Identifying perceptually significant features for recognizing faces

Pawan Sinha

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02142
sinha@ai.mit.edu

The human visual system possesses a remarkable ability to detect and identify faces even under degraded viewing conditions. The fundamental challenge in understanding this ability lies in determining which facial attributes the visual system uses for these tasks. Here we describe experiments designed to probe the limits of these abilities and determine the relative contributions of internal versus external facial features for the detection and identification tasks. The results provide strong constraints and guidelines for computational models of face perception.

Introduction

Many current machine-based face-processing systems share two common characteristics: 1. they require relatively high-resolution images in order to operate satisfactorily, and 2. they use primarily the inner section of the face (eyes, nose and mouth) while disregarding the external features (hair and jaw-line) as being too variable. It is instructive to ask how the requirements and performance of these systems compares with that of human observers. The human visual system (HVS) often serves as the de-facto standard for evaluating machine vision approaches. Clearly, in order to be able to use the human visual system as a useful standard to strive towards, we need to first have a comprehensive characterization of its capabilities.

In our investigation of the HVS's capabilities, we shall focus on two key face-perception tasks: face detection ('is this a face?') and face identification ('whose face is it?'). For both, we shall describe experiments that address two questions: 1. how does human performance change as a function of image resolution? and 2. what are the relative contributions of internal and external features at different resolutions? Let us briefly consider why these two questions are worthy subjects of study.

The decision to examine recognition performance in images with limited resolution is motivated by both ecological and pragmatic considerations. In the natural environment, the brain is typically required to recognize objects when they are at a distance or viewed under sub-optimal conditions. In fact, the very survival of an animal may depend on its ability to use its recognition machinery as an early-warning system that can operate reliably with limited stimulus information. Therefore, by better capturing real-world viewing conditions, degraded images are well suited to help us understand the brain's recognition strategies. More pragmatically, impoverished images serve as 'minimalist' stimuli, which, by dispensing with unnecessary detail, can potentially simplify our quest to identify aspects of object information that the brain preferentially encodes.

The decision to focus on the two types of facial feature sets – internal and external, is motivated by the marked disparity that exists in their use by current machine-based face analysis systems. It is typically assumed that internal features (eyes, nose and mouth) are the critical constituents of a face, and the external features (hair and jaw-line) are too variable to be practically useful. It is interesting to ask whether the human visual system also employs a similar criterion in its use of the two types of features.

It is important to stress that the limits of human performance do not necessarily define upper bounds on what is achievable. Specialized identification systems (say those based on novel sensors, such as IR cameras) may well exceed human performance. However, in many real-world scenarios using conventional sensors, matching human performance remains an elusive goal. Our experiments can not only give us a better sense of what this goal is, but also what computational strategies we could employ to move towards it and, eventually, past it.

Face Detection

Previous studies of face perception in degraded images have been designed exclusively to study within-class discrimination ('whose face is it?') rather than face classification per se ('is this a face?'). Consequently, no systematic data exist about the dependence of face-detection performance on image resolution and the relative contribution of internal versus external facial attributes. We have conducted a series of experiments to address these issues with the goal of characterizing the nature of facial representations used by the human visual system. The detailed set of studies can be found in (Torralba and Sinha, 2001). Here, we report two specific experiments:

Experiment 1: How does face detection accuracy with inner facial features change as a function of available image resolution?
Experiment 2: Does the inclusion of external features improve face detection performance?

To be able to conduct these experiments, we have to confront an interesting question - what patterns should we use as non-faces? Selecting random fragments from non-face images is not a well-controlled approach. The face/non-face discrimination can be rendered unnaturally easy for certain choices of non-face images (for instance, imagine drawing non-face patterns from a sky image). We need a more principled approach to generating non-face patterns.

In very general terms, we would like to be able to draw our non-face patterns from the same general area in a high-dimensional object space where the face patterns are clustered. Morphing between face and non-face patterns is not a satisfactory strategy since all the intermediate morphs do have a contribution from a genuine face pattern and cannot, therefore, be considered true non-faces. An alternative strategy lies in using computational classification systems that operate by implicitly encoding clusters in multidimensional spaces [Yang & Huang, 1994; Sung and Poggio, 1994; Rowley et al, 1995]. Non-face patterns on which such systems make mistakes can then serve as the distractors for our psychophysical tasks. This is the approach we have used in our work. The key caveat to keep in mind here is that the multidimensional cluster implicitly used by these computational systems may be different from the cluster encoded by the human visual system. However, based on the high-level of classification accuracy that at least some of these systems exhibit, it is reasonable to assume that there is a significant amount of congruence between the clusters identified by them and human observers.

Experiment 1: Face detection at low-resolution

What is the minimum resolution needed by human observers to reliably distinguish between face and non-face patterns? More generally, how does the accuracy of face classification by human observers change as a function of available image resolution? These are the questions our first experiment is designed to answer.

Methods

Subjects were presented with randomly interleaved face and non-face patterns and, in a 'yes-no' paradigm, were asked to classify them as such. The stimuli were grouped in blocks, each having the same set of patterns, but at different resolutions. The presentation order of the blocks proceeded from the lowest resolution to the highest. Ten subjects participated in the experiment. Presentations were self-timed.

Our stimulus set comprised 200 monochrome patterns. Of these, 100 were faces of both genders under different lighting conditions (set 1), 75 were non-face patterns (set 2) derived from a well-known face-detection program (developed at the Carnegie Mellon University by Rowley et al [1995]) and the remaining 25 were patterns selected from natural images that have similar power-spectra as the face patterns (set 3). The patterns included in set 2 were false alarms (FAs) of Rowley et al's computational system, corresponding to the most conservative acceptance criterion yielding 95% hit rate. Sample non-face images used in our experiments are shown in figure 1. All of the face images were frontal and showed the face from the middle of the forehead to just below the mouth. Reduction in resolution was accomplished via convolution with Gaussians of different sizes (with standard deviations set to yield 2, 3, 4, and 6 cycles per face; these correspond to 1.3, 2, 2.5 and 3.9 cycles within the eye-to-eye distance ('ete'). All spatial resolutions henceforth are reported in terms of number of cycles between the two eyes).

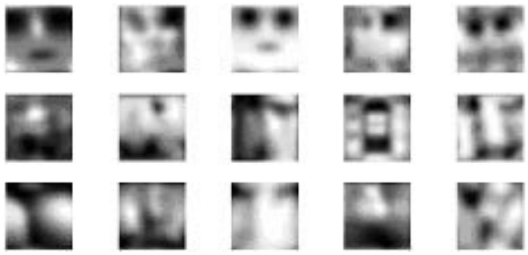

Figure 1. A few of the non-face patterns used in our experiments. The patterns comprise false alarms of a computational face-detection system and images with similar spectra as face images.

From the pooled responses of all subjects at each blur level, we computed the mean hit-rate for the true face stimuli and false alarm rates for each set of distractor patterns. These data indicated how subjects' face-classification performance changed as a function of image resolution. Also, for a given level of performance, we were able to determine the minimum image resolution required.

Results

Figure 2 shows data averaged across 10 subjects. Subjects achieved a high hit rate (96%) and a low false-alarm rate (6% with Rowley et al's FPs and 0% with the other distractors) with images having only 3.9 cycles between the eyes. Performance remained robust (90% hit-rate and 19% false-alarm rate with the Rowley et al's FA distractor set) at even higher degrees of blur (2 cycles/ete). In proceeding from 2 to 1.3 cycles/ete, the hit-rate fell appreciably, but subjects were still able to reliably distinguish between faces and non-faces.

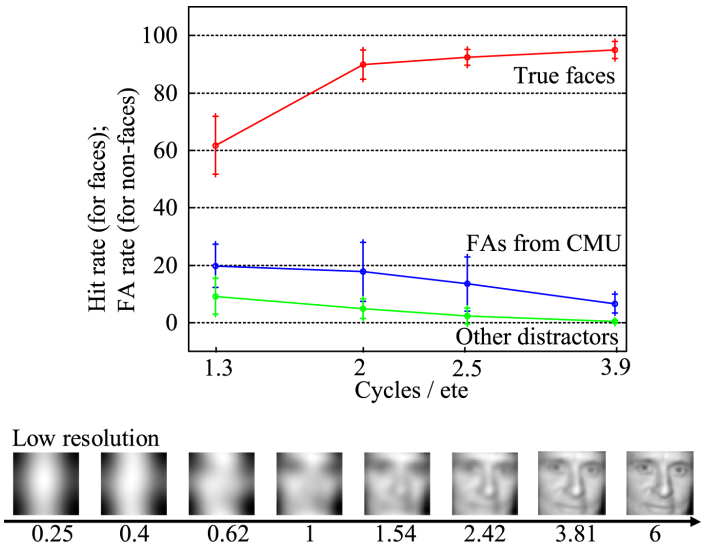

Figure 2. Results from experiment 1.The units of resolution are the number of cycles eye to eye.

The data suggest that faces can be reliably distinguished from non-faces even at just 2 cycles eye-to-eye using only the internal facial information. At even lower resolution (1.3 cycles/ete), while the hit rate falls significantly, there is still a clear distinction between true faces and distractors. Performance reaches an asymptote around 4 cycles/ete.

Impressive as this performance of the HVS is, it may be an underestimate of observers' capabilities. It is possible that the inclusion of context can improve performance further. In other words, in experiment 1, subjects made the face vs. non-face discrimination on the basis of the internal structure of faces. It has traditionally been assumed that this is the pattern that defines a face. However, it is not known whether the HVS can additionally use the external features to improve its discrimination and to better tolerate image resolution reductions. Experiment 2 addresses this issue.

Experiment 2: The role of local context in face-detection

The prototypical configuration of the eyes, nose and mouth (the 'internal features') intuitively seems to be the most diagnostic cue for distinguishing between faces and non-faces. Indeed, machine based face detection systems typically rely exclusively on internal facial structure [Sung & Poggio, 1994; Rowley et al., 1995; Leung et al. 1995]. External facial attributes such as hair, facial bounding contours and jaw-line are believed to be too variable across individuals for inclusion in a stable face representation. These attributes constitute the local context of internal facial features. To assess the contribution of local context to face-detection, we repeated experiment 1 with image fragments expanded to thrice their sizes in each dimension (see figure 3). The experimental paradigm was the same as for experiment 1. Subject pools for experiments 1 and 2 were mutually exclusive.

Figure 3. Faces (left set) and non-faces (right set) with local context.

Results

We tested 10 subjects on the 'expanded' version of images used in experiment 1. Figure 4 shows the results. Performance improved significantly following this change. Faces could be reliably distinguished from non-faces even with just 4 cycles across the entire image (which translates to 0.87 cycles/ete). At this resolution, the internal facial features become rather indistinct and, as the results from experiment 1 suggest, they lose their effectiveness as good predictors of whether a pattern is a face or not. It is also important to note that the contextual structure across different stimuli used in this experiment is very different. Faces were photographed against very different backgrounds and no effort was made to normalize the appearance of the context. Given that there is not enough consistent information within the face or outside of it for reliable classification, the likely explanation for the human visual system's impressive performance is that bounding contour information is incorporated in facial representations used for detection. As figure 4 shows, for comparable levels of performance, the use of bounding contours nearly halves the resolution lower-bounds needed for distinguishing faces from non-faces relative to the internal features only condition. Thus, the inclusion of bounding contours allows for tolerance to greater refractive errors in the eyes and/or longer viewing distances. This result also provides a useful hint for the design of artificial face detection systems. By augmenting their facial representation to include bounding contours, computational systems can be expected to improve their performance markedly.

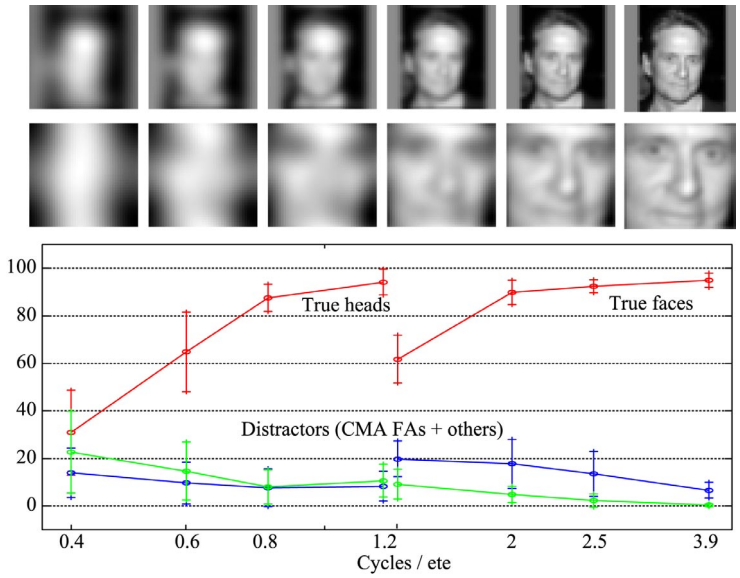

Figure 4. Results from experiment 2. The inclusion of local context (as shown in the top panel) significantly improves face detection performance and reduces resolution lower-bounds for reliable discrimination.

To the best of our knowledge, this is the first systematic study of face-detection across multiple resolutions. The data provide lower-bounds on image-resolution sufficient for reliable discrimination between faces and non-faces. They indicate that the facial representations encode, and can be matched against, facial image fragments containing merely 2 cycles between the two eyes. We can also demarcate zones on the resolution axis where specific facial attributes (internal features, bounding contours) suffice for achieving a given level of detection performance. Additionally, the results show that the inclusion of facial bounding contours substantially improves face detection performance (figure 4), suggesting that the facial representations encode this information.

The data also show that even under highly degraded conditions, humans are correctly able to reject most non-face patterns that the artificial systems confuse for faces. To further underscore the differences in capabilities of current computational face detection systems and the HVS, it is instructive to consider the minimum image resolution needed by a few of the proposed machine-based systems: 19x19 pixels for Sung and Poggio [1994]; 20x20 for Rowley et al [1995]; 24x24 for Viola and Jones [2001] and 58x58 for Heisle et al. [2001]). Thus, computational systems not only require a much larger amount of facial detail for detecting faces in real scenes, but also yield false alarms that are correctly rejected by human observers even at resolutions much lower than what they were originally detected at.

Face identification

Everyday, we are confronted with the task of face identification at a distance and must extract the critical information from the resulting low-resolution images. Precisely how does face identification performance change as a function of image resolution? Does the relative importance of facial features change as a function of image resolution? Does featural saliency become proportional to featural size, favoring more global, external features like hair and jaw-line? Or, 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? Our experiments were designed to address these open questions in face recognition by assessing face recognition performance across various resolutions and by investigating the contribution of internal and external features. Considering the importance of these issues, it is not surprising that a rich body of research has accumulated over the past few decades. Pioneering work on face recognition with low-resolution imagery was done by Harmon and Julesz [1973a, 1973b]. Working with block averaged images of familiar faces of the kind shown in figure 5, they found high recognition accuracies even with images containing just 16x16 blocks. However, this high level of performance could have been due at least in part to 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. For instance, Bachmann [1991] and Costen et al. [1996] used six high-resolution photographs during the 'training' session and low-resolution versions of the same during the test sessions. The prior 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. Also, the studies so far have not performed some important comparisons. Specifically, it is not known how performance differs across various image resolutions when subjects are presented full faces versus when they are shown the internal features alone.

Figure 5. Images such as the one shown here have been used by several researchers to assess the limits of human face identification processes.

Our experiments on face recognition were designed to build upon and correct some of the weaknesses of the work reviewed above. Here, we describe an experimental study with two goals: 1. assessing performance as a function of image resolution and 2. determining performance with internal features alone versus full faces.

The experimental paradigm we used required subjects to recognize celebrity facial images blurred by varying amounts. We used 36 color face images and subjected each to a series of blurs. The subjects were shown the blurred sets, beginning with the highest level of blur and proceeding on to the zero blur condition. We also created two other stimulus sets. The first of these contained the individual facial features (eyes, nose and mouth), placed side by side while the second had the internal features in their original spatial configuration. Three mutually exclusive groups of subjects (each containing 8 individuals) were tested on the three conditions. In all these experiments, subjects were not given any information about which celebrities they would be shown during the tests. Chance level performance was, therefore, close to zero.

Results

Figure 6 shows results from the different conditions. It is interesting to note that in the full-face condition, subjects can recognize more than half of the faces with image resolutions of merely 7x10 pixels. Recognition reaches almost ceiling level at a resolution of 19x27 pixels.

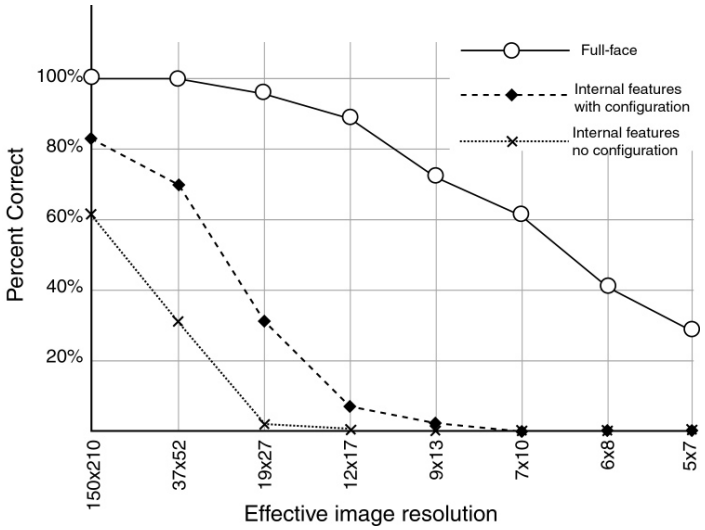

Figure 6. Recognition performance with internal features (with and without configural cues). Performance obtained with whole head images is also included for comparison.

Performance of subjects with the other two stimulus sets is quite poor even with relatively small amounts of blur. This clearly demonstrates the perceptual importance of the overall head configuration for face recognition. The internal features on their own and even their mutual configuration is insufficient to account for the impressive recognition performance of subjects with full face images at high blur levels. This result suggests that feature-based approaches to recognition are likely to be less robust that those based on the overall head configuration. Figure 7 shows an image that underscores the importance of overall head shape in determining identity.

Figure 7. Although this image appears to be a fairly run-of-the-mill picture of Bill Clinton and Al Gore, a closer inspection reveals that both men have been digitally given identical inner face features and their mutual configuration. Only the external features are different. It appears, therefore, that the human visual system makes strong use of the overall head shape in order to determine facial identity. (From Sinha and Poggio, 1996)

Conclusion

Progressive improvements in camera resolutions provide ever-greater temptation to use increasing amounts of detail in face representations in machine vision systems. Higher image resolutions allow recognition systems to discriminate between individuals on the basis of fine differences in their facial features. The advent of iris based biometric systems is a case in point. However, the problem that such details-based schemes often have to contend with is that high-resolution images are not always available. This is particularly true in situations where individuals have to be recognized at a distance. In order to design systems more robust against image degradations, we can turn to the human visual system for inspiration. Many of the factors that the human visual system has to be tolerant to are the same ones that today's computer vision systems are trying to grapple with. It makes sense, therefore, for us to turn to the human visual system in our search for clues about effective processing schemes. This is the key motivation underlying the experimental studies we have described here. The goal is to establish a common performance standard that different computer vision schemes can be evaluated against, and also to gain insights into what kinds of image information the human brain relies on to accomplish its feats of recognition. Our experimental results suggest some surprising lower-bounds on image resolutions sufficient for different face-perception tasks and also indicate the types of facial attributes the HVS relies on.

Our current work focuses on computationally modeling the experimental data reported here. We have made some headway on the task of face detection (Thoresz and Sinha, 2001; Sadr et al., 2001). The key question we have addressed is: What kind of internal representations can support robust detection performance across different illumination conditions and resolutions? We suggest that a candidate answer may be found in the response properties of early visual neurons. Based on available neuro-physiological evidence, we have developed a scheme that conceptualizes early visual neurons as rapidly saturating contrast edge detectors with large supports. This idealization leads to a representation scheme wherein objects are encoded as sets of qualitative image measurements over coarse image regions. The use of qualitative measurements leads not only to a reduction in the problem's computational complexity, but also renders the representations invariant to sensor noise and significant changes in object appearance.

Our approach uses qualitative photometric measurements to construct a face signature that is largely invariant to illumination changes and can operate on very low resolution images. We have tested a computer implementation of this scheme on a large database of real images containing frontal faces and have found the results to be encouraging (70% hit rate and 10% false alarm rate). Figure 8 shows some results of using the qualitative representation for face detection.

Figure 8. Results of using a qualitative representation scheme to encode and detect faces in images. Each box corresponds to an image fragment that the system believes is a face. The representation scheme is able to tolerate wide appearance and resolution variations.

References

Bachmann, T. (1991). Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology*, 3, 85-103.

Costen, N. P., Parker, D. M., & Craw, I. (1994). Spatial content and spatial quantization effects in face recognition. *Perception*, **23**, 129-146.

Harmon, L. D. & Julesz, B. (1973a). Masking in visual recognition: Effects of two-dimensional noise. *Science*, 180, 1194-1197.

Harmon, L. D. (1973b). The recognition of faces. Scientific American, 229(5), 70-83.

Heisle, B., T. Serre, S. Mukherjee and T. Poggio. (2001) Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images. In: *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, IEEE Computer Society Press, Jauai, Hawaii, December 8-14.

Leung, T. K., Burl, M. C., & Perona, P. (1995). Finding faces in cluttered scenes using random labeled graph matching. *Proc. Intl. Conf. On Comp. Vis.,* 637-644.

Rowley, H. A., Baluja, S., Kanade, T. (1995) Human face detection in visual scenes. CMU technical report# CS-95-158R. *Advances in Neural Information Processing Systems* ***8***, 1996, pp. 875 – 881.

Sadr, J., Mukherjee, S., Thoresz, K., and Sinha, P. (2001) Fidelty of Local Ordinal Encodings. In *Advances in Neural Information Processing Systems* (in press).

Sinha, P. and Poggio, T. (1996) I think I know that face…, *Nature*, **384**, 404.

Sung, K. K., and Poggio, T. (1994) Example based learning for view-based human face detection, AI Laboratory memo # 1521, MIT.

Thoresz, K. and Sinha, P. (2001) Qualitative representations for recognition. *Proceedings of the Annual Meeting of the Vision Sciences Society,* Florida.

Torralba, A. and Sinha, P. (2001). Detecting faces in impoverished images. MIT AI Laboratory Memo, Cambridge, MA.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, IEEE Computer Society Press, Jauai, Hawaii, December 8-14.

Yang, G. and Huang, T. S. (1994) Human face detection in a complex background. *Pattern Recognition*, 27(1) pp. 53-63.