Face Recognition by Humans:
20 Results all Computer Vision Researchers Should Know About

Pawan Sinha[*], *Benjamin Balas, Yuri Ostrovsky, Richard Russell*
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

[*]Correspondence to: <psinha@mit.edu>

A key goal of computer vision researchers is to create automated face recognition systems that can equal, and eventually surpass, human performance. To this end, it is imperative that computational researchers know of the key findings from experimental studies of face recognition by humans. These findings provide insights into the nature of cues that the human visual system relies upon for achieving its impressive performance, and serve as the building blocks for efforts to artificially emulate these abilities. In this paper, we present what we believe are 20 basic results, with direct implications for the design of computational systems. Each result is described briefly, illustrated, and appropriate pointers are provided to permit an in-depth study of any particular result.

Introduction

Notwithstanding the extensive research effort that has gone into computational face recognition algorithms, we are yet to see a system that can be deployed effectively in an unconstrained setting, with all of the attendant variability in imaging parameters such as sensor noise, viewing distance, and illumination. The only system that does seem to work well in the face of these challenges is the human visual system. It makes eminent sense, therefore, to attempt to understand the strategies this biological system employs, as a first step towards eventually translating them into machine-based algorithms. With this objective in mind, we review here twenty important results regarding face recognition by humans. While these observations do not constitute a coherent theory of face recognition in primate vision (we simply do not have all the pieces yet to construct such a theory), they do provide useful hints and constraints for one. We believe that for this reason, they are likely to be useful to computer vision researchers in guiding their ongoing efforts.

We have endeavored to bring together in one place several diverse results, so as to be able to provide the reader a fairly comprehensive picture of our current understanding regarding how humans recognize faces. Each of the results is briefly described and, whenever possible, accompanied by its implications for computer vision. While the descriptions here are not extensive, for reasons of space, we have provided relevant pointers to the literature for a more in-depth study.

Recognition as a function of available spatial resolution

Result 1: Humans can recognize faces in extremely low-resolution images

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. 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? 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, they found high recognition accuracies even with images containing just 16x16 blocks. Yip and Sinha (2002) found that subjects can recognize more than half of an unprimed set of familiar faces with image resolutions of merely 7x10 pixels, and recognition performance reaches ceiling level at a resolution of 19x27 pixels. While the remarkable tolerance of the human visual system to resolution reduction is now indisputable, we do not have a clear idea of exactly how this is accomplished. At the very least, this result demonstrates that fine featural details are not necessary to obtain good face recognition performance. Furthermore, given the indistinctness of the individual features at low resolutions, it appears likely that diagnosticity resides in their overall configuration. However, precisely which aspects of this configuration are important, and how we can computationally encode them, are open questions.

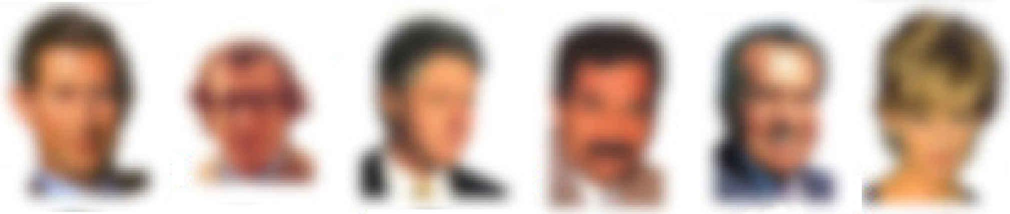
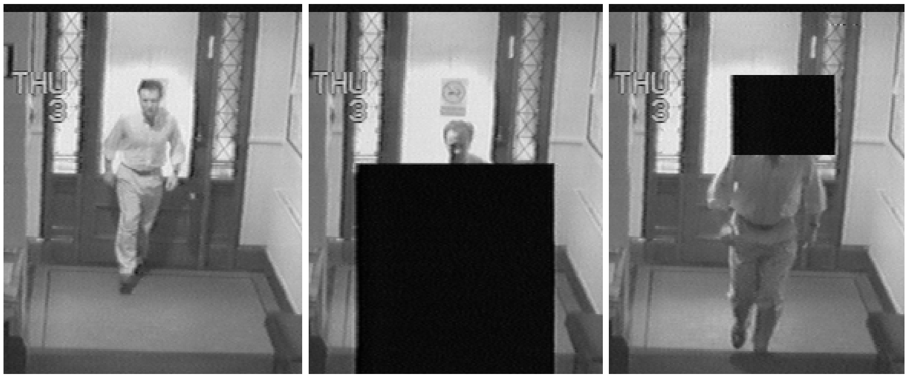


Figure 1. *Unlike current machine based systems, human observers are able to handle significant degradations in face images. For instance, subjects are able to recognize more than half of all familiar faces shown to them at the resolution depicted here. The individuals shown from left to right, are: Prince Charles, Woody Allen, Bill Clinton, Saddam Hussein, Richard Nixon and Princess Diana.*

Result 2: The ability to tolerate degradations increases with familiarity

In trying to uncover the mechanisms underlying the human ability to recognize highly degraded face images, we might wonder whether this is the result of some general purpose compensatory processes, i.e. a biological instantiation of model-free 'super-resolution'. However, the story appears to be more complicated. The ability to handle degradations increases dramatically with amount of familiarity. Burton et al (1999) have shown that observers' recognition performance with low-quality surveillance video is much better when the individuals pictured are familiar colleagues, rather than those with whom the observers have interacted infrequently. Additionally, body structure and gait information are much less useful for identification than facial information, even though the effective resolution in that region is very limited. Recognition performance changes only slightly after obscuring the gait or body, but is affected dramatically when the face is hidden, as illustrated in figure 2. This does not appear to be a skill that can be acquired through general experience; even police officers with extensive forensic experience perform poorly unless they are familiar with the target individuals. The fundamental question this finding, and others like it (Roark et al, 2003; Liu et al, 2003), bring up is the following: How does the facial representation and matching strategy used by the visual system change with increasing familiarity, so as to yield greater tolerance to degradations? We do not yet know exactly what aspect of the increased experience with a given individual leads to an increase in the robustness of the encoding; is it the greater number of views seen or is the robustness an epiphenomenon related to some biological limitations such as slow memory consolidation rates? Notwithstanding our limited understanding, some implications for computer vision are already evident. In considering which aspects of human performance to take as benchmarks, we ought to draw a distinction between familiar and unfamiliar face recognition. The latter may end up being a much more modest goal than the former, and might constitute a false goal towards which to strive. The appropriate benchmark for evaluating machine-based face recognition systems is human performance with familiar faces.



(a) (b) (c)

***Figure 2**. Frames from the video-sequences used in the Burton et al (1999) study. (a) Original input (b) Body obscured (c) Face obscured. Based on results from such manipulations, the researchers concluded that recognition of familiar individuals in low-resolution video is based largely on facial information.*

Result 3: High-frequency information by itself does not lead to good face recognition performance

We have long been enamored of edge-maps as a powerful initial representation for visual inputs. The belief is that edges capture the most important aspects of images (the discontinuities), while being largely invariant to shallow shading gradients that are often the result of illumination variations. In the context of human vision as well, line-drawings appear to be sufficient for recognition purposes. Caricatures and quick pen portraits are often highly recognizable. Do these observations mean that high spatial frequencies are critical, or at least sufficient, for face recognition? There is reason to doubt this assertion. Intuitively, line-drawings appear to contain primarily contour information and very little photometric information over which to define the luminance relations. However, experimental data suggest otherwise. Graham Davies and his colleagues have reported (Davies et al, 1978) that images which contain exclusively contour information are very difficult to recognize (specifically, they found that subjects could recognize only 47% of the line-drawings compared to 90% of the original photographs; see figure 3). How can we reconcile such findings with the observed recognizability of line-drawings in everyday experience? Bruce and colleagues (Bruce and Young, 1998; Bruce et al, 1992) have convincingly argued that such depictions do in fact contain significant photometric cues and that the contours included in such a depiction by an accomplished artist correspond not just to a low-level edge-map, but in fact embody a face's photometric structure. It is the skillful inclusion of these photometric cues that is believed to make human generated line-drawings more recognizable than computer generated ones (Pearson and Robinson, 1985). The idea that 'line-drawings' contain important photometric cues leads to the prediction that recognition performance with line-drawings would be susceptible to contrast negation, just as for gray-scale images. This prediction is indeed supported by experimental data (Pearson, Hanna and Martinez, 1990).

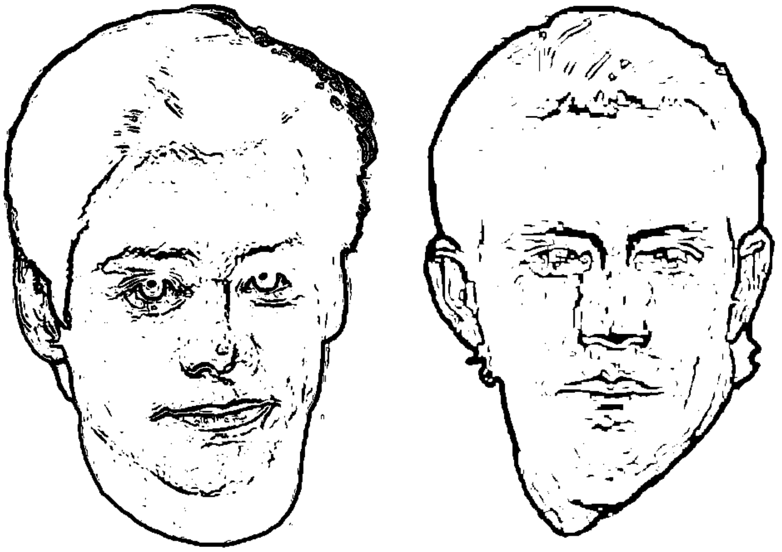


Figure 3. *Images which contain exclusively contour information are very difficult to recognize, suggesting that high-spatial frequency information, by itself, is not an adequate cue for human face recognition processes. Shown here are Jim Carrey (left) and Kevin Costner.*

The nature of processing: Piecemeal versus Holistic

Result 4: Facial features are processed holistically

Can facial features (eyes, nose, mouth, eyebrows, etc.) be processed independently from the rest of the face? Faces can often be identified from very little information. Sadr et al (2003) and others (Davies et al, 1977; Fraser et al, 1990) have shown that just one feature (such as the eyes or, notably, the eyebrows) can be enough for recognition of many famous faces. However, when features on the top half of one face are combined with the bottom half of another face, the two distinct identities are very difficult to recognize (Young et al, 1987) (see Figure 4). The holistic context seems to affect how individual features are processed. When the two halves of the face are misaligned, presumably disrupting normal holistic processing, the two identities are easily recognized. These results suggest that when taken alone, features are sometimes sufficient for facial recognition. In the context of a face, however, the geometric relationship between each feature and the rest of the face can override the diagnosticity of that feature. Although feature processing is important for facial recognition, this pattern of results suggests that configural processing is at least as important, and that facial recognition is dependent on 'holistic' processes involving an interdependency between featural and configural information. Recent work has explored how one might learn to use holistic information (Robbins and McKone, 2003) and the contribution of holistic processing to the analysis of facial expressions (Calder et al, 2000).

Figure 4. *Try to name the famous faces depicted in the two halves of the left image. Now try the right image. Subjects find it much more difficult to perform this task when the halves are aligned (left) compared to misaligned halves (right), presumably because holistic processing interacts (and in this case, interferes) with feature-based processing.*

Result 5: Of the different facial features, eyebrows are amongst the most important for recognition.

Not all facial features are created equal in terms of their role in helping identify a face. Experimental results typically indicate the importance of eyes followed by the mouth and then the nose. However, one facial feature has, surprisingly, received little attention from researchers in this domain – the eyebrows. Sadr et al (2003) have presented striking new evidence suggesting that the eyebrows might not only be important features, but that they might well be the most important, eclipsing even the eyes. These researchers digitally erased the eyebrows from a set of 50 celebrity face images (figure 5). Subjects were shown these images individually and asked to name them. Subsequently, they were asked to recognize the original set of (unaltered) images. Performance was recorded as the proportion of faces a subject was able to recognize. Performance with the images lacking eyebrows was significantly worse relative to that with the originals, and even with the images lacking eyes. These results suggest that the eyebrows may contribute in an important way to the representations underlying identity assessments.

How might one reasonably explain the perceptual significance of eyebrows in face recognition? There are several possibilities. First, eyebrows appear to be very important for conveying emotions and other nonverbal signals. Since the visual system may already be biased to attend to the eyebrows in order to detect and interpret such signals, it may be that this bias also extends to the task of facial identification. Second, for a number of reasons, eyebrows may serve as a very "stable" facial feature. Because they tend to be relatively high-contrast and large facial features, eyebrows can survive substantial image degradations. For instance, when faces are viewed at a distance, the eyebrows continue to make an important contribution to the geometric and photometric structure of the observed image. Also, since eyebrows sit atop a convexity (the brow ridge separating the forehead and orbit), as compared to some other parts of the face, they may be less susceptible to shadow and illumination changes. Further, although the eyebrows can undergo a wide range of movements, the corresponding variations in the appearance of the eyebrows themselves do not rival those observed within the eyes and mouth, for example, as they run through the gamut of their own movements and deformations.

Figure 5. *Sample stimuli from Sadr et al's (2003) experiment assessing the contribution of eyebrows to face recognition: original images of President Richard M. Nixon and actor Winona Ryder, along with modified versions lacking either eyebrows or eyes.*

Result 6: Both internal and external facial cues are important and they exhibit non-linear interactions

A marked disparity exists in the use of 'internal' and 'external' facial features by current machine-based face analysis systems. It is typically assumed that internal features (eyes, nose and mouth), and their mutual spatial configuration, 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. Some recent experiments from our lab have investigated the contribution of internal and external features as a function of effective image resolution. The experimental paradigm we used required subjects to recognize celebrity facial images blurred by varying amounts (a sample set is shown in figure 1). 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 were tested on the three conditions.

Performance of subjects with internal features (whether side by side, or in the correct spatial configuration) is quite poor even with relatively small amounts of blur (see figure 6(a)). Performance with external features, by themselves, is also poor, not exceeding 40% at any resolution level. However, performance with full heads is impressively robust to degradation, and remains high even when the independent contributions from the internal and external feature sets are close to zero, suggesting a highly non-linear interaction between the two sets of cues. Overall, the results demonstrate the insufficiency of internal features and even their mutual configuration, while highlighting the perceptual importance of the full head configuration for face recognition. Figure 6(b) shows an image that underscores the importance of overall head shape in determining identity.

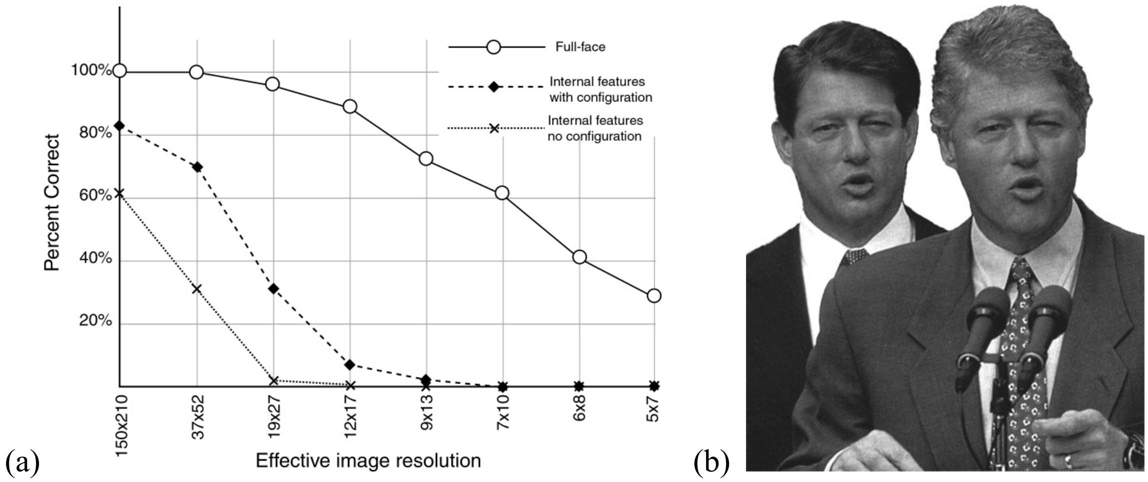


Figure 6. *(a) Recognition performance with internal features (with and without configural cues). Performance obtained with whole head images is also included for comparison.*

(b) 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)

Result 7: The important configural relationships appear to be independent across the width and height dimensions

Taking up where the previous result left off, we can ask what aspects of the spatial structure of a head are important for judgments of identity? At least a few computer vision systems involve precise measurements of attributes such as the inter-eye distance, width of the mouth and length of nose. However, it appears that the human visual system does not depend critically on these measurements. Evidence in favor of this claim comes from investigations of recognition with distorted face images. In our work, we have found a remarkable tolerance of recognition processes to compressive distortions. A face can be compressed down to 25% of its original height or width, with absolutely no loss in its recognizability (see figure 7). Clearly, such compressions play havoc with absolute inter-feature distance measurements, and also distance ratios across the x and y dimensions. Nevertheless, recognition performance stays invariant. One set of spatial attributes that stay unchanged with compressions, are ratios of distances within the same dimension. It is possible then that human encoding of faces utilizes such ratios (we refer to them as iso-dimension ratios), and this might constitute a useful strategy for computer vision systems as well. Why might the human visual system have adopted such a strategy, given that image compressions were not particularly commonplace until the recent advent of photography? To a limited extent, rotations in depth around the x and y axes approximate 2-D compressions. Perhaps the human visual system has adopted an iso-dimension ratio encoding strategy to obtain a measure of tolerance to such transformations.

Figure 7. *Even drastic compressions of faces do not render them unrecognizable. Here, the celebrity faces have been compressed to 25% of their original width. Yet, recognition performance with this set is the same as that obtained with the original faces.*

Result 8: Vertical inversion dramatically reduces recognition performance

Upside down ('inverted') faces are harder to recognize than right-side up faces, despite the fact that the same information is present in both images. Yin (1970) trained adults on a series of faces, which later had to be identified from pairs made up of seen and unseen faces. Performance in the test phase was high (90 percent) when these faces were presented upright but suffered remarkably (62 percent) when all faces were inverted. The difference in performance was much smaller (10 percentage points) when houses were used instead of faces in the two conditions, suggesting that this is not a characteristic of general object recognition but may be face-specific. The dominant explanation for the decrement in face recognition performance, induced by vertical inversion, is that this transformation selectively impairs our ability to extract configural information from faces, while leaving featural processing largely intact. Partial support for this assertion comes from experiments showing that while faces differing in individual features (such as eyes and mouth) can be readily distinguished even when vertically inverted, configurally different faces are much harder to tell apart upon inversion. The predominantly featural style of analysis with inverted faces is illustrated by the 'Thatcher illusion' (see figure 8) (Thompson, 1980). However, this notion of a clear separation of featural and configural analyses has been challenged by some recent experimental evidence (Riesenhuber et al, 2004). In the light of these experimental data, while we cannot say precisely what accounts for the difficulty in recognizing inverted faces, we can be sure that vertical inversion has a dramatic adverse effect on human performance, and therein may lie clues regarding the nature of face encoding strategies used by the visual system. For machine based systems, inverted faces might be as easy as upright ones assuming the existence of a prior step of normalization. It is interesting that such a 'normalization', while cognitively feasible, does not appear to help human performance.

Figure 8. *The Thatcher Illusion. The eyes and mouth of the image on the right have been vertically inverted. When the whole face is inverted as well, this manipulation is not apparent. If the reader turns this page around, however, the manipulation is grotesquely obvious.*

The nature of cues used: Pigmentation and shape

Result 9: Face-shape appears to be encoded in a slightly caricatured manner

Intuitively, successful face recognition requires that the human visual system should encode previously seen faces veridically. Errors in the stored representation of a face obviously weaken the potential to match new inputs to old.

However, it has been demonstrated that some departures from veridicality are actually beneficial for human face recognition. Specifically, "caricatured" versions of faces have been demonstrated to support recognition performance at least equal to or better than that achieved with veridical faces (Rhodes, 1996). Caricatured faces can be created to exaggerate deviations in shape alone (Brennan, 1985) or a combination of deviations in both shape and pigmentation cues (Benson & Perrett, 1991). In both cases, subjects display small, but consistent, preferences for caricatured faces as determined by several different measures (Lee & Perrett, 1997, 2000). Shape caricaturing is evident for objects other than faces as well (Gibson, 1947) suggesting that caricatured representations may be a widely applied strategy.

These results have been taken to suggest a norm-based representational space for faces, often referred to in the literature as "face space" (Valentine, 1999). This hypothesis may usefully constrain the kinds of encoding strategies employed by computational face recognition systems. At the very least, an important test for any recognition scheme is whether or not it displays "caricature effects" similar to those found in human recognition.

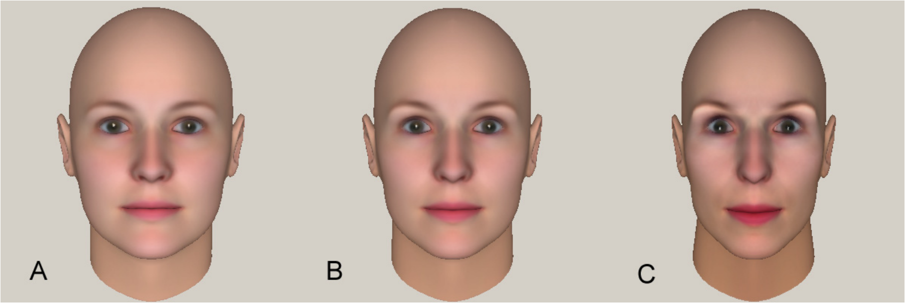


Figure 9. *An example of a face caricature. The average female face for a particular face population is displayed (A), as well as a "veridical" image of an exemplar face (B). We create a caricatured version of the exemplar by moving away from the norm, thus exaggerating differences between the average face and the exemplar. The result is a face with "caricatured" shape and pigmentation (C). Such caricatures are recognized as well or better than veridical images.*

Result 10: Pigmentation cues are at least as important as shape cues

There are two basic ways in which faces can differ—in terms of their shape, and in terms of how they reflect light, or their pigmentation. By 'pigmentation', we refer to all surface reflectance properties, including albedo, hue, specularity, translucency, and spatial variation in these properties. When referring to *all* surface reflectance properties of faces, we prefer the term 'pigmentation' to the terms 'texture' or 'color', which invite confusion because they are commonly used to refer to specific subsets of surface reflectance properties (spatial variation in albedo and greater reflectance of particular wavelengths, respectively).

Recent studies have investigated whether shape or pigmentation cues are more important for face recognition. The approach taken has been to create sets of faces that differ from one another in terms of only their shape or only their pigmentation, using either laser-scanned models of faces (pictured in Figure 10) (O'Toole et al 1999), artificial faces (Russell et al 2004), or morphing photographs of faces (in which case shape is defined in terms of the 2-dimensional outlines of the face and individual features) (Russell et al 2004). With each of these classes of stimuli, subjects have performed about equally well using either shape or pigmentation cues. This provides evidence that the two kinds of cues are used about equally by humans to recognize faces. A current study in our laboratory investigating the use of these cues for the recognition of familiar faces is also finding that both shape and pigmentation are about equally important. An implication of this work is that artificial face recognition systems would benefit from representing pigmentation as well as shape cues.

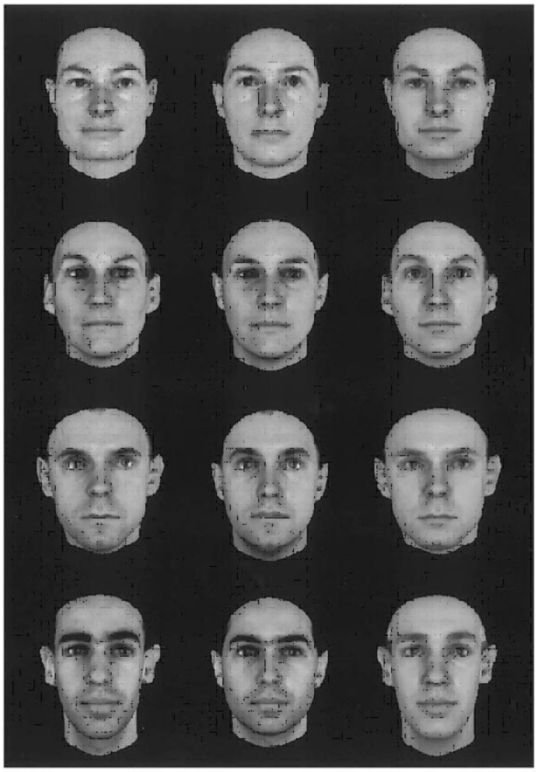


Figure 10. *Stimuli from O'Toole et al. 1999. The faces in the left column are all images of laser-scanned faces. They differ from one another in terms of both shape and pigmentation. The faces in the center column differ from one another in terms of their pigmentation but not their shape, while the faces in the right column differ from one another in terms of their shape but not their pigmentation. From the fact that the faces in either the center or right column do not look the same as each other, it is evident that both shape and pigmentation cues play a role in facial identity.*

Result 11: Color cues play a significant role especially when shape cues are degraded

The luminance structure of face images is undoubtedly of great significance for recognition. Past research has suggested that the use of these cues may adequately account for face-identification performance with little remaining need to posit a role for color information. Furthermore, people tend to accurately identify faces that are artificially colored. However, recent evidence (Yip and Sinha, 2002) counters the notion that color is unimportant for human face recognition, and suggests instead that when shape cues in images are compromised (say, by reductions in resolution), the brain relies on color cues to pinpoint identity. In such circumstances, recognition performance with color images is significantly better than with gray-scale images. Precisely how does color information facilitate face recognition? One possibility is that color provides diagnostic information. The expression 'diagnostic information' refers to color cues that are specific to an individual, for instance the particular hue of their hair or skin that may allow us to identify them. On the other hand, color might facilitate low-level image analysis, and thus indirectly aid face recognition. An example of such a low-level task is image segmentation – determining where one region ends and the other starts. As many years of work in computer vision has shown (Haralick, 1985; Felzenszwalb and Huttenlocher, 1998), this task is notoriously difficult and becomes even more intractable as images are degraded. Color may facilitate this task by supplementing the luminance-based cues and thereby lead to a better parsing of a degraded face image in terms of its constituent regions. Experimental data favor the second possibility. Recognition performance with pseudo-colored face images (which do not contain diagnostic hue information) is just as high as with natural color images (and both are significantly better than grayscale images, when shape cues are degraded). Figure 11 illustrates this idea. The images show the luminance and color components of sample face inputs. They suggest that color distributions can supplement luminance information to allow for a better estimation of the boundaries, shapes and sizes of facial attributes such as eyes and hair-lines.

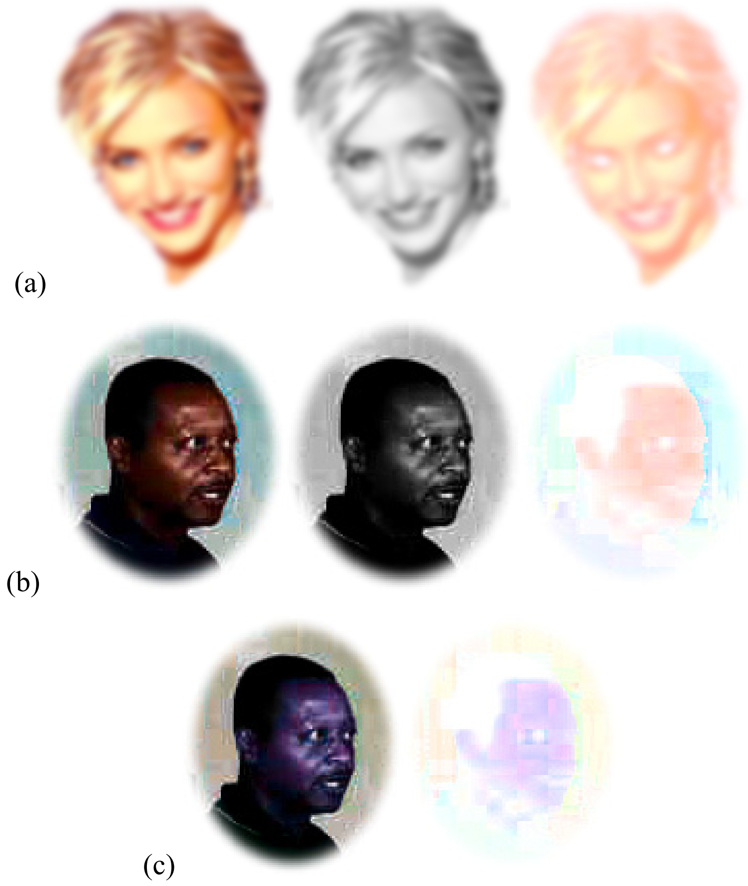


Figure 11. *Examples that illustrate how color information may facilitate some important low-level image analysis tasks such as segmentation. In (a), the hue distribution (right panel) allows for a better estimation of the shape and size of the eyes than the luminance information alone (middle panel). Left panel shows the original image. Similarly, in (b), hue information (right panel) allows for a better segmentation and estimation of the location and shape of the hair-line than just the luminance information (middle panel). This facilitation of low-level analysis happens with other choices of colors as well, such as in the pseudo-color image shown on the left in (c). The hue distribution here, as in (b) aids in estimating the position of facial attributes such as the hair-line.*

Result 12: Contrast polarity inversion dramatically impairs recognition performance, possibly due to compromised ability to use pigmentation cues

Skilled darkroom technicians working in the photo re-touching industry several decades ago noticed that faces were particularly difficult to recognize when viewed in reversed contrast, as in photographic negatives. Subsequently the phenomenon has been studied extensively in the vision science community, with the belief that determining how recognition can be impaired helps us understand how it works under normal conditions. Contrast negation is a reversible manipulation that does not remove any information from the image. Though no information is lost, our *ability* to use the information in the image is severely compromised. This suggests that some normally useful information is rendered unusable by negation.

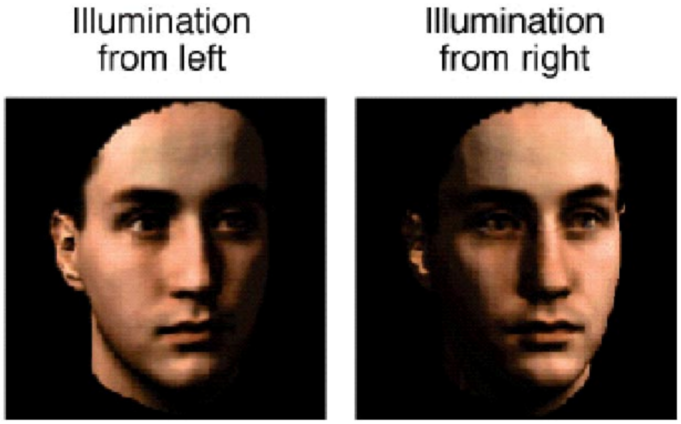
When pigmentation cues are unavailable, as in uniformly pigmented 3-dimensional face models (derived from laser scans) or in other stimuli for which pigmentation cues are unavailable (see Result 10 for examples), recognition is not significantly worse with negative contrast (Bruce and Langton 1994; Russell et al (Under review)). This suggests that pigmentation cues might be disrupted by negation. Other work with uniformly pigmented face models has found evidence that shading cues are disrupted by contrast negation, but only for faces lit from above (Liu et al 1999). These findings suggest that human face recognition uses representations that are sensitive to contrast direction, and that pigmentation and shading play important roles in recognition.

Figure 12. *A selection from the cover of the Beatles' "Sgt. Pepper Lonely Hearts Club Band" album, presented in negative contrast negative. This image contains numerous well-known celebrities, whose likenesses would be easily recognizable to many readers of this publication. However, when presented in negative contrast, it is difficult, if not impossible, to recognize most of the faces.*

Result 13: Illumination changes influence generalization

Some computational models of recognition assume that a face must be viewed under many different illumination conditions for robust representations. However, there is evidence that humans are capable of generalizing representations of a face to radically novel illumination conditions. In one recent study (Braje et al 1998), subjects shown a laser scanned image of a face with illumination coming from one side, were subsequently shown a face illuminated strongly from the other side, and were asked whether both images were of the same face. Subjects were well above chance at deciding whether the second face was the same as the first, indicating significant ability to generalize the representation of the face to novel illumination conditions. However, the subjects were significantly impaired at this task relative to when the two faces were presented under the same illumination, indicating that the generalization to novel illumination conditions is not perfect.

An implications of this result is that human recognition of faces is sensitive to illumination direction, but is capable of significant generalization to novel illumination conditions even after viewing only a single image.



Illumination from left

Illumination from right

Figure 13. *Stimuli from Braje et al. (1998). These two images demonstrate the kind of lighting used in this experiment. After being shown an image like the one on the left, subjects were well above chance at determining whether a subsequently presented image such as the one on the right represented the same or a different individual (in this case the same).*

Role of temporal cues

Result 14: View-generalization appears to be mediated by temporal association

Recognizing a face across variations in viewing angle is a very challenging computational task that the human visual system can solve with remarkable ease. Despite the fact that image-level differences between two views of the same face are much larger than those between two different faces viewed at the same angle (Moses, Adini, & Ullman, 1994), human observers are somehow able to link the correct images together.

It has been suggested that temporal association serves as the “perceptual glue” that binds different images of the same object into a useful whole. Indeed, close temporal association of novel images viewed in sequence is sufficient to induce some IT neurons to respond similarly to arbitrary image pairs (Miyashita, 1993). Behavioral evidence from human observers exposed to rotating "paperclip" objects supports rapid learning of image sequences as well (Sinha & Poggio, 1996).

In terms of human face recognition, temporal association of two unique faces (one frontally viewed, the other viewed in profile) has been demonstrated to have intriguing consequences for recognition. Brief exposure to movies containing a rotating head which morphs between one individual and another as it rotates from frontal to profile views can impair observers' ability to distinguish between the two faces contained in the sequence (Wallis & Bulthoff, 2001).

Taken together, these results suggest that the temporal proximity of images is a powerful tool for establishing object representations. Studying recognition performance using images that lack a temporal context may be a profound handicap to our understanding of how view invariance is achieved. Exploring image sequences using mechanisms that make explicit temporal associations (Foldiak, 1991) may be a powerful means for view generalization.

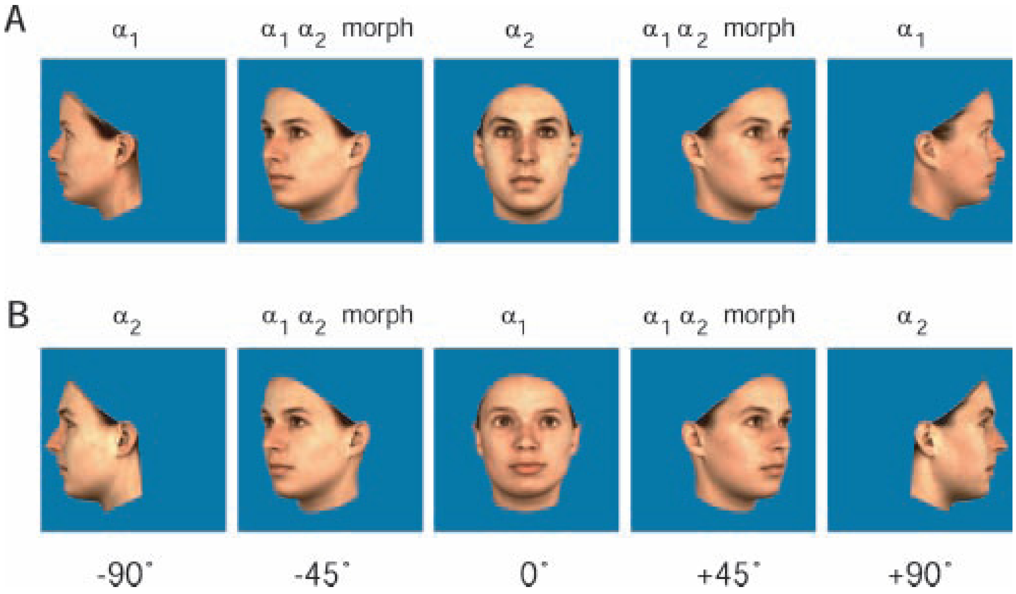


Figure 14. *The time course of the sequences shown to observers in Wallis & Bulthoff (2001). Faces α1 and α2 are each used as the frontally viewed face in separate sequences, and combined with the other face profile in their respective movies. ¾ morphs between α1 and α2 are used to interpolate between the frontally-viewed faces and the profiles to create a smooth motion sequence. Same/Different performance for faces appearing in the same sequence is impaired relative to pairs of faces appearing in different sequences.*

Result 15: Motion of faces appears to facilitate subsequent recognition

Do dynamic cues aid face recognition? The answer is 'yes', but only in some cases. Rigid motion (such as that obtained from a camera rotating around a motionless head) can facilitate recognition of previously viewed faces (Schiff et al, 1986; O'Toole et al, 2002) but there seems to be very little, if any, benefit of seeing these views during the learning phase. By contrast, non-rigid motion (where the individuals exhibit emotive facial expressions or speech movements) plays a greater role. Experiments in (Knappmeyer et al, 2003), using subtle morphs of form and facial motion in novel (i.e., unfamiliar) faces, showed that non-rigid facial motion from one face applied to the form of another face can bias an observer to misidentify the latter as the former (see Figure 15). Experiments with famous (i.e., highly familiar) faces (Lander and Chuang, 2005) again showed a facilitation in recognition with dynamic cues from expressive or talking movements, but not from rigid motion. Facilitation was most pronounced for faces whose movement was judged as 'distinctive'. Note also that facilitation comes from a natural sequence of moving images, not merely from having more views available: The facilitation is greatly lessened when the same frames are presented in random order or in a static array.

These results suggest that face motion is more than just a sequence of viewpoints to the face recognition system. The dynamic cues from expressive and talking movements provide information about aspects of facial structure that transcend the gains of simply having multiple viewpoints.

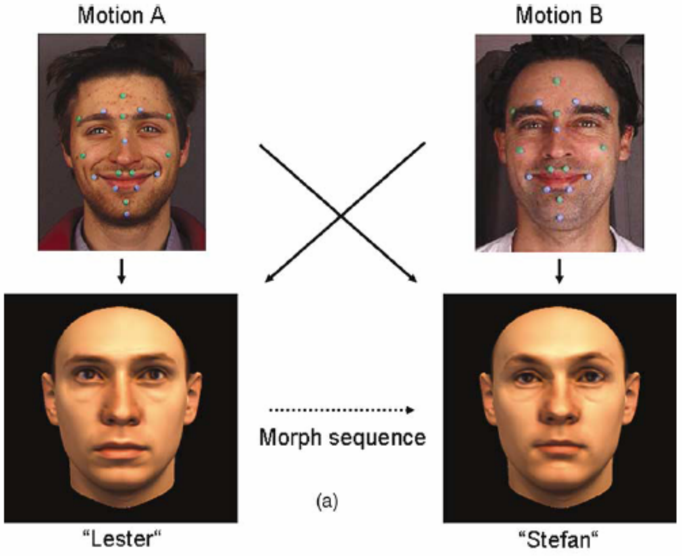


Figure 15. *Facial motion from expressions and talking were morphed onto the forms of 'Lester' and 'Stefan'. Subjects could be biased to identify an anti-caricatured (morphed towards the average) form of Lester as Stefan when Stefan's movements were imposed onto Lester's form. (From Knappmeyer et al, 2003.)*

Developmental progression

Result 16: The visual system starts with a rudimentary preference for face-like patterns

What, if any, are the face-specific biases that the human visual system starts out with? Newborns selectively gaze at 'face-like' patterns only hours after birth. A pattern that is face-like can be something as simple as that shown in figure 16(a): three dots within an oval that represent the two eyes and a mouth. An impossible face (created by vertically inverting the triad of dots) does not attract the newborn's attention as much as the more normal face. However, the specificity of the response to the three-dot arrangement has been called into question. More recent work (Simion et al, 2001) suggests that newborns simply prefer ‘top-heaviness’ (figure 16(b)). Thus, it remains unclear whether this is a general preference (perhaps with no practical significance) or a face-specific orienting response to prime the infant in bootstrapping its nascent face recognition system. Even if this preference really is an innate face-orienting mechanism, it may be more for the benefit of the mother (e.g., to form the mother-child bond) than the infant’s face processing capabilities.

A simple arrangement of three dots within an oval may serve as an appropriate template for detecting faces in the bootstrapping stages of a face-learning system. Similar templates have been used with reasonable success in some applications (for example, Sinha, 2002) of face detection.

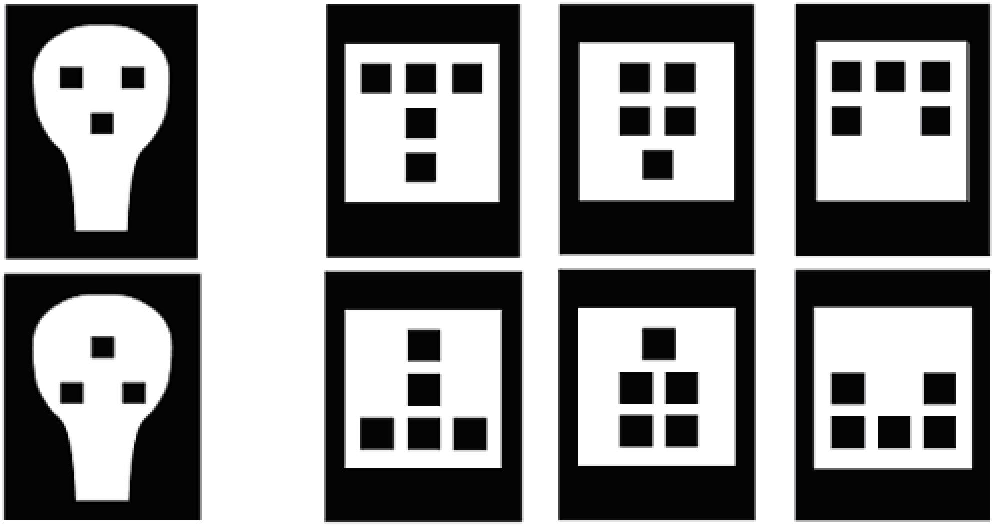
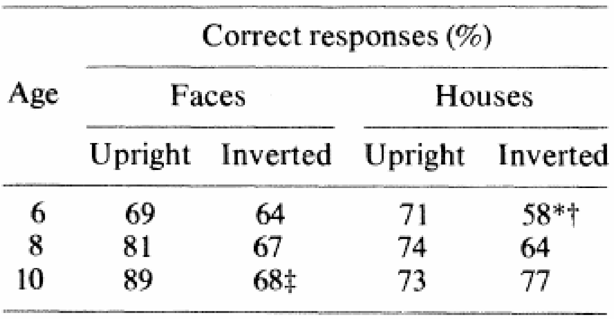

(a) (b)

Figure 16 (a). *Newborns preferentially orient their gaze to the face-like pattern on the left, rather than the one shown on the right, suggesting some innately specified representation for faces. (From Johnson et al, 1991.) (b) As a counterpoint to the idea of innate preferences for faces, Simion et al (2001) have shown that newborns consistently prefer top-heavy patterns (left column) over bottom-heavy ones (right column). It is unclear whether this is the same preference exhibited in earlier work, and if it is, whether it is face-specific or some other general-purpose or artifactual preference.*

Result 17: The visual system progresses from a piecemeal to a holistic strategy over the first several years of life

As discussed in Result 8, normal adults show a remarkable deficit in recognition of inverted faces, but no such deficit for inverted images of non-face objects such as houses. A number of studies have shown, however, that this pattern of results takes many years to develop (Carey and Diamond, 1977; Hay and Cox, 2000; Maurer et al, 2002; Mondloch et al, 2002, 2003; Pellicano and Rhodes, 2003; Schwarzer, 2003). Six year old children are *not* affected by inversion when it comes to recognizing seen faces in a seen-unseen pair [16]; 8 year olds show some inversion effect and 10 year olds exhibit near adult-like performance (see Fig. 17). Experimenters in (Mondloch et al, 2002) selectively manipulated spacing (moving the location of features on a face) versus features (taking eyes or mouth from different faces) and found that it is specifically sensitivity to *spacing* manipulations that is impaired when faces are inverted. Interestingly, although six year old children are not sensitive to inversion in the tests mentioned above, they *are* susceptible to the Thatcher Illusion (Thompson, 1980; Lewis, 2003), suggesting, perhaps, that configural information is important for face *processing* throughout child development, but that this information has not been assimilated into the face *recognition* system.

This pattern of behavior suggests that over the course of several years, a shift in strategy occurs. Initially, infants and toddlers adopt a largely piecemeal, feature-based strategy for recognizing faces. Gradually, a more sophisticated holistic strategy involving configural information evolves. This is indirect evidence for the role of configural information in achieving the robust face recognition performance that adults exhibit.



Age	Correct responses (%)			
	Faces		Houses	
	Upright	Inverted	Upright	Inverted
6	69	64	71	58*†
8	81	67	74	64
10	89	68‡	73	77

Figure 17. *Generally, six year olds are rather poor at upright and inverted faces. As their age approaches ten years, their performance improves dramatically on upright faces, but hardly any improvement is exhibited on inverted faces. From Carey and Diamond, 1971.*

Neural underpinnings

Result 18: The human visual system appears to devote specialized neural resources for face perception

Whether or not faces constitute a "special" class of visual stimuli has been the subject of much debate for many years. Since the first demonstrations of the "inversion effect" described above (Yin, 1969), it has been suspected that unique cognitive and neural mechanisms may exist for face processing in the human visual system.

Indeed, there is a great deal of evidence that the primary locus for human face processing may be found on the fusiform gyrus of the extra-striate visual cortex (Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997). This region shows an intriguing pattern of selectivity (schematic faces do not give rise to much activity) and generality (animal faces do elicit a good response) (Tong, Nakayama, Moscovitch, Weinrib, & Kanwisher, 2000), suggesting a strong domain-specific response for faces. In keeping with behavioral results, the "fusiform face area" (FFA) also appears to exhibit an "inversion effect" (Kanwisher, Tong, & Nakayama, 1998). Overall, the characterization of the FFA as a dedicated face processing module appears very strong.

However, it must be noted that the debate over faces being "special" is far from over. It has been suggested that rather than being a true "face module," the FFA may be responsible for performing either subordinate or "expert-level" categorization of generic objects. There are results from both behavioral studies (Diamond & Carey, 1986;

Gauthier & Tarr, 1997) and neuroimaging studies (Gauthier, Anderson, Tarr, Skudlarski, & Gore, 1997) that lend some support to this "perceptual expertise" account. Recent findings appear to favor the original "face module" account of the FFA's function, however (Grill-Spector, Knouf, & Kanwisher, 2004).

The full breadth and depth of the arguments supporting both positions are beyond the scope of this review (see (McKone & Kanwisher, 2005) for a more thorough treatment), but it is important to recognize that specialized face processing mechanisms in the human visual system are a very real possibility. Whatever its ultimate status, the response profile of the FFA provides a potentially valuable set of constraints for computational systems, indicating the extent of selectivity and generality we should expect from face recognition systems.

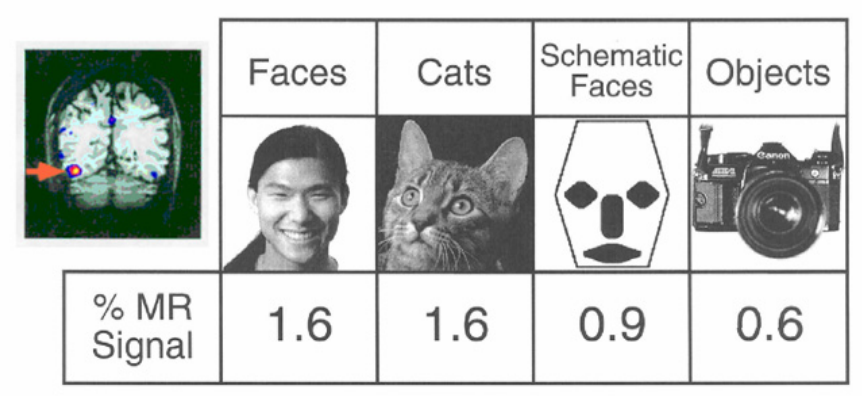


Figure 18. *At upper left, an example of the FFA in one subject, showing right-hemisphere lateralization. Also included here are example stimuli from Tong et al. 2000, together with the amount of percent signal change observed in the FFA for each type of image. Photographs of human and animal faces elicit strong responses, while schematic faces and objects do not. This response profile helps place constraints on the selectivity and generality we might expect from computational models of human face recognition.*

Result 19: Latency of responses to faces in IT cortex is about 120 ms, suggesting a largely feed-forward computation

Human observers can carry out visual recognition tasks very rapidly. Behavioral reaction times (RTs) are already quite fast, and represent a potentially large overestimate of the time required for recognition due to the motor component of signaling a response. Indeed, when a neural marker of recognition is used, accurate performance on such seemingly complex tasks as determining the presence/absence of an animal in a natural scene appears to require as little as 50ms (Thorpe, Fize, & Marlot, 1996).

Recently it has been shown that although this particular task (animal/no animal) seems quite complicated, it may be solvable using very low-level visual representations (Johnson & Olshausen, 2003). That said, there is neurophysiological evidence that truly complex tasks, such as face recognition, may be carried out over a surprisingly short period of time.

Neurons in primate inferotemporal (IT) cortex can exhibit selectivity to stimuli that are more complicated than the simple gratings and bars that elicit responses from cells in early visual areas. In particular, it has been noted that there are some cells in IT cortex that are selective for faces (Desimone, Albright, Gross, & Bruce, 1984). Moreover, the latency of response in these cells is in the neighborhood of 80-160ms (Perrett, Rolls, & Caan, 1982). Although slightly longer than the 50ms “super-RT” reported in the animal/no-animal task, such rapid emergence of a selective response indicates that face recognition may indeed proceed very rapidly.

The computational relevance of these results is that recognition as it is performed up to the level of IT cortex probably requires only one feed-forward pass through the visual system. Feedback and iterative processing are likely not major factors in the responses recorded in these studies. This is a very important constraint on recognition algorithms, as it indicates that sufficient information must be extracted immediately from the image without the luxury of resorting to slowly converging iterative computations.

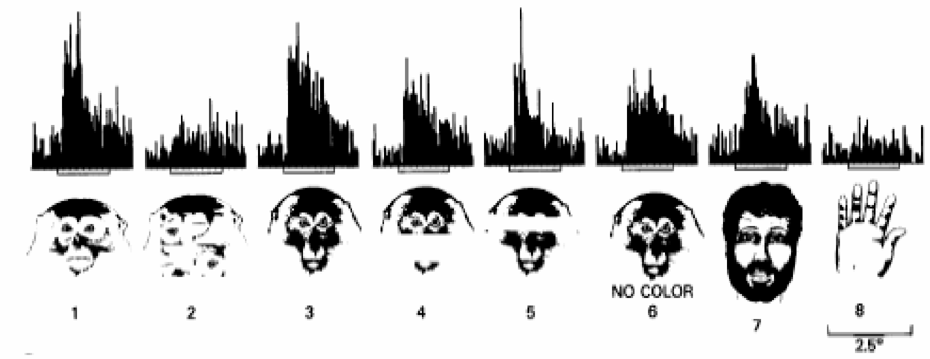


Figure 19. *Examples of an IT cell’s responses to variations on a face stimulus (from Desimone et al. 1984). The response is robust to many degradations of the primate face (save for scrambling) and also responds very well to a human face. The lack of a response to the hand indicates that this cell is not just interested in body parts, but is specific to faces. Cells in IT cortex can produce responses such as these with a latency of about 120ms.*

Limitations of human performance

Result 20: Human memory for briefly seen faces (as in eyewitness testimony) is rather poor

A central finding from the study of human memory is that people are generally better at recognizing something that they have seen then at recalling it when cued. For example, if given a set of words to study, people are better at recognizing whether presented words were a part of the studied set than at recalling specific items from the set. Most of the experiments described in this review deal with the problem of face recognition. Face recall is much more difficult, and the problem is compounded by the typically poor ability of individuals to externalize an image. However, the quality of face recall has important consequences in the context of eyewitness testimony in criminal cases (Sporer et al, 1996).

Recollection of criminal faces is made particularly difficult by several factors related to criminal activity. Most importantly, crime victims are typically in a highly aroused emotional state during the crime, and are unable to dispassionately study the details of the face of the perpetrator. Related to this emotional issue is the phenomenon of "weapon focus", wherein victims are much more aware of weapons being used by the perpetrator than the actual perpetrator (illustrated in figure 20). These problems with eyewitness recall of faces illustrate one key way in which human recognition of faces differs from that of most computational systems: for humans, face recognition is conducted in the context of ecological priorities, goals, and emotions that have large effects on the saliency of a given face and its context, and hence on subsequent ability to recognize or recall it. Machine based systems, by being immune to such modulatory influences, stand a good chance of exceeding human performance under stressful circumstances.

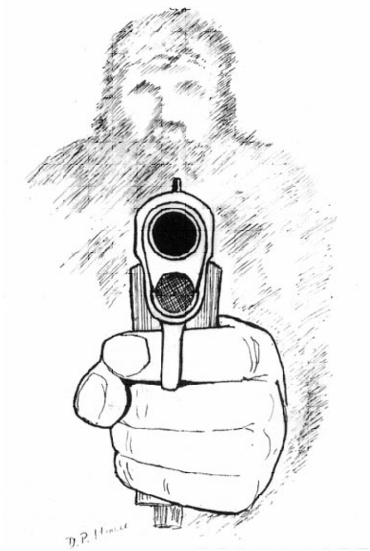


Figure 20. *Figure reprinted from Hinkle (1989). This is a representation of what crime victims often perceive—the threatening weapon in vivid detail, but only a vague sense of the details of the perpetrator. This phenomenon is called "weapon focus".*

Conclusions

The twin enterprises of visual neuroscience and computer vision have deeply synergistic objectives. An understanding of human visual processes involved in face recognition can facilitate and, in turn be facilitated by, better computational models. Our presentation of results in this paper is driven by the goal of furthering cross-talk between the two disciplines. The observations included here constitute twenty brief vignettes into what is surely a most impressive and rather complex biological system. We hope that these vignettes will help in the ongoing computer vision initiatives to create face recognition systems that can match, and eventually exceed, the capabilities of their human counterparts.

Bibliography

Benson, P. J., & Perrett, D. I. (1991). Perception and Recognition of Photographic Quality Facial Caricatures: Implications for the Recognition of Natural Images. *European Journal of Cognitive Psychology,* **3**(1), 105-135.

Braje W L, Kersten D, Tarr M J, Troje N F, (1998). “Illumination Effects in Face Recognition” *Psychobiology,* **26** 371-380.

Brennan, S. E. (1985). The caricature generator. *Leonardo,* **18**, 170-178.

Bruce, V. and Young, A. W. (1998). '*In the eye of the beholder*', Oxford Univ. Press.

Bruce, V., Hanna, E., Dench, N., Healey, P., & Burton, M. (1992). The importance of ‘mass’ in line drawings of faces. *Applied Cogniti6e Psychology*, **6**, 619–628.

Bruce V, Langton S, (1994). “The use of pigmentation and shading information in recognizing the sex and identities of faces” *Perception* **23** 803-22.

Burton, A. M., Wilson, S., Cowan, M. & Bruce, V. (1999). Face recognition in poor-quality video, *Psychological Science*, **10**, 243-248.

Calder, A. J., Young, A. W., Keane, J. and Dean, M. (2000). Configural information in facial expression perception, *J Exp Psychol Hum Percept Perform*, vol. **26**, pp. 527-51.

S. Carey and R. Diamond, (1977). "From Piecemeal to Configurational Representation of Faces," *Science*, vol. **195**, pp. 312-314.

Davies, G., Ellis, H., and Shepherd, J. (1977) Cue saliency in faces as assessed by the 'Photofit' technique. *Perception*, **6**, 263-269.

Davies, G., Ellis, H. D., & Shepherd, J. (1978). Face recognition accuracy as a function of mode of representation. *Journal of Applied Psychology*, **63**, 180–187.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience,* **4**(8), 2051-2062.

Diamond, R., & Carey, S. (1986). Why Faces Are and Are Not Special: An Effect of Expertise. *Journal of Experimental Psychology: General,* **115**(2), 107-117.

Felzenszwalb, P., and Huttenlocher, D. (1998). Image Segmentation Using Local Variation. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 98-104.

Foldiak, P. (1991). Learning Invariance from Transformation Sequences. *Neural Computation,* **3**, 194-200.

Fraser, I. H., Craig, G. L., and Parker, D. M. (1990) Reaction time measures of feature saliency in schematic faces. *Perception*, **19**(5), 661-673.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring the face recognition mechanism. *Vision Research,* **37**(12), 1673-1682.

Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., & Gore, J. C. (1997). Levels of categorization in visual objects studied with functional MRI. *Current Biology,* **7**, 645-651.

Gibson, J. J. (1947). *Motion picture testing and research* (7). Washington D.C.: AAF Aviation Psychology Program.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience,* **7**(5), 555-562.

Haralick, R. (1985). Survey, image segmentation techniques, *Computer Vision, Graphics, and Image Processing,* **29**, 100-135.

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.

Hay, D. C. and Cox, R. (2000). "Developmental changes in the recognition of faces and facial features," *Infant and Child Development*, vol. **9**, pp. 199-212.

Hinkle, D P, (1989). "Faces of Crime", (Atlanta, GA: Peachtree Publishers)

Johnson, M. H., Dziurawiec, S., Ellis, H. and Morton, J. (1991). "Newborns Preferential Tracking of Face-Like Stimuli and Its Subsequent Decline," *Cognition*, vol. **40**, pp. 1-19.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of Neural Signatures of Object Recognition. *Journal of Vision,* **3**, 499-512.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for the Perception of Faces. *Journal of Neuroscience,* **17**, 4302-4311.

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The Effect of Face Inversion on the Human Fusiform Face Area. *Cognition,* **68**, B1-B11.

Knappmeyer, B., Thornton, I. M. and Bulthoff, H. H. (2003). "The use of facial motion and facial form during the processing of identity," *Vision Res*, vol. **43**, pp. 1921-36.

Lander, K. and Chuang, L. (2005). "Why are moving faces easier to recognize?," *Visual Cognition*, vol. **12**, pp. 429-442.

Lee, K. J., & Perrett, D. (1997). Presentation-time measures of the effects of manipulations in colour space on discrimination of famous faces. *Perception,* **26**, 733-752.

Lee, K. J., & Perrett, D. I. (2000). Manipulation of colour and shape information and its consequence upon recognition and best-likeness judgments. *Perception,* **29**, 1291-1312.

Lewis, M. B. (2003). "Thatcher's children: development and the Thatcher illusion," *Perception*, vol. **32**, pp. 1415-21.

Liu, C. H., Collin, C. A., Burton, A. M., Chaurdhuri, A., (1999). "Lighting direction affects recognition of untextured faces in photographic positive and negative" *Vision Research* **39,** 4003-4009.

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

Maurer, D., Le Grand, R. and Mondloch, C. J. (2002). "The many faces of configural processing," *Trends in Cognitive Sciences*, vol. **6**, pp. 255-260.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience,* **9**, 605-610.

McKone, E., & Kanwisher, N. (2005). Does the human brain process objects of expertise like faces? A review of the evidence. In S. Dehaene, J. R. Duhamel, M. Hauser, & G. Rizzolatti (Eds.), *From Monkey Brain to Human Brain*. Cambridge, MA: MIT Press.

Miyashita, Y. (1993). Inferior temporal cortex: where visual perception meets memory**.** *Annual Reviews of Neuroscience,* **16**, 245-263.

Mondloch, C. J., Le Grand, R. and Maurer, D. (2002). "Configural face processing develops more slowly than featural face processing," *Perception*, vol. **31**, pp. 553-566.

Mondloch, C. J., Geldart, S., Maurer, D. and Le Grand, R. (2003). "Developmental changes in face processing skills," *Journal of Experimental Child Psychology*, vol. **86**, pp. 67-84.

Moses, Y., Adini, Y., & Ullman, S. (1994). Face recognition: the problem of compensating for illumination changes. *Proceedings of the European Conference on Computer Vision*, 286-296.

O'Toole A J, Vetter T, Blanz V, (1999). "Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing" *Vision Research* **39** 3145-3155.

O'Toole, A. J., Roark, D. A. and Abdi, H. (2002). "Recognizing moving faces: a psychological and neural synthesis," *Trends Cogn Sci*, vol. **6**, pp. 261-266.

Pearson, D. E. and Robinson, J. A. (1985). Visual communication at very low data rates, *Proceedings of the IEEE*, vol. **74**, no. 4, 795-812.

Pearson, D., Hanna, E. and Martinez, K. (1990). Computer generated cartoons, in Barlow, H., Blakemore, C. and Weston-Smith, M., Eds. *Images and Understanding*, chapter 3, pages pp. 46-60. Cambridge University Press.

Pellicano, E. and Rhodes, G. (2003). "Holistic processing of faces in preschool children and adults," *Psychological Science*, vol. **14**, pp. 618-622.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research,* **47**(3), 329-342.

Rhodes, G. (1996). *Superportraits: Caricatures and Recognition*. East Sussex, UK.: Psychology Press Publishers.

Riesenhuber, M., Jarudi, I., Gilad, S., and Sinha P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. *Proc. R. Soc. Lond.* B (Suppl.) **271**, S448-450.

Roark, D. A., O'Toole, A. J., and Abdi, H. (2003). Human recognition of familiar and unfamiliar people in naturalistic video. *IEEE International Workshop on Analysis and Modeling of Faces*, 36-43, Nice, France.

Robbins, R. and McKone, E. (2003). Can holistic processing be learned for inverted faces?, *Cognition*, vol. **88**, pp. 79-107.

Russell R, Sinha P, Biederman I, Nederhouser M, (2004). "The importance of pigmentation for face recognition" *Journal of Vision* **4** 418a.

Russell R, Sinha P, Biederman I, Nederhouser M, (Under review) "Is pigmentation important for face recognition? Evidence from contrast negation"

Sadr, J., Jarudi, I. and Sinha, P. (2003). The role of eyebrows in face recognition, *Perception*, vol. **32**, pp. 285-93.

Schiff, W., Banka, L. and de Bordes Galdi (1986). Recognizing people seen in events via dynamic "mug shots", *Am J Psychol*, vol. **99**, pp. 219-31.

Schwarzer, G. (2000). "Development of face processing: The effect of face inversion," *Child Development*, vol. 71, pp. 391-401.

Simion, F., Cassia, V. M., Turati, C. and Valenza, E. (2001). "The origins of face perception: Specific versus non-specific mechanisms," *Infant and Child Development*, vol. **10**, pp. 59-65.

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

Sinha, P., & Poggio, T. (1996). The role of learning in 3-D form perception. *Nature,* **384**, 460-463.

Sinha, P. (2002). "Qualitative representations for recognition," *Biologically Motivated Computer Vision, Proceedings*, vol. **2525**, pp. 249-262.

Sporer S L, Malpass R S, Koehnken G, (1996). "Psychological Issues in Eyewitness Identification", (Mahwah, NJ: Lawrence Erlbaum Associates)

Thompson, P. (1980) "Margaret Thatcher: a new illusion." *Perception*. **9**(4):483-4.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature,* **381**, 520-522.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response Properties of Human Fusiform Face Area. *Cognitive Neuropsychology,* **17**(1), 257-279.

Valentine, T. (Ed.). (1999). *Face-Space Models of Face Recognition*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences,* **98**(8), 4800-4804.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology,* **81**, 141-145.

Yin, R. K. (1970). Face recognition by brain-injured patients: a dissociable ability?, *Neuropsychologia*, vol. **8**, pp. 395-402.

Yip, A. and Sinha, P. (2002). Role of color in face recognition. *Perception*. Vol. **31**, 995-1003.

Young, A. W., Hellawell, D. and Hay, D. C. (1987). Configurational information in face perception, *Perception*, vol. **16**, pp. 747-59.