Generalization from a single view in face recognition

Maria Lando Shimon Edelman†
Department of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
† edelman@wisdom.weizmann.ac.il

Abstract

We describe a computational model of face recognition, which generalizes from single views of faces, by taking advantage of prior experience with other faces, seen under a wider range of viewing conditions. The model represents face images by vectors of activities of graded overlapping receptive fields (RFs) [10]. It relies on high spatial frequency information to estimate the viewing conditions, which are then used to normalize (via a transformation specific for faces), and identify, the low spatial frequency representation of the input. The class-specific transformation approach allows the model to replicate a series of psychophysical findings on face recognition [7], and constitutes an advance over current face recognition methods, which are incapable of generalization from a single example.

1 Introduction

The ability to recognize a novel view of a face previously seen under a restricted range of conditions is one of the more amazing feats of human vision. We describe a computational model of generalization from a single view in face recognition, built around the assumption that such generalization is made possible by the previous experience of the visual system with similar objects (i.e., other faces). In particular, we assume that the visual system stores information regarding the appearance of a considerable number of faces under a relatively wide range of conditions, and ask how such information can be put to use in generalizing to novel views of an unfamiliar face. The answer suggested by our results provides an explanation of a number of recent psychophysical findings, and is likely to lead to a significant enhancement in the performance of the present-day computer vision systems for face recognition, which are by and large incapable of generalization from a single view.

1.1 Psychophysical background

A new insight into the computational basis of human generalization performance in face recognition has been achieved as a result of a recent study that compared generalization for upright and inverted faces [7]. Human subjects performing an upright face discrimination task were found to generalize nearly perfectly to face images obtained under novel illumination and viewpoint direction. In comparison, for inverted faces, the generalization to novel views was significantly worse. The control provided by inverted faces allowed the researchers to rule out a universal (i.e., valid for all images) basis for generalization. Furthermore, the subjects' ability to generalize from a single image of an upright face made the possibility of strictly model-based generalization (which, in principle, requires more that one image to form a full 3D model) unlikely. The researchers concluded that the generalization occurs at an intermediate or class-based level, where upright faces constitute a class distinct from that of inverted faces (despite the pixel by pixel equivalence in complexity of the two types of stimuli).

The central difference between upright and inverted faces, which presumably allows the former, but not the latter, to be recognized easily under a wide range of unfamiliar conditions, is the extensive experience of the human visual system with upright faces in everyday life [4]. Indeed, face recognition appears to be, to a considerable extent, an acquired ability. Children's recognition of upright faces improves steadily from age 6 to 10, dips temporarily at $11-12$, then climbs to an adult level [3]. In comparison, recognition performance on inverted faces does not change throughout the life and remains significantly worse [2, 3].

1.2 Computational approach

Our approach to the problem of generalization from a single view is based on a formalization of the notions of class-based processing and of experience, outlined below.

RF-based representation. A basis for the similarity of face images must be provided before one can discuss class-related issues. In biological visual systems, such a basis can be derived naturally from the concept of a unit with a localized receptive field (RF), known in psychophysics as spatial frequency channel [11]; activities of the RF units at the previous stages is the only input available to any processing stage in the visual pathway past the retina.

Clustering by face identity. We assume that the metrics of the internal representation space for faces reflects the true metrics prevailing in the objective "face space," in the sense that different appearances of the same face (i.e., views taken under different viewing positions and illumination conditions) tend to cluster together. This assumption relies on a recent computational investigation [10], which showed that images of the same face form tighter clusters at the higher levels of an RF-based representation hierarchy resembling the one found in mammalian vision. The same study found that, although higher overlap between the constituent RFs improves the utility of a representation, this improvement saturates when the number of RFs reaches a few hundred (i.e., a relatively small set of RFs is nearly as useful as a full dimensionality-preserving coverage of the retinal space).

Class-based generalization. We assume that applying the same transformation (change of viewpoint or illumination) to images of different faces results in similar changes in the internal representation of each face. Thus, the transformation that maps a prototypical view of a face to its appearance under specific viewing conditions is assumed to be similar across different faces.

This last assumption constitutes the main working hypothesis of the present paper. Consider two viewing positions, V_0 and V_1. We conjecture that a face previously seen from V_0 can be recognized when seen from V_1 because the system stores a snapshot of the RF activity evoked by the exposure to the face in viewing position V_0, and compares the stored activity vector to the present one, taking into account experience with numerous other faces in both positions, V_0 and V_1. Note that the same approach can be applied to the case of varying illumination, face expressions and different combinations of these parameters.

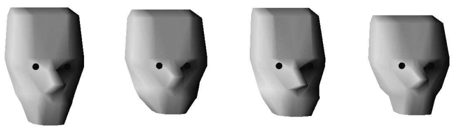

Figure 1: Faces obtained by random variation of the parameters in the face geometry model (section 2).

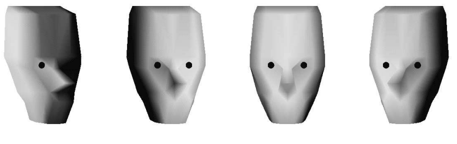

Figure 2: One of the face models rendered under different viewpoint and illumination conditions.

2 Class-specific transformations in RF space

To substantiate the notion of class-based generalization, we examined the expected behavior of RF-space representations of face images under various transformations of faces, and compared the resulting computational predictions with data derived from artificial (computer graphics) and real (human) faces.

We first define geometric similarity of human faces, via a shape variation parameter δ that expresses the upper bound across individuals on the changes in each component of the normal at any point of a face. Assuming that the specular component of the face reflectance function is relatively small, the intensity at each point of the face image will depend on the angle between the normal at this point and the illumination vector. Thus, the resemblance of different faces to a certain average one can be described in terms of the parameter δ, which bounds the variation in the normal at each point of the face.

Class-specific transformations for artificial faces. How do changes in viewing conditions affect RF-space representation of faces? Let $\mathbf{P}^{(f_1)} \in R^k$ and $\mathbf{P}^{(f_1)} + \Delta\mathbf{P}^{(f_1)}$ be the RF-space representations of face f_1 under two different illuminations. The i'th component of $\Delta\mathbf{P}^{(f_1)}$ is then proportional to the change in the normal direction at the i'th RF, and can differ at most by δ from person to person. As shown in [6], for two persons f_1 and f_2 for whom the corresponding normals differ at most by δ, the length of the difference vector $\|\Delta\mathbf{P}^{(f_1)} - \Delta\mathbf{P}^{(f_2)}\|$ is bounded from above by δ. Furthermore, the cosine of the angle between $\Delta\mathbf{P}^{(f_1)}$ and $\Delta\mathbf{P}^{(f_2)}$ is bounded from below by $\sqrt{1-\delta^2}$.

If the transformation includes a shift in the viewpoint, the vector $\Delta\mathbf{P}^{(f_1)} - \Delta\mathbf{P}^{(f_2)}$ depends also on the interaction between changes in illumination direction and viewpoint. The lower bound on the cosine of the angle between $\Delta\mathbf{P}^{(f_1)}$ and $\Delta\mathbf{P}^{(f_2)}$ depends then both on δ and on the minimal change of the projection of the normal onto the illumination direction under a shift in the viewing position (i.e., $\min\|(\mathbf{L}\cdot\Delta\mathrm{N})\|$, where $\mathbf{L}$ is the illumination vector and $\mathbf{N}$ is the normal vector). As shown in [6], this bound is equal to: $\sqrt{1-(2\delta/\min|(\mathbf{L}\cdot\Delta\mathrm{N})|)^2}$.

We have tested the validity of the above bounds on class-specific transformations in RF space empirically, using a 19-parameter model of face geometry, constructed with a 3D graphics toolkit (SGI Inventor). Figure 1 depicts eight of the 50 faces that were obtained by randomly varying the 19 parameters. Images of each of the 50 face models were computed under five different viewing positions and three illumination directions, and were represented by activities of 500 RFs, as described in section 1 (see also [10]). The calculated values of of $\|\Delta\mathbf{P}^{(f_i)} - \Delta\mathbf{P}^{(f_j)}\|$ and $\angle(\Delta\mathbf{P}^{(f_i)}, \Delta\mathbf{P}^{(f_j)})$ were nearly identical for the different face models (i.e., different choices of f_1, f_2), and in all cases lay well within the analytically derived bounds [6].

Class-specific transformations for real faces. We next assessed the behavior of $\|\Delta\mathbf{P}^{(f_i)} - \Delta\mathbf{P}^{(f_j)}\|$ and $\angle(\Delta\mathbf{P}^{(f_i)}, \Delta\mathbf{P}^{(f_j)})$ on real face images, taken from the Weizmann FaceBase [7]. The results for the length and the angle differences appear, respectively, in Figures 3 and 4, which illustrate the correlated changes induced in the RF-space representations of different faces by changes in the viewing conditions. These results (especially, the tight correlation exhibited by $\angle(\Delta\mathbf{P}^{(f_i)}, \Delta\mathbf{P}^{(f_j)})$ for the different face pairs i,j) support the hypothesis that view transformations induce similar changes in RF-space representation, regardless of face identity.

3 Class vs. transformation-based clustering in RF space

The results of the previous section provide the foundations for a class-specific approach to face processing, by demonstrating that same-class objects (faces) induce similar patterns in the RF representation space across changes in viewing conditions. This finding can only be put to use in face identification if it is possible to distinguish, in the RF space, between points representing images of the same face

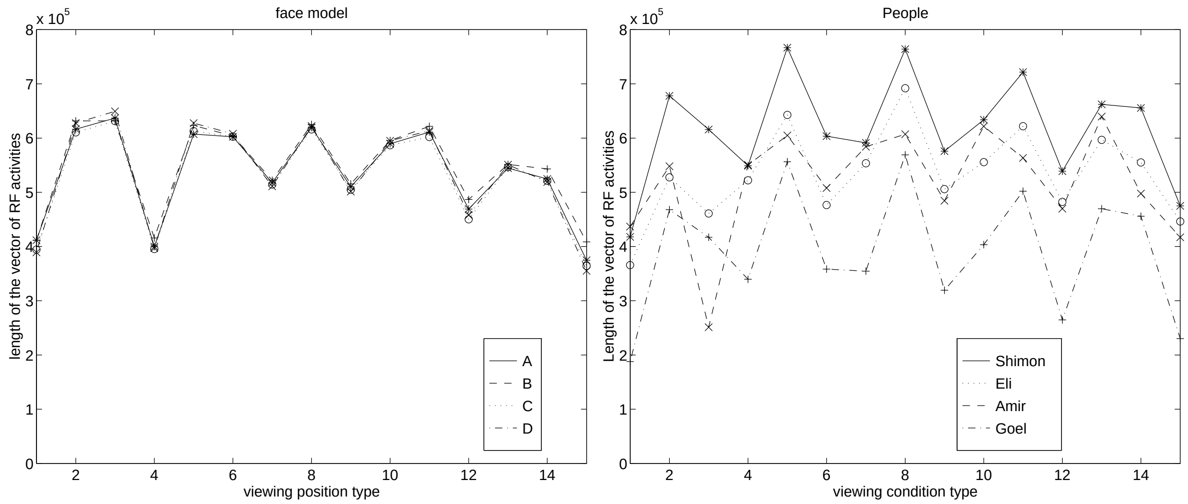

Figure 3: Length changes of the RF-space vectors corresponding to 4 different face models (left) and 4 real faces (right), over 15 different viewing conditions. Note that the changes for different faces are correlated with each other.

from points representing all other faces. In principle, the clustering in the RF space can take one of two possible forms. The first possibility is that points representing the same face are clustered together. Alternatively, points representing the same viewing conditions can be clustered together. In both cases, the correlations apparent from Figures 3 and 4 would stem from the common displacement of all clusters following a certain transformation (either of view or of face identity).

The spatial frequency analysis of face-difference images reported in [10] suggests that the nature of clustering of faces in RF-space should depend on the size of the RFs used in the representation. We tested this conjecture, by convolving face images with Difference of Gaussian RFs, corresponding in size to the different spatial frequency channels described in [11]. Eight values of RF size, between 0.8 cycles per degree (*cpd*) and 16 *cpd*, were used with images of different faces, under different viewing conditions. For each RF size, we conducted 60 trials, with faces chosen randomly from the Weizmann database. The degree of clustering was defined as the ratio R of mean within-cluster and between-clusters distances (under this measure, $R << 1$ indicates good clustering).

Two experiments were carried out. In the first experiment only the illumination was varied (Figure 5, left). Clustering by face identity improved with increasing RF size, while clustering by viewing condition became better when the RF size decreased. In the second experiment, viewing position varied (over 68^o of visual angle) in addition to illumination (Figure 5, right). Here, clustering by face identity did not improve with increasing RF size as much as in the previous case, although it did become better than clustering by viewing position.

The principle of class-based generalization, stated in the introduction, requires tighter clustering by identity than that available even with the low spatial frequency RFs. It turns out, however, that one can effectively combine the viewpoint information carried by the high-frequency (16 *cpd*) RFs with the face identity information in the low-frequency (0.8 *cpd*) RFs, by training appropriately constructed function approximation modules (we used for this purpose radial basis function, or RBF, classifiers [8, 5]). The resulting recognition scheme is described in the next section.

4 A complete model for face recognition

The model consists of the following stages:

1. **Detection of viewing conditions.** This is implemented by an RBF classifier, which accepts image representations in the high-frequency RF space.
2. **Normalization to the prototype.** According to the detected viewing conditions, a class-specific transformation transforms low-frequency RF-space representation into a prototypical form predicted for the input image.
3. **Face identification.** This is implemented by an RBF classifier, which, in the low-frequency RF space, compares the hypothesized prototypical representation with those of known faces, and identifies the input face.

In this model, the amount of "prior experience" with face transformations corresponds to the size of the set of individual face images, paired with the images of the same faces obtained under different but known viewing conditions (in our experiments, each face was represented by 15 images, taken under all combinations of 5 viewpoints and 3 illuminations). The first-stage RBF classifier (viewpoint detector) is trained on these transformation examples. The training set of the second-stage RBF classifier (identity detector)

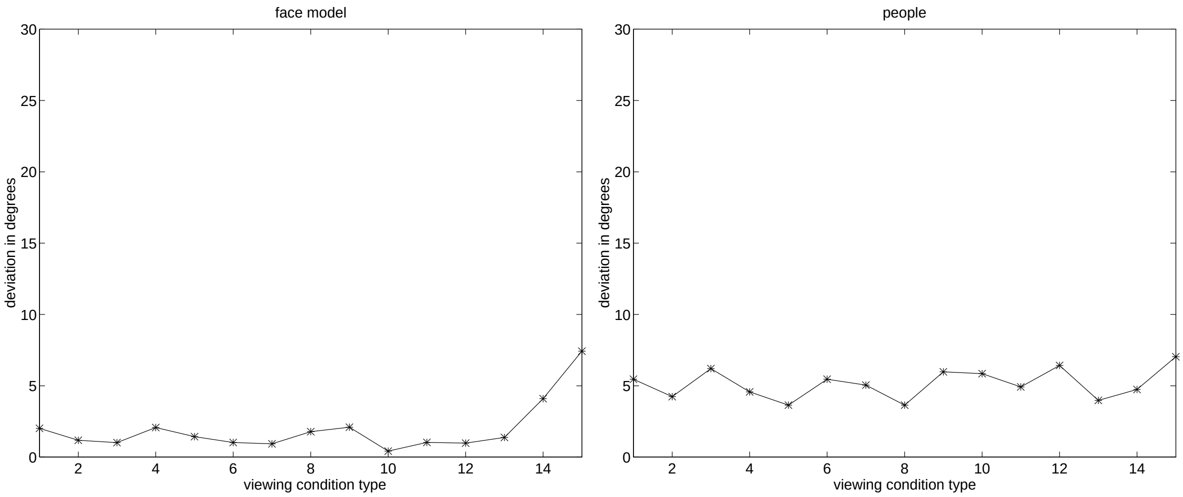

Figure 4: Direction (angle) changes of the RF-space vectors corresponding to 4 different face models (left) and 4 real faces (right), over 15 different viewing conditions. The changes here are even better correlated across faces than in Figure 3.

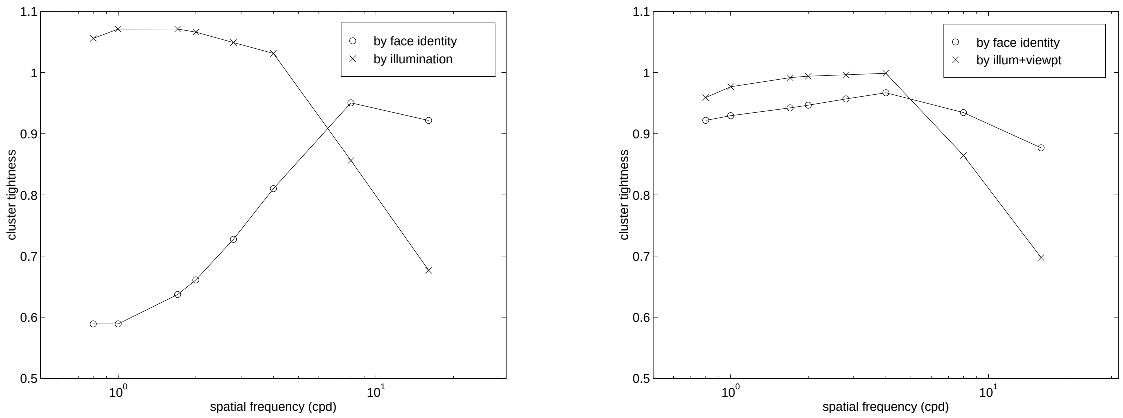

Figure 5: *Left:* The degree of clustering (R; see section 3) by face identity and by illumination, for different RF sizes. *Right:* Clustering by face identity and by viewing conditions (illumination and viewpoint), for different RF sizes.

is a collection of (RF-space representations of) prototypical images of known faces, paired with identity information for those faces.

Given an input image, the system first constructs its high and low spatial frequency RF representations. Once the viewing conditions are determined by the first-stage classifier from the high-frequency representation, the proper class-specific transformation is applied in the low-frequency RF space, yielding the predicted prototype. The class-specific transformation here is calculated as the average of the transformations corresponding to the different known faces, stored as the RBF centers in the first-stage classifier.

The estimated prototype is passed on to the second classifier, which, by interpolation among stored examples, identifies the image with one of the familiar faces. After each recognition trial, the class-specific transformation that corresponds to the recognized view is used to update the model's first-stage database. In addition, the identity database (stored in the second stage) can also be updated. In this manner, the model is capable of improving its performance with experience.

5 Comparing model and human performance

We compared the performance of the model with that of human subjects, by replicating the upright/inverted face recognition experiments described in [7]. In the simulated experiment, the model was trained on a varying number of upright faces (each shown under a range of conditions, as described above). It was then confronted with the same task performed by the human subjects, namely, discrimination between diverse images of three faces previously seen from a single viewpoint and under a fixed illumination. The difficulty of this task is noteworthy: the system was provided with just one view

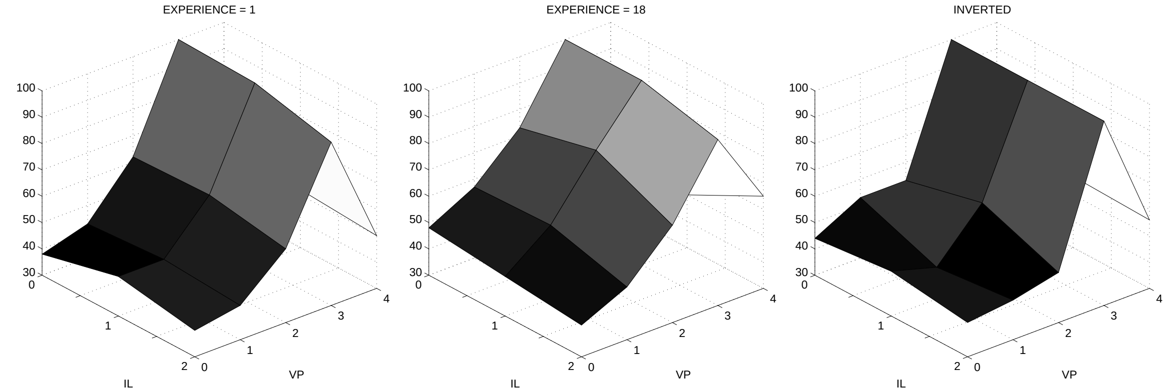

Figure 6: The improvement in the model's generalization performance with experience. *Left:* experience= 1, upright faces. *Middle:* experience= 18, upright. *Right:* experience= 18 with upright faces, testing with inverted faces (cf. [7]). The performance is similar to that of human subjects (Figure 7) in several respects (upright/inverted comparison; better generalization over illumination vs. viewpoint; see [6] for details).

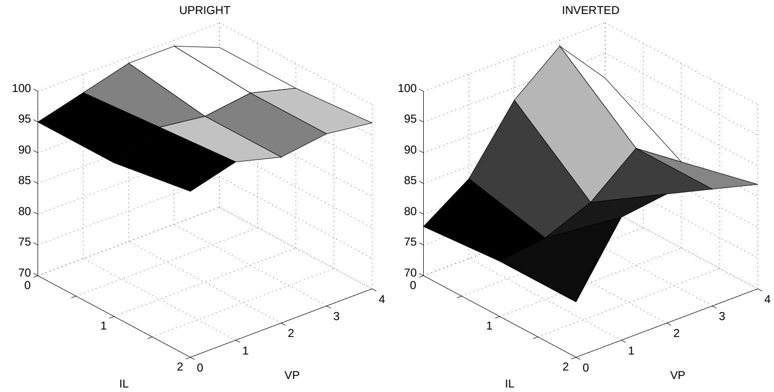

Figure 7: Human performance for upright (left) and inverted (right) face images (replotted from [7]).

of each face, and had to recognize 15 other images of the same face, taken from very different (up to 68^o) viewpoints, and under widely varying illuminations.

The extent of the prior experience of the model was varied between 1 (seeing a single individual at the 15 training orientations/illuminations) and 18 (seeing nearly all the individuals in the database). This caused the model's mean generalization performance (correct recognition rate for novel views) to grow from 59% to 66% correct. When the knowledge of bilateral symmetry of faces was taken into account, as suggested (albeit in a different context) in [9], an additional training view became effectively available to the system, boosting the generalization performance in the high experience case to 76%. Now, this figure is similar to the performance of the first stage of the system of [5] (which, however, was only capable of generalization from *multiple* examples). Adding an ensemble-based classifier improved the performance of that system by over 20%, and should have a similar effect on the performance of the present model. Furthermore, we presume that a larger database of individuals (comparable to the extent of an adult's lifetime exposure to faces) would allow the system to parallel the essentially perfect ($\approx$ 97% correct) generalization performance of human subjects for upright faces.

Figure 6 shows two plots of the generalization performance for upright test images (with two extreme levels of experience with upright faces – 1 and 18), alongside a plot of a system exposed to 18 upright faces and tested on the generalization over images of inverted faces (paralleling the experiments of [7]). A comparison of these plots with those of human subjects (Figure 7) shows that the contrast between generalization patterns for upright and inverted faces — the main point of the present experiment, as well as of the experiments described in [7] — is similar in the model and the human data.

6 Summary and discussion

We have presented a model of the human ability to generalize face recognition from a single image. The model is based on a computational analysis of the notion of class-specific transformations, and is supported by computer simulations, in which it exhibited a considerable ability to generalize from single images. The model also replicated the central findings of a recent psychophysical study which examined

V IL	-34	-17	0	17	34
0	38	43	62	100	55
1	45	45	63	99	52
2	40	43	58	92	50

V IL	-34	-17	0	17	34
0	48	57	73	100	58
1	45	58	80	100	52
2	42	50	67	93	65

Table 1: Generalization performance for test images, following exposure to a single training image (see also Figure 6). *Left:* experience=1; *Right:* experience=18.

generalization in upright and inverted faces in human subjects. From the practical standpoint, the model constitutes a significant advance over the approach of Edelman et al. [5], which also used RBF classifiers to learn face recognition from examples, but was incapable of generalization from a single image of a face. From the standpoint of theoretical neurobiology, the model suggests a promising approach to the understanding of the computational basis of class-based generalization in human vision.

The relevance of our model to the understanding of human vision stems not only from its performance in the upright/inverted face recognition experiments, but also from its predictions regarding the effect of face distinctiveness on generalization. The less similar a face is to an average human face (in a geometrical sense), the more different are the relationships among the RF-space representations of its views, compared to those of an average face. In such a case, one would tend to predict worse recognition results. However, the third step of the recognition in our model — comparing the predicted prototypical representation with those of known prototypes — is actually easier for a distinctive face, because the more unusual the face is, the larger the RF-space distance between its representation and those of the other faces. Consequently, it should be more difficult to generalize over novel views of distinctive faces, but also more difficult to misrecognize them under more familiar viewing conditions. These predictions are consistent with the results of recent experiments carried out by F. Newell (personal communication), who found better performance for distinctive faces in the mismatch trials, but not in the match trials in a face matching experiment.

The interpretation offered by our model for the results obtained with distinctive faces can be extended to account for the peculiarities of recognition of faces across race [1]. It is well known that people used to see predominantly Caucasian faces find it more difficult to distinguish among Oriental faces than people living in the Orient, and vice versa. We conjecture that this happens because of the limited applicability of class-specific transformations to a radically different population of face shapes. With practice, the relevant portion of the representation space may become populated by the proper prototypes, and the discrimination performance may improve (as it indeed happens in cross-racial face recognition).

As a conclusion, we point out that the class-specific transformation approach adopted in the present work can be extended to classes of objects other than faces, and is probably worth exploring in a wider context of computer vision.

Acknowledgements. We thank Ronen Basri and Yael Moses for useful discussions and for constructive comments on an earlier version of this work.

References

[1] J. C. Brigham. The influence of race on face recognition. In H. D. Ellis, M. A. Jeeves, and F. Newcombe, editors, *Aspects of face processing*, pages 170–177. Martinus Nijhoff, Dordrecht, 1986.

[2] S. Carey and R. Diamond. From piecemeal to configurational representation of faces. *Science*, 195:312–314, 1977.

[3] S. Carey, R. Diamond, and B. Woods. Development of face recognition-a maturational component? *Developmental Psychology*, 16:257–269, 1980.

[4] R. Diamond and S. Carey. Why faces are and are not special: an effect of expertise. *Journal of experimental psychology*, 115(2):107–117, 1986.

[5] S. Edelman, D. Reisfeld, and Y. Yeshurun. Learning to recognize faces from examples. In G. Sandini, editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 787–791. Springer Verlag, 1992.

[6] M. Lando. A computational model of upright and inverted human face recognition. Master's thesis, Feinberg Graduate School of the Weizmann Institute of Science, Rehovot, Israel, November 1994.

[7] Y. Moses, S. Edelman, and S. Ullman. Generalization across illumination and orientation changes for inverted and upright faces. CS-TR 14, Weizmann Institute of Science, 1993.

[8] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[9] T. Poggio and T. Vetter. Recognition and structure from one 2d model view: observations on prototypes, object classes, and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[10] Y. Weiss and S. Edelman. Representation of similarity as a goal of early visual processing. *Network*, 6:19–41, 1995.

[11] H. R. Wilson and J. R. Bergen. A four mechanism model for threshold spatial vision. *Vision Research*, 19:19–32, 1979.