massachusetts institute of technology — artificial intelligence laboratory

Relative Contributions of Internal and External Features to Face Recognition

Izzat N. Jarudi and Pawan Sinha

AI Memo 2003-004 March 2003
CBCL Memo 225

Abstract

The central challenge in face recognition lies in understanding the role different facial features play in our judgments of identity. Notable in this regard are the relative contributions of the internal (eyes, nose and mouth) and external (hair and jaw-line) features. Past studies that have investigated this issue have typically used high-resolution images or good-quality line drawings as facial stimuli. The results obtained are therefore most relevant for understanding the identification of faces at close range. However, given that real-world viewing conditions are rarely optimal, it is also important to know how image degradations, such as loss of resolution caused by large viewing distances, influence our ability to use internal and external features. Here, we report experiments designed to address this issue. Our data characterize how the relative contributions of internal and external features change as a function of image resolution. While we replicated results of previous studies that have shown internal features of familiar faces to be more useful for recognition than external features at high resolution, we found that the two feature sets reverse in importance as resolution decreases. These results suggest that the visual system uses a highly non-linear cue-fusion strategy in combining internal and external features along the dimension of image resolution and that the configural cues that relate the two feature sets play an important role in judgments of facial identity.

This work was supported by grants from the DARPA HumanID Program and the Alfred P. Sloan Fellowship in Neuroscience to PS. IJ is supported by the John Reed and the Meryl and Stewart Roberston Funds.

Introduction

In order to understand the basis for the human visual system's remarkable proficiency at the task of recognizing faces, we need to assess the contribution of different facial cues to judgments of identity. Amongst the most prominent of such cues are the most obvious ones: eyes, nose, mouth, hair and jaw. For any given face, these attributes have typically been placed into two groups: 'internal' attributes comprising the eyes, nose and mouth, and 'external' attributes comprising the hair and jaw-line. Several studies, reviewed below, have examined the relative roles of these groups. Our goal is to extend these studies by examining how the contributions of these sets of attributes change as a function of image resolution.

Image resolution is an important dimension along which to characterize face recognition performance. The change in image information that accompanies a reduction in image resolution mimics the information decrease caused by increasing viewing distances or common refractive errors in the optics of the eye. Understanding recognition under such conditions is of great ecological significance given their prevalence in the real world. In order for our visual apparatus to function effectively as an early alerting system, it necessarily has to be able to identify faces, and other objects, at large distances. Many automated vision systems too need to have the ability to interpret degraded images. For instance, images derived from present-day security equipment are often of poor resolution due both to hardware limitations and large viewing distances. Figure 1 is a case in point. It shows a frame from a video sequence of Mohammad Atta, a perpetrator of the World Trade Center bombing, at a Maine airport on the morning of September 11, 2001. As the inset shows, the resolution in the face region is quite poor. For the security systems to be effective, they need to be able to recognize suspected terrorists from such surveillance videos. This provides strong motivation for our work. In order to understand how the human visual system interprets such images and how a machine-based system could do the same, it is imperative that we study face recognition with such degraded images.

Fig. 1. *A frame from a surveillance video showing Mohammad Atta at an airport in Maine on the morning of the 11th of September, 2001. As the inset shows, the resolution available in the face region is very limited. Understanding the recognition of faces under such conditions remains an open challenge and provides the motivation for the work reported here.*

In this paper, we explore the relative importance of internal and external facial features as a function of resolution. While a rich body of past work has explored the contributions of these sets of features, most of the experiments have been conducted with high-resolution face images or good quality line drawings. One cannot be certain that the visual strategies revealed through the use of such images would continue to be valid when the available information is degraded as in common real-world viewing situations. Before proceeding further, it is worth noting that the issue of what constitutes a 'facial feature' is not entirely unambiguous. For instance, an image patch that included just part of the nose can be considered a feature in the same way that an eye on its own is. Indeed, this idea is supported by recent computational work (24, 25) that attempts to automatically learn face components. Along the same lines, the grouping of facial attributes into 'internal' and 'external' sets is open to interpretation. However, in this paper, we adopt this terminology to be consistent with conventional usage in the literature and to have our results be comparable with those from past studies.

Past work on the role of internal and external features in face recognition has indirectly suggested their relative contributions by proposing a "feature hierarchy" (1). By measuring reaction-time changes caused by the omission or substitution of facial features in schematic line-drawn faces, Fraser et al (1) reproduced previous findings by Haig (2) with photographic images that certain features are more important to face recognition than others. In particular, a feature hierarchy was observed with the head outline as the most significant, followed by the eyes, mouth and then nose. Other studies using different techniques have supported this general pattern of results suggesting that for the recognition of unfamiliar faces external features are more important that internal features (Bruce et al, 7; see also ref. 3 for a review on cue saliency). Ellis (4), for example, has even named external features like hair and facial shape "cardinal" features. With increasing familiarity, however, internal features become more significant. A number of studies have suggested that familiar faces are recognized better from their internal rather than external features. Ellis et al (5) found higher identification rates of famous faces from internal versus external features. Young et al (6) found that subjects were faster at matching a picture of a complete face with a picture showing only its internal features when the face was familiar rather than unfamiliar. Bruce and Young (8) hypothesize that we rely on internal features in recognizing familiar faces because external features like hair are more variable than internal features like eyes.

In the context of familiar face recognition, these studies raise some interesting questions. How does the relative significance of the internal and external features change when the stimuli are degraded to mimic real-world viewing conditions? At low resolutions, does feature saliency become proportional to feature size, favoring more global, external features like hair and jaw-line? Or, as in short-range recognition, are we still better at identifying familiar faces from internal features like the eyes, nose, and mouth? Even if we prefer internal features, does additional information from external features facilitate recognition? How does overall recognition performance relate to that obtained with the internal and external features independently? In other words, what kind of a cue-fusion strategy, linear or non-linear, does the visual system use for combining information from these two sets of features? The experiments we describe below are intended to address these questions.

Methods

Applying the methods of previous studies that have explored the influence of image degradation on face recognition in general (9, 10), we ran four experiments to determine the relative contributions of internal and external features to familiar face recognition as a function of image resolution. The four experiments characterized the resolution dependence of recognition performance under the following conditions:

Experiment A: Internal features placed in a row, thus not preserving their mutual spatial configuration.

Experiment B: Internal features in their correct spatial configuration.

Experiment C: External features alone.

Experiment D: Internal and external features together.

Stimuli

We collected twenty-four high-resolution color images of famous individuals in frontal views. The celebrities were movie or television actors and politicians. All twenty-four faces used were scale-normalized to have an inter-pupillary distance of 50 pixels. The heads were rotoscoped so that all images had a uniform white background. For each experiment, we created six 6 x 4 grids. One of the grids ('reference') was the same across all experiments and showed high-resolution full-face images. The remaining five were 'test' grids and differed for each experiment, as described below. Performance on the reference set was used to normalize our data since it indicated which of the celebrities pictured the subject was actually familiar with. Subsequent data analysis considered recognition data only for the individuals that the subject was able to identify in the reference grid.

Samples of stimuli used for each of the experiments are shown in Fig. 2. Except for one copy that was maintained at the original resolution, the test grids in each experiment were subjected to several levels of Gaussian blur with the resulting resolutions ranging from 1 cycle between the eyes to 4 cycles. In addition, all subjects were shown a full-face reference grid to normalize their performance.[*]

[*] In the full-face condition, the test grid at high resolution and the full-face reference grid were the same so there was a total of five grids in that experiment.

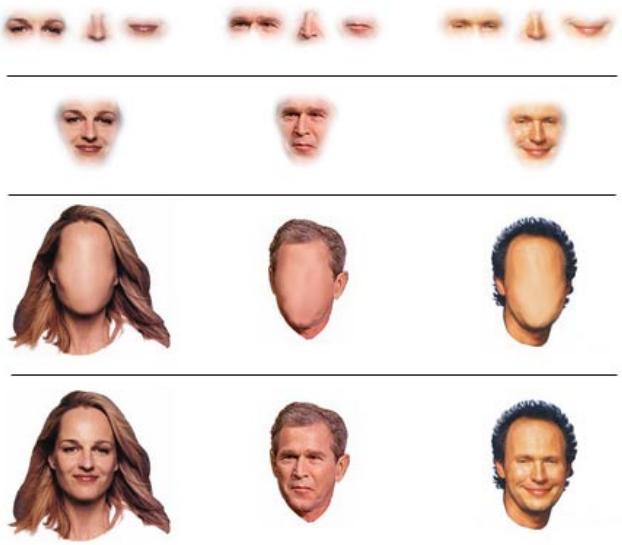

Fig. 2. *Sample stimuli from the four conditions used in our experiments. From top to bottom, they are:*
Experiment A: Internal features for each face placed in a row.
Experiment B: Internal features for each face in their correct spatial configuration.
Experiment C: External features alone with the internal features digitally erased.
Experiment D: Full faces, including both internal and external features.

Subjects

Thirty subjects, ranging in age from 18 to 38 participated in the study either for payment or course credit. Subjects were randomly placed in four non-overlapping groups corresponding to the four experiments (eight each in experiments A through C, and six in experiment D). The mutual exclusion was enforced to prevent any transfer of information from one condition to another.

Procedure

In each experiment, the grids were presented sequentially, proceeding from the most degraded to the least degraded conditions. The reference grid was shown subsequent to the presentation of all the test grids. For each grid, subjects were asked to identify as many of the celebrities as they could either by recording their name or some uniquely identifying information (such an identifier could simply specify a job for which the celebrity may have been famous; for example, for actors this would include the name of a movie, television show or character with which he or she may have been associated) on a response sheet. Subjects were permitted to guess and change their identifications in progressing from one grid to the next (but not in the reverse direction).

Images were presented on a color CRT monitor measuring 19" diagonally. Screen resolution was set to 1280 x 1024 pixels with a refresh rate of 85 Hz. Head movements and viewing time were not restricted. Average viewing distance was 60 cm.

Results

Figure 3 shows the experimental results. The graphs plot performance (proportion of faces correctly identified) as a function of image blur. Chance-level performance in all these

conditions is close to zero since subjects were unaware of which individuals they might see in the session.

The graph shows five plots: four correspond to each of the experimental conditions tested (A through D) and the fifth corresponds to the sum of performances obtained in conditions B and C (this is relevant for our discussion of cue-combination strategies). Performance in the full-face condition (D) is remarkably robust to reductions in image quality and declines gradually with increasing image blur. Even at a blur level of just 3 cycles between the eyes, performance is greater than 80%. This is in contrast to performance with the rearranged internal features. Even at the highest resolution, performance in this condition (A) is rather modest, averaging just a little over 50%. Furthermore, the curve drops sharply with decreasing image resolution. When the internal features are placed in their correct spatial configuration (condition B), performance improves relative to condition A, but continues to be extremely sensitive to the amount of blur applied to the stimulus images. The absence of external features, therefore, severely compromises performance in condition B relative to condition D. In contrast to the rapid fall-off for conditions A and B, the shallow slope of the curve for external features alone (condition C) indicates gradual performance change with increasing blur; however, the absolute level of performance all along this curve is poor, not exceeding 40% even at the highest resolution. Interestingly, at a resolution of approximately 3.5 cycles between the eyes, we observe a change in the rank-ordering of the curve for condition C relative to that for conditions A and B. Finally, Fig. 3 shows the fifth curve corresponding to the sum of performances obtained in condition B and condition C. Although the stimuli used in condition D can be obtained by a superposition of the stimuli in conditions B and C, it is evident that the result of summing performances with conditions B and C falls significantly short of the plot corresponding to condition D. Before we discuss the implications of these results, we briefly quantify the statistical significance of the key performance differences observed across conditions.

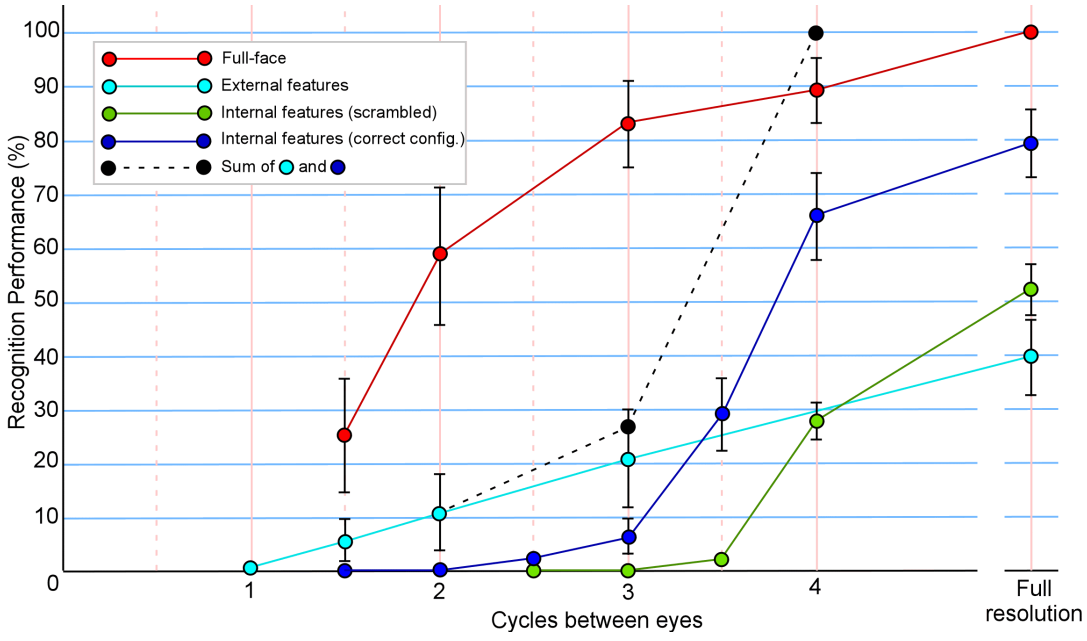

Fig. 3. *Recognition performance as a function of image resolution across the different conditions we tested. The dashed curve shows the sum of performances obtained with internal features (condition B) on the one hand and external features (condition C) on the other.*

A 2-factor ANOVA on combined data across blur levels and the four experimental conditions showed highly significant main effects of blur level ($F_{2,78}$=30.2, p<.001) and feature condition ($F_{3,78}$=174.8, p<.001) as well as a highly significant blur level-feature condition interaction ($F_{6,78}$=14.2, p<.001). At one extreme of the blur scale corresponding to high blurs (levels 1.5, 2 and 3), one-tailed t-tests revealed that performance in condition D was better than in condition C, which, in turn, was better than in conditions A and B (all $p < 0.02$). Means for performance in condition A and B were both identically zero at blur levels 1.5 and 2, while at level 3, condition B was better than condition A ($p < 0.01$).

At the other end of the blur scale, i.e. full resolution, one-tailed t-tests revealed that the rank-ordering of condition C relative to conditions A and B was reversed in favor of the latter ($p < 0.01$). Overall, performance in condition C (external features only) was higher than in condition A and B (internal features only) at low resolutions, while conditions A and B were better than condition C at full resolution (with condition B's configured features better than condition A's rearranged ones). Performance in condition D (full faces) was superior across all blur levels.

Discussion

Although the experiments we have described above are simple in their design, they allow us to make some interesting inferences about face recognition. First, they characterize full-face identification performance (condition D) as a function of available image resolution. This kind of result has been previously reported in the literature by a few researchers, including Harmon (10), Harmon and Julesz (11), Bachmann (9) and Costen et al. (12). However, our results from condition D are not merely a replication of past findings. Experiment D was designed to address some of the important limitations of earlier studies. For instance, Harmon and Julesz's (10, 11) results of recognition performance with block-averaged images of familiar faces were confounded by the fact that subjects were told which of a small set of people they were going to be shown in the experiment. More recent studies too have suffered from this problem. Bachmann (9) and Costen et al. (12) used six high-resolution photographs during the 'training' session and low-resolution versions of the same photographs during the 'test' sessions. The subject priming about stimulus set and the use of the same base photographs across the training and test sessions renders these experiments somewhat non-representative of real-world recognition situations. Another drawback of these studies that widens the gulf between the experiments and real-world settings, is that the images used were exclusively monochrome. Recent experiments show that color increasingly contributes to face recognition with decreasing image resolution (13). Therefore, we believe that the results reported here with full-color images are more representative of performance in real-world viewing conditions.

Besides characterizing full-face recognition performance with degraded images, our results show how the relative contributions of internal and external features to face recognition change as a function of image resolution. As borne out by the results in condition A, the appearance of the internal features independently provides a very limited cue to identity. Even at

the highest resolution tested, subjects' performance was quite compromised, averaging only half of the performance they obtained with full faces. Surprisingly, the inclusion of correct configural information (condition B), improved performance only marginally. As with condition A, performance remained very sensitive to image resolution, declining rapidly with increasing image blur.

On the other hand, performance with external features only shows a more gradual fall-off with increasing blur, suggesting external features are more useful for face identification than internal features in poor viewing conditions. This saliency of external features at low resolutions could be accounted for by appealing to an image property as simple as feature size. It may be that the human visual system relies more on large features like the hair and jaw-line when the facial image is blurry because information about their appearance survives greater image degradations than the smaller features like the eyes, nose and mouth. This explanation is supported by the results for condition C at full resolution where the size advantage for external features disappears and consequently, the presentation of internal features alone (condition A and B) leads to better performance. Our data under these optimal viewing conditions agree with previous studies that have also found that internal features of familiar faces in high-resolution images are more useful for recognition than external features (5, 6).

Our finding of a greater reliance on external features for face recognition at low resolutions has an interesting analogue in the developmental literature. Reports from several researchers studying face recognition by neonates (14-16) suggest that infants initially depend more on external features than on internal ones for discriminating between individuals. For instance, Pascalis et al (16) found that although four-day old infants could reliably discriminate their mother's face when all the facial information was present, they were unable to make the distinction when their mother and a stranger wore scarves around their heads. In conjunction with the fact that infant visual acuity starts out being very poor and improves over time (17-19), these results echo our finding with adult observers. We speculate, therefore, that infant reliance on external features may, as for our adult subjects, be driven at least in part by considerations of which subset of facial information provides more useful cues to identity at a given resolution.

Beyond helping to identify which of the two feature sets is more important at different resolutions, the results from our experiments also answer the question of whether a linear combination of performances obtained with internal and external features separately can account for the performance obtained with the full face. This does not seem to be the case since a sum of performances with conditions B and C does not replicate the results with condition D. It thus appears that the visual system does not use a simple linear cue fusion strategy for combining information from the internal and external facial attributes. This non-linearity is particularly evident at high levels of blur. Even when internal and external features stop being useful independently (leading to performance close to zero), together the two sets yield a high level of performance. This aspect of the results has an important implication. It suggests that at least under conditions of low resolution, it is not the internal or external configurations on their own that subserve recognition, but rather measurements corresponding to how internal features are placed relative to the external features. This idea conflicts with conventional notions of facial configuration, especially prominent in the computational vision community, which primarily involve 'internal' measurements such as inter-eye, eye to nose-tip and nose-tip to mouth

distances (20, 21). Thus, external features, even though poor indicators of identity on their own, provide an important frame of reference for analyzing facial configuration.

Additional support for this idea comes from the observation illustrated in Fig. 4. In a separate set of experiments to be detailed in a forthcoming paper, we have found that independently transforming internal and external features disrupts recognition performance more significantly than transforming both together. Fig. 4 schematically summarizes the basic finding. Independent scaling of the internal and external features (equally or unequally in X and Y dimensions) leads to faces that are harder to recognize than those resulting from simultaneous scaling.

A visual illusion that was developed a few years ago (22) also serves to illustrate this idea. Fig. 5a shows what appears to be a picture of former US President Bill Clinton and Vice President Al Gore. Upon closer examination, it becomes apparent that Gore's internal features have been supplanted by Clinton's (in the configuration that they have on Clinton's face). If the appearance and mutual configuration of the internal features were the primary determinants of facial identity, then this illusion would have been much less effective because we would readily have seen the image as two identical faces. However, given that most observers do not, it seems valid to conclude that external features play a very significant role in judgments of identity. Furthermore, their contribution becomes evident only in concert with the internal features because on their own they do not permit reliable recognition. Fig. 5b shows this illusion updated to incorporate the current team in the White House (23).

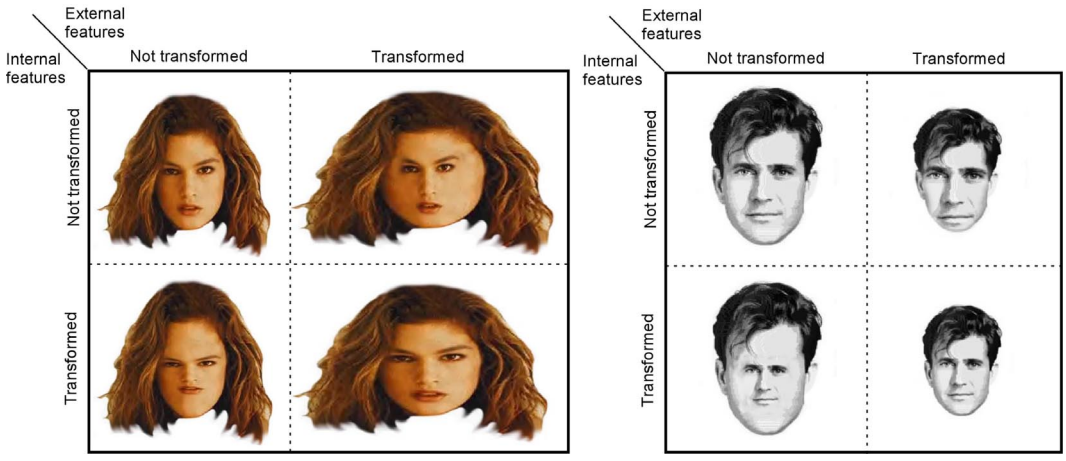

Fig. 4. *Independent transformation of internal and external features (top-right and lower-left boxes in the two grids above) disrupts recognition performance more than simultaneous transformation of both (lower right boxes). This points to the perceptual importance of the relationships between the two sets of features rather than just the configural cues within either set alone. The two individuals shown here are the model Cindy Crawford and actor Mel Gibson.*

(a) (b)

Fig. 5. *Two versions of the 'Presidential Illusion' devised by Sinha and Poggio (22, 23) that highlight the significance of external features in face recognition.*

By simulating the limiting conditions of face recognition at a distance, this study sheds light on the relative significance of internal and external features to the demands of everyday face recall. The pragmatic significance of such understanding lies in helping to design artificial recognition systems that may be better suited to dealing with the kinds of image degradations common to real settings.

References

1. Fraser, I. H., Craig, G. L. & Parker, D. M. (1990) *Perception* **19,** 661-673.

2. Haig, N. D. (1986) *Perception* **15,** 235-247.

3. Shepherd, J., Davies, G. & Ellis, H. (1981) in *Perceiving and remembering faces*, eds. Davies, G., Ellis, H. & Shepherd, J. (Academic Press, London), pp. 105-131.

4. Ellis, H. D. (1986) *Human Learning* **5,** 189-196.

5. Ellis, H. D., Shepherd, J. W. & Davies, G. M. (1979) *Perception* **8,** 431-439.

6. Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M. & Ellis, A. W. (1985) *Perception* **14,** 737-746.

7. Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M. & Miller, P. (1999) *J. Exp. Psychol. Applied* **5,** 339-360.

8. Bruce, V. & Young, A. (1998) *In the eye of the beholder: the science of face perception* (Oxford University Press, Oxford, England).

9. Bachmann, T. (1991) *Eur. J. Cogn. Psychol.* **3,** 87-103.

10. Harmon, L. D. (1973) *Scientific American* **229,** 70-83.

11. Harmon, L. D. & Julesz, B. (1973) *Science* **180,** 1194-1197.

12. Costen, N. P., Parker, D. M. & Craw, I. (1996) *Percept. Psychophys.* **58,** 602-612.

13. Yip, A. W. & Sinha, P. (2002) *Perception* **31,** 995-1003.

14. Field, T. M., Cohen, D., Garcia, R. & Greenberg, R. (1984) *Inf. Behav. Dev.* **7,** 19-25.

15. Bushnell, I. W. R., Sai, F. & Mullin, J. T. (1989) *Br. J. Dev. Psychol.* **7,** 3-15.

16. Pascalis, O., de Schonen, S., Morton, J., Deruelle, C. & Rabre-Grenet, M. (1995) *Inf. Behav. Dev.* **18,** 79-85.

17. Dobson, V. & Teller, D. Y. (1978) in *Visual Psychophysics and Physiology*, eds. Armington, J. C., Krauskopf, J. & Wooten, B. R. (Academic Press, New York).

18. Dobson, V. (1993) in *Early Visual Development, Normal and Abnormal*, ed. Simons, K. (Oxford University Press, New York).

19. Hainlin, L. & Abramov, I. (1992) in *Advances in Infancy Research*, eds. Rovee-Collier, C. & Lipsitt, L. P. (Ablex, Norwood, NJ), Vol. 7.

20. Doi, M., Sato, K. & Chihara, K. (1998) in *Third IEEE Intl. Conf. on Automatic Face and Gesture Recog.* (IEEE Computer Society Press, Nara, Japan).

21. Brunelli, R. & Poggio, T. (1993) *IEEE Trans. Patt. Anal. & Mach. Intell.* **15,** 1042-1052.

22. Sinha, P. & Poggio, T. (1996) *Nature* **384,** 404-404.

23. Sinha, P. & Poggio, T. (2002) *Perception* **31,** 133-133.

24. Heisele, B., Serre, T., Pontil, M., and Poggio, T. (2001) Component-based face detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society Press, vol **1**, 657-662.

25. Serre, T., Riesenhuber, M., Louie, J., and Poggio, T. (2002). On the role of object-specific features for real-world object recognition in biological vision. In *Biologically Motivated Computer Vision*, Buelthoff et al. eds., Lecture Notes in Computer Science 2525, Springer.