Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-010
AIM-2004-006
CBCL-236

March 5, 2004

Face processing in humans is compatible with a simple shape-based model of vision

Riesenhuber, Jarudi, Gilad, and Sinha

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Face processing in humans is compatible with a simple shape-based model of vision

Maximilian Riesenhuber[1,2,3], Izzat Jarudi[2], Sharon Gilad[2], Pawan Sinha[2]

[1] McGovern Institute for Brain Research and Center for Biological and Computational Learning, [2]Department of Brain and Cognitive Sciences Massachusetts Institute of Technology Cambridge, MA 02142, [3]Department of Neuroscience, Georgetown University Medical Center, Washington, DC 20007

Correspondence and requests for materials should be addressed to M.R. (e-mail: mr287@georgetown.edu).

keywords: object recognition, faces, psychophysics, inversion effect, neuroscience, computational neuroscience

short title: Shape-based face processing in humans

Understanding how the human visual system recognizes objects is one of the key challenges in neuroscience. Inspired by a large body of physiological evidence (Felleman and Van Essen, 1991; Hubel and Wiesel, 1962; Livingstone and Hubel, 1988; Tso et al., 2001; Zeki, 1993), a general class of recognition models has emerged which is based on a hierarchical organization of visual processing, with succeeding stages being sensitive to image features of increasing complexity (Hummel and Biederman, 1992; Riesenhuber and Poggio, 1999; Selfridge, 1959). However, these models appear to be incompatible with some well-known psychophysical results. Prominent among these are experiments investigating recognition impairments caused by vertical inversion of images, especially those of faces. It has been reported that faces that differ "featurally" are much easier to distinguish when inverted than those that differ "configurally" (Freire et al., 2000; Le Grand et al., 2001; Mondloch et al., 2002) – a finding that is difficult to reconcile with the aforementioned models. Here we show that after controlling for subjects' expectations, there is no difference between "featurally" and "configurally" transformed faces in terms of inversion effect. This result reinforces the plausibility of simple hierarchical models of object representation and recognition in cortex.

M.R. was supported by a McDonnell-Pew Award in Cognitive Neuroscience. I.J. was supported by the Meryl and Stewart Robertson UROP fund at MIT. P.S. is supported by DARPA's HumanID Program and an Alfred P. Sloan Fellowship.

Since its discovery by Yin in 1969 (Yin, 1969), the Face Inversion Effect (FIE) has acquired the status of a cornerstone finding in the domain of visual recognition. The dominant explanation of this effect is that the human visual system's strategy for facial representation is primarily 'configural', i.e. it involves encoding the 2nd order spatial relationships between face parts such as eyes, nose and mouth (Carey and Diamond, 1986; Farah et al., 1995; Freire et al., 2000; Le Grand et al., 2001; Mondloch et al., 2002; Tanaka and Farah, 1993). Configural analysis is believed to be compromised with vertically inverted faces, and, under these circumstances, the visual system is forced to resort to a 'featural' mode of processing. Directly testing this hypothesis, several studies have reported that changes to the "features" of a face (commonly defined as consisting of the eyes, mouth, and nose) can be detected equally well in upright as in inverted faces, while changes to the "configuration" of a face (defined as the "distinctive relations among the elements that define the shared configuration" (Carey and Diamond, 1986) of face features) cause an inversion effect, with much better detectability in upright than in inverted faces (Freire et al., 2000; Le Grand et al., 2001; Mondloch et al., 2002).

However, the models suggested by physiology are agnostic about the source of differences between two faces; they make no explicit distinction between featural and configural changes. According to the models, therefore, if two modifications to the shape of a face – be they due to changes in the "configuration" or in the "features" – influence discrimination performance to an equal degree for upright faces, they should also have an equal effect on the discrimination of inverted faces, *i.e.,* there is no special role for "configuration" or "features". There is, thus, an important inconsistency between reported psychophysical data and predictions from the hierarchical models of recognition.

A potentially significant shortcoming of the psychophysical studies mentioned above is that they have used blocked designs (trials were either grouped by change type (Mondloch et al., 2002) or used a different subject group for each change type (Freire et al., 2000)) where "featural" and "configural" changes were separated into different groups, making it possible for subjects to use change type-specific recognition strategies different from generic face processing strategies. For instance, in a blocked design, it is conceivable for subjects in featural trials to use a strategy that does not rely on the visual system's face representation but rather focuses just on detecting local changes in the image, *e.g.,* in the eye region. This would then lead to performance less affected by inversion. However, such a local strategy would not be optimal for configural trials as there the eye itself does not change but only its position with respect to the rest of the face. Thus, configural trials can profit from a "holistic" strategy (*i.e.,* looking at the whole face, which for upright but not for inverted faces presumably engages the learned (upright) face representation), which would in turn predict a strong effect of inversion.

We therefore performed a same/different face matching experiment using face pairs differing in features or configuration, in which subjects were not able to predict change type (see Fig. 1 and Methods; for additional details, see Supplementary Information).

Materials and Methods

Subjects performed a same/different task using pairs of faces differing either by a "configural" (Fig. 1a) or a "featural" (Fig. 1b) change. Photorealistic stimuli were created using a custom-built morphing system (see Supplementary Material) that allowed us to freely move and exchange face parts of 200 face prototypes (Blanz and Vetter, 1999). Subjects performed a total of 160 trials, based on 80 image pairs, each presented upright and inverted on different trials. 40 face pairs were associated with "featural trials": 20 face pairs with the faces in each pair differing by a feature change (replacement of eyes and mouth regions with those from other faces prototypes (Freire et al., 2000; Le Grand et al., 2001; Mondloch et al., 2002)) and 20 "same" face pairs composed of the same outlines and face component positions as the corresponding "different" faces, with both faces having the same eyes/mouth regions. Another 40 face pairs were used in the "configural change" trials, consisting of 20 face pairs differing by a configural change (displacement of the eyes and/or mouth regions (Freire et al., 2000; Le Grand et al., 2001; Mondloch et al., 2002)), plus 20 "same" face pairs composed of faces with the same face outlines and parts as the corresponding "different" pairs, with both faces in the pair having the same configuration. Faces were selected so that performance in upright featural and configural trials was comparable. In the "unblocked" version of the experiment, configural and featural trials were presented in random (and counterbalanced) order. In the "blocked" version, trials were grouped by change type.

Results

Results for 15 subjects are shown in Fig. 2. As expected, inversion adversely affects performance. More importantly, we find the effect of inversion to be comparable for "featural" and "configural" changes. This is also borne out by an ANOVA that shows a highly significant main effect of orientation ($p<0.0001$) but no main effect for change type ($p>0.29$) and no interaction ($p>0.17$). These results are compatible with a shape-based representation specialized for upright faces, but not with a representation that explicitly encodes facial "configuration."

These results are thus in notable contrast to previous experiments that have found an Inversion Effect for featural but not for configural changes (Freire et al., 2000; Le Grand et al., 2001; Mondloch et al., 2002). To investigate whether the earlier results might have been an artefact of blocking trials by change type, we tested additional subjects on a modified version of our experiment with identical trials, but this time blocked according to change type. Thus, one group of subjects ($n=12$, the "configural first" group) was first exposed to all trials containing the "configural" images (including "same" and "different" trials, see Methods), and then all "featural" trials, while the blocks were reversed for another group of subjects ($n=13$, the "featural first" group). Subjects were not informed that images in the two blocks differed in any way. None of the subjects had participated in the unblocked experiment.

Interestingly, blocking the trials caused subjects' performance to vary substantially depending on which group they belonged to: In the "configural first" group, subjects

showed no difference in performance over all "configural" *vs.* all "featural" trials (t-test, $p=0.19$), compatible with the hypothesis that subjects used the same holistic, face-based strategy for all trials, as in the original unblocked experiment (where there was no difference between average subject performance on featural and configural trials, $p>0.7$). The situation was very different in the "featural first" group, where performance on featural and configural trials was highly significantly different ($p=0.001$). This was due to poor performance on the configural trials (63%, vs. 73% on featural trials), as would be expected if subjects used a strategy based on local, part-based image comparisons. Indeed, ANOVAs for the different subject groups showed a significant main effect of orientation in both groups ($p<0.001$), but a highly significant effect of change type only in the "featural first" group ($p<0.001$). The effect of change type missed significance for the "configural first" group ($p>0.05$), similar to the original, unblocked design. This suggests that blocking trials can cause subjects to adopt artifactual visual strategies.

Discussion

Our psychophysical results therefore help reconcile an important inconsistency between past experimental data and predictions from modeling and physiology. They strongly support a simple shape-based model of visual processing, in agreement with physiological data. Furthermore, they suggest that the representation of facial shape information, while holistic, is not explicitly configural. This hypothesis makes interesting predictions regarding the response properties of face-selective neurons in the primate inferotemporal cortex, a brain area crucial for object recognition in primates (Logothetis and Sheinberg, 1996). Many "face neurons" have already been shown to exhibit "holistic" tuning (Farah et al., 1995; Tanaka and Farah, 1993) given that they "require nearly all the essential features of a face" for activation (Tanaka, 2003). Based on our experimental results here, we would predict that "featural" and "configural" changes of a face stimulus that cause an equal activation change for upright faces should also have an equal effect for inverted faces (but likely of lower magnitude given the preferred tuning of most face neurons to upright faces (Tanaka et al., 1991)), in marked contrast to configural theories.

References

Blanz, V., and Vetter, T. (1999). *A morphable model for the synthesis of 3D faces*. In: SIGGRAPH '99 (ACM Computer Soc. Press), 187-194.

Carey, S., and Diamond, R. (1986). Why faces are and are not special: An effect of expertise. *J Exp Psychol Gen* 115, 107-117.

Farah, M. J., Tanaka, J. W., and Drain, H. M. (1995). What causes the face inversion effect? *J Exp Psychol Hum Percept Perform* 21, 628-634.

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1, 1-47.

Freire, A., Lee, K., and Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: direct evidence. *Perception* 29, 159-170.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160, 106-154.
Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol Rev* 99, 480-517.
Le Grand, R., Mondloch, C. J., Maurer, D., and Brent, H. P. (2001). Neuroperception. Early visual experience and face processing. *Nature* 410, 890.
Livingstone, M. S., and Hubel, D. H. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* 240, 740-749.
Logothetis, N. K., and Sheinberg, D. L. (1996). Visual Object Recognition. *Annual Review of Neuroscience* 19, 577-621.
Mondloch, C. J., Le Grand, R., and Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception* 31, 553-566.
Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2, 1019-1025.
Selfridge, O. G. (1959). *Pandemonium: A paradigm for learning*. In: Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physics Laboratory (HMSO, London).
Tanaka, J. W., and Farah, M. J. (1993). Parts and wholes in face recognition. *Q J Exp Psychol A* 46, 225-245.
Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb Cortex* 13, 90-99.
Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J Neurophysiol* 66, 170-189.
Tso, D. Y., Roe, A. W., and Gilbert, C. D. (2001). A hierarchy of the functional organization for color, form, and disparity in primate visual area V2. *Vision Res* 41, 1333-1349.
Yin, R. K. (1969). Looking at upside-down faces. *J Exp Psychol* 81, 141-145.
Zeki, S. (1993). *A vision of the brain* (Oxford, England, Blackwell Scientific Publishing).

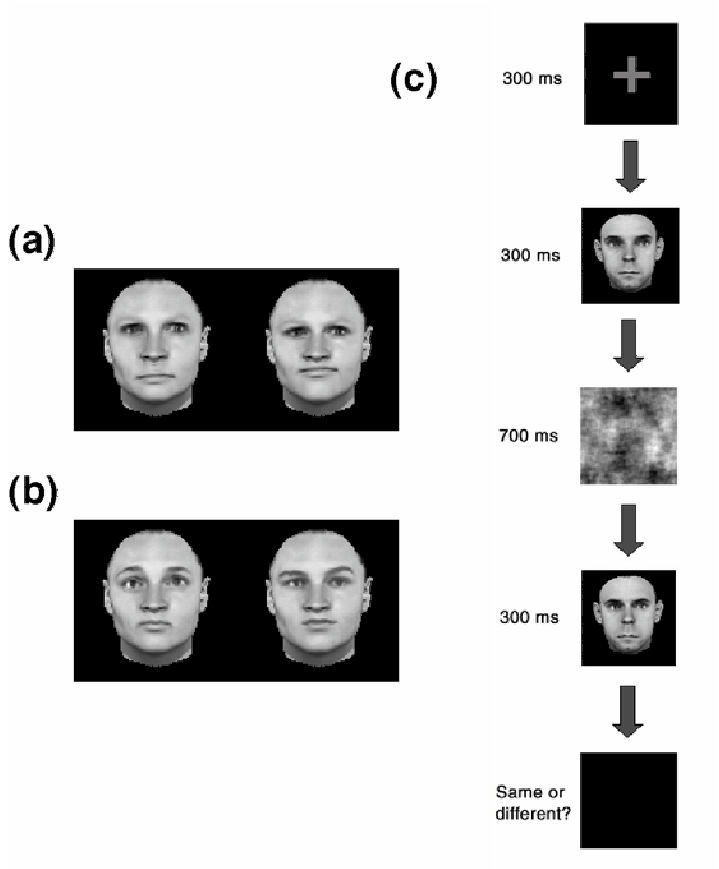

Figure 1: Example stimuli and task. (a) Example of a “configural” change stimulus pair. The two images show two versions of the same face, with the right face’s mouth moved up and the eyes moved together relative to the left one. (b) Example of a “featural” change stimulus pair. The two images show two versions of the same face, with the right face’s mouth and eyes replaced by the mouth and eyes from another, randomly selected face. The face outline and the nose are the same as in (a). (c) Experimental paradigm. Subjects first fixated on a cross for 300 ms, then viewed one of two pictures in a face pair (here, a “configural change” pair) for 300 ms, followed by a noise mask for 700 ms, and the second picture in the pair for 300 ms, and finally a blank screen until subjects made their same/difference judgment by pressing a specific keyboard button.

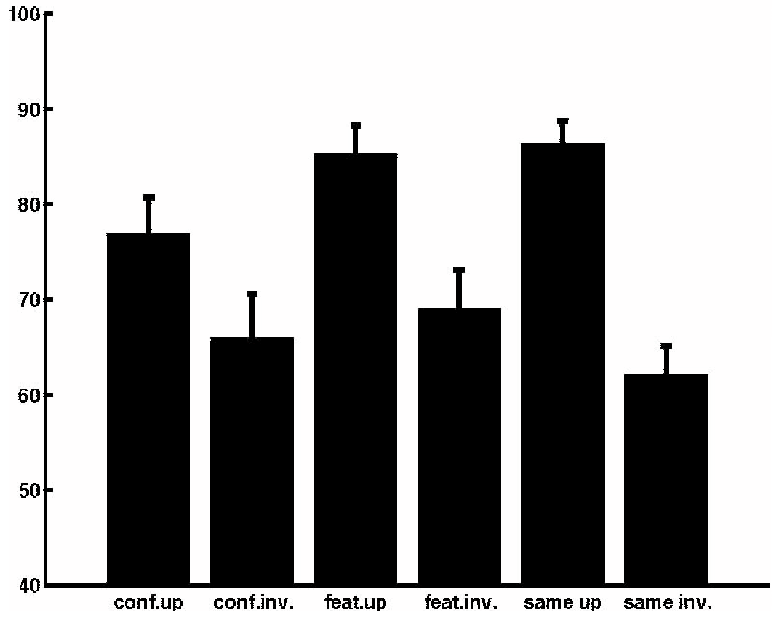

Figure 2: Subject Performance in the unblocked discrimination experiment. The bars show subjects' performance (n=15) for the different trial types ("conf." refers to trials where the two images differed by a configural change, "feat." to trials with featural changes, and "same" to trials with two identical images; "up" refers to upright images, "inv." to inverted images). Error bars show standard error of the mean.

Supplementary Information

This document provides more details on the experimental design described in the main paper.

1 Experiment 1: Selection of Face Pairs

Experiment 1 was devoted to the collection of a well-controlled yet substantially large stimulus set for Experiments 2 and 3. To consistently measure discrimination impairment as a result of inversion, this stimulus set needed to consist of images for which subjects were equally good at detecting configural and featural changes when upright. Any significant difference in upright detection would be a confounding variable since it could also affect how featural and configural cues were processed when inverted.

1.1 Subjects

A total of thirty adults were randomly assigned to one of two subject groups. Thus, each subject group contained 15 mutually exclusive participants. All subjects were naive as to the exact purpose of the experiment. Subjects had normal or corrected-to-normal vision.

1.2 Stimuli

Two hundred original male and female face images (the face prototypes from [1]) were manipulated to compile 2,428 pairs of gray-scale, frontal faces used in this experiment. To avoid providing easy cues for identification, the faces were cut at the hairline, and did not include distinctive features such as beards or spectacles. The 200 images were divided into two groups: source faces and target faces. Using a customized morphing program written in MATLAB (The Mathworks, MA), image manipulations were applied as follows: Each source face was duplicated and prepared in 10 configural versions and 10 featural versions, differing in the position (for the configural set) or shape (for the featural set) of the eye and/or mouth region [2, 3, 4]. Configural changes involved spatial shifts of the eyes and/or mouth such that the basic symmetry of a face was preserved. Eyes were displaced a maximum of two pixels along the x-axis (in or out) and/or two pixels along the y-axis (up or down). The mouth was displaced a maximum of two pixels along the y-axis (up or down). Both magnitude and direction of movement were randomly selected. Featural changes were accomplished by replacing the eyes and mouth of a source face with the corresponding features from a randomly selected target face. Target features used on a given source face were selected from two different faces (for instance, source A received eyes from Y and mouth from X), and the combination was unique. Source faces were matched with target features of the same gender. Note that all eye alterations treated the eyes and eyebrows as a single unit.

Each of the 20 created images were then paired once with themselves, and once with another image that underwent the same type of manipulation. The above procedure was repeated for each of the 100 source faces, resulting in a total of 4,000 unique pairs. These

pairs were then screened for morph artefaces. The resulting stimulus set was comprised of 2,428 image pairs, with an equal number of those pairs (607) belonging to each of the four stimulus categories: pairs with the same features (SF), same configuration (SC), different features (DF), and different configuration (DC).

1.3 Procedure

Two subject groups underwent the exact same experiment, with the exception that the stimulus set presented to each group consisted of different images. This design allowed us to collect a large stimulus set while avoiding noise due to subject exhaustion. The resulting data consisted of 15 different responses for each of the 2,428 image pairs. Prior to beginning the experiment, both written and oral instructions were given. Subjects were told that the goal of the experiment was to investigate how well they could distinguish between similar faces, and warned about the possible/potential similarity of the two faces (like identical twins with different expressions).

Subjects were presented with approximately 10 blocks, each with 100 image pairs (numbers differed somewhat for each subject groups). The order of presentation was randomly selected at the beginning of each experiment. In a given trial, a selected pair was presented as a sequence of two images using the following design: cross (300ms) → image 1 (300ms) → random noise mask (700ms) → image 2 (300ms) → blank screen until subjects respond by pressing one of two labeled keys. Image selection, presentation, and the recording of subject input were controlled using a MATLAB (The Mathworks, MA) program.

2 Experiment 2: Face Discrimination (Unblocked)

In this experiment, featural and configural image pairs were randomly presented both upright and inverted. The aim was to determine how well subjects could judge two inverted pictures to be the same or different depending upon whether the images differed by a featural or a configural change.

2.1 Subjects

Fifteen subjects, ranging in age from 19 to 56 (mean age 33), participated in the study for payment.

2.2 Stimuli

Through Experiment 1, the initial stimulus set of 2,428 image pairs was reduced to 80 for Experiment 2 by choosing only among those that were correctly judged to be the same or different two images approximately 80% of the time (78.0% on average for the entire set). The 80 pairs were evenly divided into each of the four stimulus categories: SC, DC, SF, and DF.

2.3 Procedure

Each subject viewed all 80 image pairs in the stimulus set both upright and inverted, thus leading to 160 trials in the entire experiment. Both the presentation order of stimulus type (SC, DC, SF, or DF) and orientation (whether upright or inverted) was randomized and counterbalanced.

In each trial, subjects were shown the two pictures in a face pair in a sequence as before (see Expt. 1), either both upright or both inverted (see Fig. 1 in the paper). Each subject received the same oral instructions from the same experimenter who ran all the subjects in Experiments 2 and 3. Subjects were told that the goal of the experiment was to measure their abilities to differentiate between two similar faces. They were warned that the two pictures in each face pair were going to be very similar (like pictures of two twins or different expressions). Their task was simply to determine whether the two pictures were identical or not. They were explicitly instructed that they did not have to judge whether the two images were of the same face or person, just whether there was any perceptible difference between the two images. Subjects were also informed that there were 160 trials in all with a break halfway through the experiment for them to rest if they liked.

Images were displayed on a 21" Trinitron monitor, at a true size of 186 x 186 pixels (the remainder of the screen remained neutral black throughout the experiment). The screen resolution was set to 1024 x 768 pixels, with a refresh rate of 60 Hz. Image selection, presentation, and the recording of subject input were automated and controlled using a MATLAB program. Subjects freely viewed the monitor from a distance of about 60 cm.

3 Experiment 3: Face Discrimination (Blocked)

Experiment 3 was identical to Experiment 2 in all the details of its stimuli and procedure except that featural trials (SF,DF) and configural trials (SC,DC) were grouped together in different stimulus blocks corresponding to the first and second halves of the experiment. Within each block, presentation order of stimulus type and orientation was randomized. If the featural block was first, for example, subjects would see 80 featural trials, half upright, half inverted and half with the same features (SF), half different (DF) in the first half of the experiment. Whether any particular subject saw the featural or configural block in the first half of the experiment and the other in the second half was randomized. Subjects received no indication that the trials were blocked by stimulus type and were told, as in Experiment 2, that the break halfway through the experiment was intended for them to rest if they liked.

3.1 Subjects

Twenty-five subjects, ranging in age from 18 to 52, participated in the study for payment. Subjects were randomly placed in two non-overlapping groups corresponding to the two blocked presentation orders (featural or configural block first). The mutual exclusion was enforced to prevent any transfer of information from one condition to another. Twelve subjects were tested in the configural block first condition with the remaining thirteen viewing the featural block first.

References

[1] Blanz, V. and Vetter, T. In *SIGGRAPH '99 Proceedings*, 187–194. ACM Computer Soc. Press, (1999).
[2] Freire, A., Lee, K., and Symons, L. A. *Perception* 29, 159–170 (2000).
[3] LeGrand, R., Mondloch, C. J., Maurer, D., and Brent, H. P. *Nature* 410, 890 (2001).
[4] Mondloch, C., LeGrand, R., and Maurer, D. *Perception* 31, 553–566 (2002).