person, was located beside the first face on the right hand side.

On every trial, a bar moving across the top of the screen indicated the time remaining for the participant to make their choice. The high time pressure condition imposed a time limit of six seconds, while the low time pressure condition imposed a limit of 15 seconds. These values were determined on the basis of pilot studies.

On every trial within an experimental condition involving the additional task, prior to making a 'same' or 'different' decision, participants were required to answer a question relating to the details on the identification card. The question was multiple choice, with five different response options, and related (at random) to one of the identification card fields. Participants were required to answer the question correctly before proceeding to the face recognition, making their response by clicking the button at the bottom of the interface corresponding to the correct answer. There was a delay of one second for each incorrect answer, to discourage guessing.

Testing sessions were held in a designated computer room at various times over a period of two weeks. An example screen shot, similar to Figure 1, displaying images that were not used in the experiment, was shown to participants to familiarize them with the experimental interface. This was accompanied by a detailed verbal explanation. Participants were asked to respond as quickly and accurately as possible, and the whole procedure took 30 to 60 minutes.

Results

Precision and Recall Decisions are naturally partitioned into four classes. 'True accepts' are correct decisions that two faces are of the same person; 'true rejects' are correct decisions that two faces are of different people; 'false accepts' are incorrect decisions that two faces are of the same person, when in fact they are of different people; 'false rejects' are incorrect decisions that two faces are of different people, when in fact they are of the same person.

In the current context, where the interest is in people's ability to detect that two faces are different, it is useful to define indices of precision and recall relative to 'different' decisions. Formally, precision is defined as the proportion of true rejects relative to the total of true rejects and false rejects, while recall is defined as the proportion of true rejects relative to the total of true rejects and false accepts. Intuitively, precision measures the proportion of face pairs that were actually different out of those decided to be different, while recall measures the proportion of face pairs decided to be different out of those that actually were different. In other words, precision indicates how accurately people find mismatched face pairs, while recall indicates how thoroughly people find mismatched face pairs.

Because it was possible for participants to fail to make a decision, two types of precision and recall measures are possible. Under a 'stringent' policy, non-decisions are regarded as 'different' decisions, while under a 'lenient' policy, non-decisions are regarded as 'same' decisions. Intuitively, stringent conditions correspond to a scenario where guilt is presumed, and face pairs are regarded as different unless an explicit 'same' decision is made, while lenient conditions correspond to the assumption of innocence, with face pairs regarded as the same unless an explicit 'different' decision is made. Accordingly, stringency emphasizes recall at the expense of precision, but leniency emphasizes precision at the expense of recall.

Figure 2 shows both stringent and lenient precision and recall performance for every participant in every condition, as well as aggregated performance across all participants in each condition. Stringent measures are represented by crosses, while lenient measures are represented by circles. Aggregate measures are shown in bold. The most striking feature of Figure 2 is that there are large individual differences in performance,

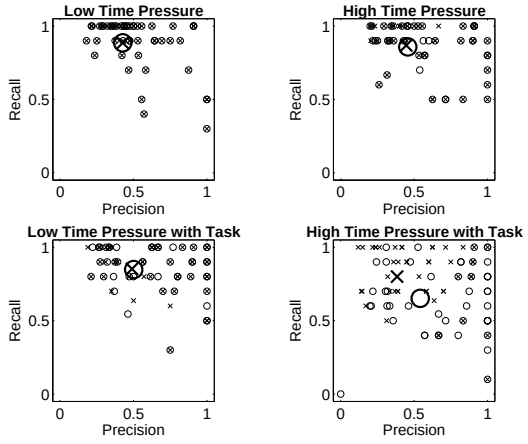

Figure 2: Precision and recall performance for every participant in every condition, as well as aggregated performance across all participants in each condition, under both stringent and lenient policies.

even in the condition with low time pressure and no additional task, and the variation across individuals increases with greater time pressure or the introduction of the additional task. It is clear that recall performance is generally better than precision performance, and it is noteworthy that, even with low time pressure and no additional task to perform, not a single participant made all decisions correctly.

In terms of the aggregate measures of precision and recall, Figure 2 suggests that performance across the four conditions is remarkably similar. This conclusion is borne out by Table 1, which shows the aggregate precision and recall measures, together with ranges corresponding to 95% confidence intervals. These confidence intervals generally overlap one another across conditions, except for the condition with high time pressure and an additional task. In this case, precision and recall measures diverge depending on whether a stringent or lenient policy is adopted. This finding suggests that there are relatively greater number of trials in the final condition where no decision was made.

Non-Decisions Table 2, which shows the total number of non-decisions in each condition, confirms this intuition. There were more than twice as many non-decisions in the condition with high time pressure and the additional task. Table 2 also partitions the non-decisions into those involving 'same face' versus 'different faces' pairs. Recalling that there are 90 'same face' pairs and only 10 'different faces' pairs in each condition, the percentages provide the most useful account of any potential differences. While the percent-

Table 1: Aggregate recall and precision, with ranges showing 95% confidence intervals, for all four experimental conditions, under both stringent and lenient policies.

Condition	Stringent Precision	Lenient Precision	Stringent Recall	Lenient Recall
low	0.43 (0.40-0.46)	0.43 (0.40-0.46)	0.89 (0.84-0.93)	0.89 (0.84-0.93)
high	0.45 (0.41-0.48)	0.46 (0.42-0.49)	0.87 (0.82-0.91)	0.86 (0.81-0.90)
low+task	0.49 (0.46-0.52)	0.50 (0.47-0.54)	0.85 (0.81-0.90)	0.85 (0.80-0.89)
high+task	0.38 (0.35-0.41)	0.54 (0.50-0.58)	0.80 (0.76-0.84)	0.65 (0.61-0.70)

Table 2: Total number and percentage of non-decisions in each condition, together with numbers and percentages for 'same face' and 'different faces' pairs.

Condition	Same	Different	Total
low	55 (1.27%)	0 (0%)	55 (1.27%)
high	63 (1.46%)	5 (1.04%)	68 (1.42%)
low+task	71 (1.64%)	2 (0.42%)	73 (1.52%)
high+task	96 (2.22%)	71 (14.79%)	167 (3.48%)

ages are all low in the first three conditions, and so both 'same face' and 'different faces' pairs have similar performance, there seems to be a clear difference in the condition with high time pressure and the additional task. In this condition, more than 14% of 'different faces' pairs resulted in no decision, compared to only about 2% for the 'same face' pairs.

Response Times An analysis of response times is provided in Figure 3, which shows the distributions for true accepts, true rejects, false accepts and false rejects in all four conditions. In the condition with low time pressure and no additional task, the majority of participants responded in approximately three seconds, with few participants taking more than five seconds. Under the high time pressure condition, participants responded slightly faster, particularly for true accepts; the majority of responses being made between one and three seconds. Response times lengthened in the low time pressure with task condition, but still rarely exceeded ten seconds.These results indicate that even when extra time was available for participants to view the faces, it was not used. Interestingly, there do not generally seem to be large differences in response time distributions between the correct and incorrect decisions, or the accept and reject decisions. The only possible difference might be slightly faster responding for true accepts in some of the conditions.

Figure 4 shows how total response times were divided between the additional task and the face matching decision, for the condition with low time pressure. It is clear that the additional task took at most nine or ten seconds for almost all trials, leaving five or six

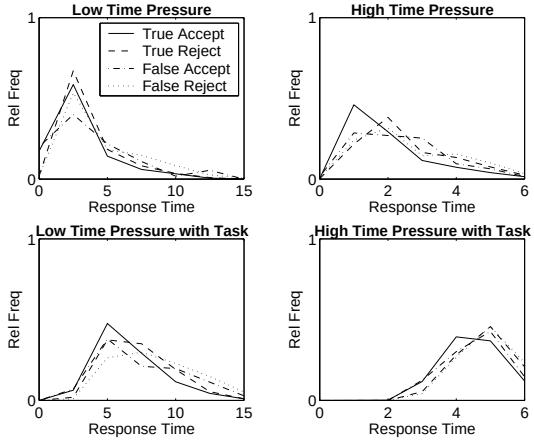

Figure 3: Response time distributions for each type of decision in each condition.

seconds to make a face matching decision. This means participants had about the same available time to complete the face matching decision task as they did in the condition with high time pressure but no additional task. The extremely similar precision and recall measures in these conditions (i.e., low time pressure with task and high time pressure without task), therefore, suggests that participants were not able to attend to the information relevant to the face matching decision while completing the additional task.

Learning Figure 5 shows the stringent and lenient recall and precision scores for the first 50 trials in each condition, compared to the last 50 trials. From the first to second half, all of the measures move upwards and to the right for all conditions. This means that precision and recall scores either remained stable or improved from the first to the second half trials in each condition, under both policies. The consistent improvement suggests a learning effect, where participants were able to adapt over time and familiarize themselves with the demands of each specific condition. It is also worth noting that improvement was greatest in the two conditions with high time pressure.

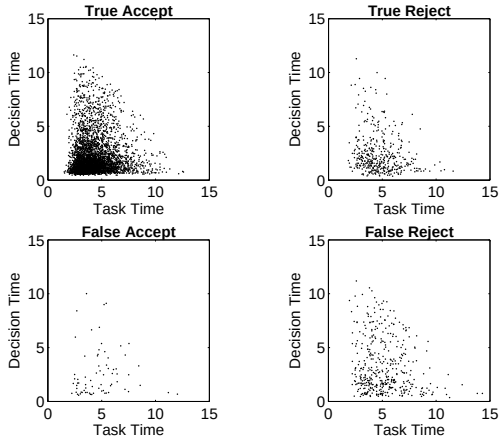

Figure 4: Relationship between time taken to complete the additional task and time taken to make the face matching decision, for each decision type in the condition with low time pressure.

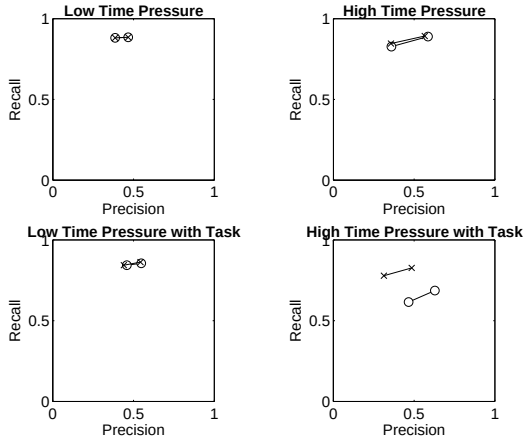

Figure 5: Recall and precision from the first half to the second half of trials in each condition. Crosses represent stringent scores, while circles represent lenient scores.

Face Properties There were large variations in the type and characteristics of the facial images that produced recognition errors. For example, 87% of the 400 face pairs used in the experiment produced at least one error. The three face pairs that produced the highest number of false reject errors used two facial images of the same person with either a change in hairstyle, clothing, glasses or pose. It is also interested to note that the two face pairs producing the highest number of false accept errors displayed the facial images of four different people of African-American appearance.

Discussion

In operational environments involving face matching, both recall and precision are important. High recall is vital to insure mismatched pairs are identified to prevent unlawful entry or access to an area, and precision is important to insure that there are few false alarms to prevent inconvenience, and minimize resources and transaction time. Neither recall nor precision were emphasized as the better measure to optimize in our experimental instructions, but participants clearly focused on recall over precision. That said, overall performance was unimpressive in even the most generous condition (i.e., low time pressure and no additional task) with a recall of about 90% and a precision below 50%. This means, even under these conditions, 10% of mismatched faces remained undetected, and over half of the faces that were rejected by participants as different, were in fact the same. These failures seem unlikely to be due to the 15 second time limit, since most decisions were made in less than half that time. Indeed, starting from this low initial level of performance, the introduction of increased time pressure or an additional task had little overall effect on performance. Only when both were combined did performance deteriorate further, largely as a result of an increased number of non-responses, particularly for face pairs that displayed the images of two different people.

The decisions associated with the different face stimuli are consistent with previous literature highlighting that recognition performance decreases with particular changes in appearance (Kemp et al. 1997; Luckman et al. 1995; Patterson & Baddeley 1977; Terry 2001). Since the majority of participants were Caucasian, the errors on African-American face pairs is also consistent with the previously reported own-race effect for face recognition (e.g., Furl, Phillips & O'Toole 2002), which suggests that people are better at recognising faces of their own race because they are more familiar with them, through every-day social interactions (Elliott, Wills & Goldstein 1973).

In applied settings, it seems unlikely the levels of performance observed in our experiment would be acceptable. As a result, these findings have significant implications for the design of new training solutions, or changed work practices to enhance face recognition performance in these environments. The performance found in the current experiment may provide justification for the introduction of automated face recognition biometric technologies (Burton et al. 2001; Hancock, Bruce & Burton 1998; Luckman et al. 1995).

Beyond the general findings, one of the most striking results is the presence of large individual differences in performance. Participants ranged from performing almost perfectly, to having recall or precision at or

below 50%. This strongly suggests that some people are better at face matching than other. Theoretically, it would be a worthwhile exercise understanding the nature of these differences. They could relate to representation, decision-making, or some other combination of basic cognitive processes. Practically, the observed differences highlight the importance of careful assessment in recruiting for applied environments if human decision-making is relied upon for face matching.

Ultimately, understanding human face matching performance, including individual differences, will require the development of cognitive models. Some of our results provide useful guidance about the form of candidate models. One good example is the lack of obvious differences in response time distributions across the decision types. If response times had been shorter for the reject decisions than the accept decisions, this would have indicated that participants were focusing predominantly on the differences between the images and were responding as soon as a sufficient degree of difference was detected. Instead, the similarities in response times suggest that participants were assessing both the features that were the same between each face pair, and the features that were different. This type of decision strategy is consistent with participants accumulating evidence directly in relation to both 'same' and 'different' decisions, and making a response once either threshold is met, or an externally imposed time limit is reached. This is essentially a sequential sampling process, of the type used extensively in other areas of cognitive modeling (e.g., Vickers 1979). The development of these and other models of human face matching is a priority for future research.

Acknowledgments

We acknowledge financial support from the Defence Science and Technology Organisation. Portions of the research in this paper use the FERET database of facial images collected under the FERET program.

References

Baudouin, J. Y., Gilibert, D., Sansone, S., & Tiberghein, G. (2000). When the smile is a cue to familiarity. *Memory, 8,* 285–292.

Burton, M. A., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: a comparison across image formats. *Vision Research, 41,* 3185–3195

DiNardo, L., & Rainey, D. W. (1991). The effects of illumination level and exposure time on facial recognition. *Psychological Record, 41,* 329–334

Elliott, E. S., Wills, E. J., & Goldstein, A. G. (1973). The effects of discrimination training on the recognition of white and oriental faces. *Bulletin of the Psychonomic Society, 2,* 71–73

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26,* 797–815

Hager, J. C., & Ekman, P. (1979). Long-distance transmission of facial affect signals. *Ethology and Sociobiology, 1,* 77–82

Hancock, P. J. B., Bruce, V., & Burton, M. A. (1998). A comparison of two computer-based face identification systems with human perceptions of faces. *Vision Research, 38,* 2277–2288

Harrigan, J. A., & Taing, K. T. (1997). Fooled by a smile: Detecting anxiety in others. *Journal of Nonverbal Behaviour, 21,* 203–221

Henderson, Z., Bruce, V., & Burton, M. A. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology, 15,* 445–464

Inui, T., & Miyamoto, K. (1984). The effect of changes in visible area on facial recognition. *Perception, 13,* 49–56

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11,* 211–222

Knight, B., & Johnston, A. (1997). The role of movement in face recognition. *Visual Cognition, 4,* 265–273

Lander, K., & Bruce, V. (2003). The role of motion in learning new faces. *Visual Cognition, 10,* 897–912

Laughery, K. R., Alexander, J. F., & Lane, A. B. (1971). Recognition of human faces: Effects of target exposure time, target position, pose position, and type of photograph. *Journal of Applied Psychology, 55,* 477–483

Lewis, M. B., & Edmonds, A. J. (2003). Face detection: Mapping human performance. *Perception, 32,* 903–920.

Liu, C. H., Seetzen, H., Buton, A. M., & Chaudhuri, A. (2003). Face recognition is robust with incongruent image resolution: Relationship to security video images. *Journal of Experimental Psychology: Applied, 9,* 33–41

Luckman, A. J., Allinson, N. M., Ellis, A. W., & Flude, B. M. (1995). Familiar face recognition: A comparative study of a connectionist model and human performance. *Neurocomputing, 7,* 3–27

Nagayama, R. (1999). The effects of facial expression and person on face recognition: A study using priming paradigm. *Japanese Journal of Psychology, 70,* 186–194

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology, 3,* 406–417.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing Journal, 16(5),* 295–306.

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100,* 139–156

Terry, R. L. (2001). Effects of facial transformations on accuracy of recognition. *The Journal of Social Psychology, 134,* 483–492

Vickers, D. (1979). *Decision Processes in Visual Perception.* New York: Academic Press.

Yip, A. W., & Sinha, P. (2002). Contribution of colour to face recognition. *Perception, 31,* 995–1003