An attractor field model of face representation: Effects of typicality and image morphing

Marian Stewart Bartlett  
University of California San Diego  
La Jolla, CA 92093-0523  
marni@salk.edu

James W. Tanaka  
University of Victoria  
Victoria, BC V8W 3P5  
jtanaka@uvic.ca

Abstract

A morphed face image at the midpoint between a typical face and an atypical face tends to be perceived as more similar to the atypical parent than the typical parent [Tanaka et al., 1998]. One account of this atypicality bias in face perception is provided by the hypothesis that face representations are characterized as basins of attraction in face space. According to this hypothesis, atypical faces should have larger basins of attraction than typical faces since they are farther from the origin of face space where the density of faces is much lower. This hypothesis is tested on a set of grayscale face images, and normalized Gabor representations of the faces. We first examine some of the assumptions about face space made by the attractor field hypothesis, and demonstrate that typical faces are indeed closer to the origin of face space than atypical faces, and the density of faces is greater near typical faces than atypical faces, for the normalized Gabor representations. We next show that a face model based on principal component analysis, which has accounted for other face perception phenomena such as "other race" effects, does not account for the atypicality bias. Finally, an attractor network model of face representations is implemented, and the basins of attraction are examined by presenting morphed faces to the network. The model exhibits an atypicality bias when presented with morphed test patterns, and upholds the prediction that "atypical" patterns have larger basins of attraction than "typical" patterns stored in an attractor network.

Introduction

Familiar faces are recognized despite deviations in the face image such as those due to changes in age [Bruck et al., 1991], orientation [Hill et al., 1997], expression [Ekman et al., 1972], and caricaturization [Rhodes and Carey, 1987]. This many-to-one mapping from multiple face images to a single identity lead Tanaka and colleagues [Tanaka et al., 1998] to propose that face representations may be characterized as basins of attraction, in which the representation is activated by a broad range of inputs that fall within the boundary of the attractor basin. See Figure 2. According to the attractor field model, faces that are located farther away from the origin face space will have broader attractor basins than faces that are closer to the origin, where the density of faces is much higher. Faces rated as typical have been proposed to be closer to the mean of face space where the density of faces is high, whereas atypical faces are farther from the mean of face space where the density of faces is much lower.
space where the density of faces is more sparse [Valentine and Ferrara, 1991]. This characterization of face space is based on behavioral evidence such as the finding that atypical faces are also rated as more memorable and are recognized more quickly and accurately than typical faces [Light et al., 1979, Bartlett et al., 1984, Valentine and Bruce, 1986]. If atypical faces are indeed located in a region of face space that is sparse, the attractor field model predicts that they will have broader basins of attraction than typical faces.

Primary support for the attractor field hypothesis comes from psychophysical studies of the perception of morphed images. A morphed face image at the midpoint between a typical face and an atypical face tends to be perceived as more similar to the atypical parent than the typical parent, suggesting that the basin of attraction for atypical faces is indeed larger [Tanaka et al., 1998]. See Figure 1.

This paper explores the attractor field model of face space through a computational analysis of a set of face images. We first test assumptions about face space using both graylevel images and normalized Gabor representations. We investigate whether there are differences in stimulus density around faces rated as typical than around faces rated as atypical, and whether typical faces are closer to the origin of face space than atypical faces in either of these representational spaces. Next we explore the ability of a computational model of face perception based on principal component analysis (PCA) to account for the atypicality bias in face perception. Finally we implement an attractor network model of face perception, and test the atypicality bias in the attractor network.

The attractor network simulations employed a Hebbian learning model, which is among the most standard update rules for attractor networks [Hopfield and Tank, 1986]. Hebbian learning is an unsupervised learning rule proposed as a model for activity-dependent modification of the synaptic strength between neurons [Hebb, 1949]. In Hebbian learning, the synaptic strength is increased between simultaneously active units in proportion to the magnitude of the pre- and post-synaptic unit activities. This property can be implemented in a learning rule by incrementing the synaptic strength in proportion to the product of the pre- and post-synaptic unit activities. Incremental Hebbian learning in a fully connected network converges on the covariance matrix of unit activities across the set of training data. Support of the Hebbian learning model comes from studies of long-term potentiation of synaptic strengths in the mammalian hippocampus (e.g.

Figure 1: 50/50 morphs between typical and atypical faces are perceived to be more similar to the atypical parent than the typical parent on 65% of trials. From Tanaka, Giles, Kreman, & Simon (1998).
Hebbian learning has successfully modeled experience dependent organization of the visual system, from center-surround receptive fields in the retina and LGN to oriented simple cells [Linsker, 1986] and ocular dominance columns, including the effects of monocular deprivation [Miller et al., 1989]. Hebbian association of temporally contiguous input patterns has been shown to account for viewpoint invariant responses to faces [Bartlett and Sejnowski, 1998]. The modeling experiments presented here employ Hebbian learning to examine the extent to which experience dependent self-organization can account for the observed effects of typicality on face perception, and the influence of competition on the resulting representations. The PCA model employs Hebbian learning in a feedforward system. The attractor network employs Hebbian learning in a recurrent system with sustained activity.

Face Images

The face stimuli used in this study were identical to those used previously to demonstrate the atypicality bias perceptually [Tanaka et al., 1998, Tanaka and Cornielle, subm]. The faces were selected from college yearbooks. The selected faces contained no glasses, little hair covering the forehead, and a closed mouth. Faces were displayed in frontal or near-frontal viewing pose. Sixty-six female faces meeting the above criteria were Xeroxed, cropped around the face outline, and glued onto 8 1/2" x 11" sheets of paper (approx. 12 faces/page). Forty subjects rated the distinctiveness of the faces on a scale from 1 to 5, with 1 being non-distinctive and 5 being very distinctive. Subjects were instructed to rate the faces according to how distinctive looking they appeared relative to the general population.

From the original set of faces, eight female faces were selected on the basis of being either high or low in rated typicality. The four atypical female faces had a mean distinctiveness rating of 3.95 and the four typical faces had a mean distinctiveness rating of 2.10. The 8 faces were digitized using a MicroTech Z flatbed scanner and Adobe PhotoshopTM. Using the MorphTM 2.5 program, a 50/50 morph face was generated by averaging each atypical face with typical face (see Figure 3). This process yielded a total 16 female morphs. Key
points for facial features were kept constant, with 12 points on the mouth, 7 points on each eye, 6 points on the nose, 5 points on each eyebrow, and 22 points for the outline of the face. Occasionally, additional points were required to eliminate shadows on the forehead region.

The morphing procedure was equivalent to the process described by Beal and Keil [Beale and Keil, 1995]. In the morphing process, a Delaunay tessellation process was applied in which neighboring control points were connected to form non-crossing triangular regions on a planar surface. The triangular regions were optimized such that pixels within a given region were closer to the control points at the triangle’s vertices than to any other control points on the surface. Applying a warping algorithm, control points for the morph face were generated by moving 50% of the total distance along the vector that connected corresponding control points in parent face 1 \((x_1, y_1)\) and parent face 2 \((x_2, y_2)\). The locations of intervening pixels were linearly interpolated across the surface based on the position of the nearest control point (see [Wolberg, 1990]). A fade process was then employed in which the brightness values for each corresponding pixel were weighted according to the contribution of each parent image. In a 50/50 morph face, pixel values were determined based on the equal contribution of each parent image.

The computational image analysis was performed on the 8 typical and atypical female faces, the 16 morphed images, and an additional 23 female faces from the above set which were rated as neither strongly typical or strongly atypical. Mirror reflected images of all faces were generated to increase the number of training images and to remove any bias in head orientation, resulting in 62 face images for training models, excluding the morphs. The pixel brightnesses of the images were linearly rescaled to fill the full range of \([0, 255]\). The centers of the two eyes and the mouth were located in each image. Images were rotated in the plane so that the eyes were horizontal, and then scaled to a constant area of the triangle generated by the two eyes and the mouth.\(^1\) The images were then aligned on the left eye and cropped to 120 × 120 pixels.

\(^1\)This measure uses both interocular distance and eye-to-nose distance to equalize scale between faces, and is more consistent across individuals than interocular distance alone.
Experiment 1: Testing assumptions about face space

Image Representation: Normalized Gabor Energy

The characteristics of face space were examined for both the graylevel images and for a representation based on the outputs of Gabor filters. A Gabor filter is a 2-dimensional Gaussian modulated by a sine or cosine wave, and closely models the receptive fields of simple cells in primary visual cortex [Daugman, 1988]. A normalized Gabor energy representation based on [Heeger, 1991] was generated, and is illustrated in Figure 4. The representation consisted of the outputs of a bank of Gabor filters at four spatial scales and 8 orientations, sampled at 255 spatial locations. The outputs of sine and cosine Gabor filters were squared and summed, and then the contrast was normalized by dividing by the activity across all orientations and scales at each spatial location. This representation reduces variations due to differences in illumination and contrast, as well as small shifts in spatial position, and is an approximate model of the responses of V1 complex cells [Daugman, 1988, Heeger, 1991, Lades et al., 1993].

Physical similarity of morph to parent images

An immediate question is whether there is some property of the morphing process, or something about the selection of feature points, that causes the morphed images to be more physically similar to the atypical parent image than the typical parent images. We verified that the morphed faces were indeed equidistant from their typical and atypical parents for both the graylevel image representation, and the normalized Gabor representation. The mean Euclidean distance of the morphed face images to each of the parent images is shown in Figure 5. There was no significant difference for either the graylevel image representation or for the normalized Gabor representation.

Distributions of typical and atypical faces in face space

Two assumptions about the distributions of typical and atypical faces in face space were next investigated: 1. Are typical faces indeed closer to the origin of face space than atypical faces? and 2. Is the density of faces greater near typical than atypical faces?

Face space was assessed using the 62 female face images, excluding the morphs. To address
Question 1, the origin of the face space was estimated as the mean face representation across the set of 62 faces. Distances of typical and atypical faces to the mean face are shown in Figure 6 for the graylevel images and for the normalized Gabor representation. There was no significant difference in the distance to the mean face for the graylevel images. However, for the normalized Gabor representation, the typical faces were closer to the mean face ($t(14) = 2.3, p < 0.05$).

To address the density of face space near typical and atypical faces, the mean distance of each face to its nearest neighbor in face space was compared for typical and atypical faces. The comparison is shown in Figure 7. There was indeed a shorter distance to the first neighbor of a typical face than an atypical face for the raw graylevel image representation ($t(14) = 2.8, p < 0.05$), and the difference was more pronounced for the normalized Gabor representation ($t(14) = 4.3, p < 0.001$).

**Experiment 2: PCA reconstruction model**

We next demonstrate that a face model based on principal component analysis (PCA) does not account for the atypicality bias in the perception of morphed faces. PCA models are based on the correlations among the image pixels, and are also termed autoassociative networks since they reconstruct the input in the output (see [Valentin et al., 1994] for a review). Network recognition rates in PCA models, as measured by the similarity of a reconstructed test image to the original image, have been shown to correlate with human ratings of typicality, human recognition accuracy, and distance of the faces from the population mean, and has successfully modeled "other-race" effects in face recognition (e.g. [O’Toole et al., 1991, Hancock et al., 1996]). Can this simple feedforward model, which has provided an account for several other effects in human face perception, also account for the atypicality bias in the perception of morphed images?

Because such models represent dimensions of variability that are common to the population very accurately, whereas dimensions of variability that are not common to the population tend to be lost, we predict that reconstructions of the morphed face images will regress towards the population mean and be more similar to the typical parent. The PCA model would therefore be unable to account for the atypicality bias, and would in fact show the opposite effect.
Figure 6: Mean distance of typical and atypical faces to the origin of face space. The origin was defined as the mean face.

Method

The network was trained on an ensemble consisting of the 8 typical and atypical female faces, the 23 additional female faces, and their mirror reflections, for a total of 62 training images. Each face image comprised a datapoint in a 14400-dimensional space given by the grayvalue at each pixel location. The principal component axes of the dataset were calculated by finding the eigenvectors of the pixelwise covariance matrix. Network output was the principal component reconstruction of the image obtained by projecting the image onto the PCA subspace and then reconstructing it as a linear combination of the PCA eigenvectors. The PCA analysis was repeated on normalized Gabor representations of face images.

Let $X$ be a 14400x62 matrix of face images in columns and $\mu$ be a 14400x1 vector consisting of the mean face in $X$. The principal component basis vectors $P$ are the eigenvectors of the pixelwise covariance matrix

$$P = \text{eig}[(X - \mu)(X - \mu)^T].$$

The eigenvectors in $P$ were normalized to unit length. Let $M$ be a 14400x16 matrix of the 16 morph images. The PCA representation of $M$ (the coordinates with respect to the new basis vectors) was calculated as $R_M = (M - \mu) * P$, and the image reconstruction as $\hat{M} = R_M * P^T + \mu$. Hence the full reconstruction model is

$$\hat{M} = (M - \mu) * PP^T + \mu. \quad (1)$$

It has been shown that PCA can be implemented in simple neural networks with Hebbian learning [Oja, 1989, Singh, 1991]. Moreover, the PCA reconstruction model of Equation 1 is equivalent to a 1-layer feedforward network with weights given by the columns of $PP^T$. This is a Hebbian learning rule on $P$, the principal component eigenvectors, which can be seen by referring to the equation for a basic Hebb rule given in Equation 2, below.
Results and Discussion

After training the network on the ensemble of 62 face images, the 16 morphed face images were presented and the PCA reconstructions $\hat{M}$ were generated according to equation 1. A measure of network recognition of the morph images was provided by the Euclidean distance between the PCA reconstructions and the original parent images. Similarity was also measured using cosine of the angle between the reconstructed images and the original parent images.

For the PCA model trained on graylevel images, there was no significant difference in similarity of the morph reconstructions to the typical and atypical parents. (For Euclidean distance, $t(62)=.01, p=.99$; For cosines, $t(62)=.09, p=.93$.)

The PCA model trained on the normalized Gabor representations of images also did not exhibit an atypicality bias. Moreover, there was a trend towards the opposite effect, in which the morph reconstructions were closer in Euclidean distance to the typical parent ($t(62)=1.9, p=.06$). This trend is consistent with the prediction that PCA reconstructions would regress towards the population mean. The Gabor representation had substantially more pronounced typicality effects in the face space analysis than the original graylevel images. It makes sense, therefore, that a reverse typicality effect would emerge with the Gabor representation and not with the graylevel images.

A similar pattern of results was obtained for a PCA model consisting of the PCA coordinates $R_M$ instead of the image reconstructions $\hat{M}$. For a PCA model trained on graylevel images, there was no difference in the similarity of the projected coordinates of the morphs to their typical or atypical parents. The model trained on the normalized Gabor representations again showed a trend towards an opposite effect from the human data. The projected coordinates of the morphs were closer to the typical than the atypical parent ($t(62)=1.8, p=.07$).

Associative memory mechanisms based on pixel-wise correlations alone therefore do not account for the atypicality bias in human face perception. Another explanation is required.
Attractor network model

We next constructed an attractor network model of face representations. The network was first trained to produce distinct patterns of sustained activity for a set of typical and atypical faces. The attractor basin hypothesis predicted that faces rated as "typical" would have smaller basins of attraction than atypical faces due to the density of points near the population mean, whereas atypical faces would have larger basins of attraction due to the lower population density. The basins of attraction for each face were examined by presenting morphed faces to the network. If the attractor basins of the atypical faces were larger than those of the typical faces, the 50/50 morphs would be more likely to fall into the basin of attraction of the atypical parent than the typical parent. The similarity space of the attractor network representations was also evaluated by measuring the Euclidean distance between the sustained patterns of activity. We predicted that the network response to morphs would be more similar to the atypical parent than the typical parent, exhibiting an atypicality bias similar to that observed in human subjects.

We first examined an extension of a basic Hopfield network [Hopfield and Tank, 1986], with 128 units, in which every unit was connected to every other unit, and there were no self-connections. The weights were symmetric, which ensured that the attractors would all be fixed-points, meaning a stable pattern of activity rather than cycling between a set of activity patterns.

The Hebbian learning rule employed in the basic Hopfield attractor network is

\[ W = \frac{1}{N} X X^T \quad (2) \]

where \( X \) is the matrix of training patterns, in which the patterns are stored in columns, and \( N \) is the number of units in the network. The update rule for the activation \( V \) of unit \( i \) at time \( t \) is given by:

\[ V_i(t + \delta t) = \phi \left[ \sum W_{ij} V_j(t) - \theta \right] \quad (3) \]

Where \( \theta \) is a threshold and \( \phi(x) = 1 \) for \( x > 0 \), and 0 otherwise. In these simulations, \( \theta = 0.0 \).

For a set of orthogonal training patterns, a basic Hebbian learning rule stores the training patterns as fixed points, meaning that after training, when the network is presented with one
of the training patterns, it produces a sustained, fixed pattern of activity that is identical to the training pattern. Small perturbations of the training pattern will also settle to the learned fixed point, and the range of initial patterns that settle to a given fixed point is called its basin of attraction.

The training patterns in these simulations, however, are not orthogonal. The face patterns are highly correlated. The simple Hebbian learning rule in Equation 2, will not store correlated patterns as separate fixed points, and instead, correlated patterns will fall into a single basin of attraction, settling to a single fixed point.

In order to store correlated patterns in separate basins of attraction, competitive interactions are necessary. A simple solution to the problem of storing correlated patterns in an attractor network was presented by Kanter & Sompolinski (1987), and consists of inserting a decorrelating matrix, $Q^{-1}$, into the learning rule:

\[
Q = \frac{1}{N}X^TX \\
W = \frac{1}{N}XQ^{-1}X^T
\]  

This learning rule is termed the “pseudo-inverse” rule because $Q^{-1}X^T$ is the pseudo-inverse of the matrix of training patterns, $X$. The matrix $Q$ is closely related to the covariance matrix of the training patterns. In the case where each pattern has zero-mean and unit variance, $Q$ is identical to the covariance matrix and multiplication of $X$ by $\sqrt{Q^{-1}}$ removes the covariances from the data matrix $X$. The learning rule in Equation 5 is equivalent to a two-step process that first transforms the training data to remove the covariances, and then performs Hebbian learning on the set of decorrelated training patterns:

\[
W = \frac{1}{N}X\sqrt{Q^{-1}}\sqrt{Q^{-1}}^TX^T.
\]  

Note that this is a Hebbian learning rule of the form $W = \frac{1}{N}AA^T$ where $A = X\sqrt{Q^{-1}}$ is the decorrelated set of training patterns. Although we can piece apart the learning rule into two steps, the learning rule in Equation 5 carries out both steps with a single weight matrix. PCA is also a decorrelating system, in that the coordinates of each image with respect to the new basis vectors ($R_M$ from Equation 1), are decorrelated versions of the original images. (See [Valentin et al., 1994].) The primary difference between PCA and the attractor network, then, is the sustained activity in the recurrent connections. Note that the activation function contains a nonlinear term $\phi$. Here $\phi$ is a step function, but sigmoidal functions can also be employed in attractor networks [Amit, 1995].

**Experiment 3: Attractor network with simple inputs**

The attractor network model was first tested with a set of synthesized data which enabled the population density properties to be exaggerated while keeping the total number of units in the network relatively small. The training patterns consisted of binomial [-1,1] patterns. Eight “atypical” patterns were generated as random, binomial 128-dimensional vectors. These patterns differed from each other on 50% of units, on average. Eight “typical” patterns were generated from a single “mother” pattern by randomly flipping 16 units of the mother pattern. The “typical” patterns differed from each other on 13% of units, on average. These 16 patterns comprised the input to an attractor network with 128 units.

Not surprisingly, the simple Hebbian learning rule in Equation 2, failed to store the typical patterns as separate fixed points, and all typical patterns fell into a single basin of attraction, settling to a single fixed point, which was the mean of the typical patterns.
Table 1: Attractor network response to 50/50 morphs of synthetic data. Left: Number of times the morph pattern was inside the basin of attraction for the typical or atypical parent pattern. Center: Number of times the morph pattern was closer to the typical or atypical parent pattern. Right: Mean distance of the sustained pattern of activity in for the morph inputs to that of the typical and atypical parent patterns.

The pseudo-inverse learning rule successfully stored the eight typical and eight atypical training patterns in separate basins of attraction. After training, 50/50 morphs between the typical and atypical training patterns were presented to the network. A set of 32 morph patterns was generated by taking means between typical and atypical patterns. Each of the morph inputs settled to a fixed point in the attractor network, and the fixed point for the morph pattern was compared to the fixed point for the typical and atypical parent patterns. The results are shown in Table 1. 24 of the 32 morph patterns fell into the basin of attraction for the atypical parent, meaning that the activity settled to the same fixed point. In contrast, only two of the morph patterns fell into the basin of attraction of the typical parent. The remaining 6 morph patterns settled into a distinct activity pattern that differed from both that of both parents. Of these 6, four were closer in Euclidean distance to the activity pattern for atypical parent, and two were closer to the activity pattern for the typical parent. Overall, the mean distance was smaller to the atypical parent than to the typical parent, (2.9 versus 14.3, t(62) = 9.5, p < 0.001).

Figure 9: Mean distance of morph fixed points to the typical and atypical parent fixed points for a sequence morphs on 0.1 intervals in Experiment 3.

The attractor network boundaries were next examined by presenting morphed patterns at 10% intervals between the atypical pattern and the typical pattern. The network responses are shown in Figure 9. For both typical and atypical parent patterns the basin of attraction extended to at least 60/40 mixtures. This network showed a form of categorical perception,
in which a range of inputs elicited identical responses. The mean basin of attraction was larger for the atypical parent patterns, with 75% extending to include 50/50 morph mixtures. The cross-over point was at approximately 45/55 mixtures of atypical and typical parents.

Experiment 4: Attractor Network Model with Face Images

The attractor network model described in Experiment 3 was next trained on real face image data. The face representation consisted of the normalized outputs of Gabor filters. This representation is 14400-dimensional, and a network that took the complete representation as input would require $14400^2$ weights. A lower dimensional representation was therefore generated by choosing the 1000 Gabor filter outputs with the highest variance. A 1000-dimensional Hopfield network was trained on 16 patterns consisting of eight typical faces and eight atypical faces. Each of the 16 training images settled to a fixed point. Because the training patterns were not binary [-1,1], the fixed points were not identical to the training patterns.

The network was then tested by presenting the image representation of the 32 female 50/50 morph faces (16 combinations of 4 typical and 4 atypical faces, plus mirror reversed images). The sustained activity pattern for each of the morph images was compared to that for its two parent images. The results are summarized in Table 2. The morphed face fell into the basin of attraction for the atypical parent face 6 times, whereas it never fell into the basin of attraction for the typical parent face. This difference in success rates was statistically significant ($Z=2.6, p=.01$). The remaining 26 morphed faces settled to a distinct pattern of sustained activity that differed from both parents. Of these, 22 were closer to the atypical than typical parent, two were equidistant, and two were closer to the typical than atypical parent. This difference in rates was highly significant ($Z=5.6, p<.0001$). Mean distance of the 32 morph faces to the typical parent was 36.0 compared to and 24.2 for the atypical parent ($t(62) = 3.9, p < .001$). Overall, the attractor network would have selected the atypical parent 87.5% of the time based on Euclidean distance of the sustained activity patterns to the parent patterns. Thus the attractor network model with real face images exhibited an atypicality bias in the sustained activity patterns for the morphed faces.

<table>
<thead>
<tr>
<th>No. in parent basin</th>
<th>No. morphs closer to parent</th>
<th>Mean dist. to parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical 0/32</td>
<td>Atypical 6/32</td>
<td>Typical 2/32</td>
</tr>
</tbody>
</table>

Table 2: Attractor network response to 50/50 morphs of real face images in Experiment 4. Left: Number of times the morph pattern was inside the basin of attraction for the typical or atypical parent pattern. Center: Number of times the morph pattern was closer to the typical or atypical parent pattern. Right: Mean distance of the sustained pattern of activity in for the morph inputs to that of the typical and atypical parent patterns.

The attractor network boundaries were next examined by presenting morphed faces with 35/65 and 65/35 mixtures of typical and atypical parents. The results are shown in Figure 10. Like the 50% morphs, the 35% and 65% morphs settled to sustained activity patterns that were closer to the typical than atypical parent. The cross-over point was beyond the 65% morph.
Figure 10: Attractor network boundries for the network trained on real face images in Experiment 4. Mean distance of the activity patterns for the morphed faces to those for the typical and atypical parents are shown for morphs containing 0, 35, 50, 65, and 100% of the typical parent.

**Discussion**

This paper explored an attractor network model of face representations proposed by Tanaka et al., 1998 on a set of graylevel face images. This model proposes that face representations are characterized by basins of attraction, in which the representation is activated by a broad range of inputs that fall within the boundary of the attractor basin. This hypothesis is consistent with the many-to-one mapping between face images and identity. The primary empirical support for this hypothesis comes from the atypicality bias in the perception of morphed faces. According to the attractor field model, atypical faces will have a broader attractor basin than typical faces, which are proposed to be closer to the origin of face space where the density of faces is much higher. Perceptual studies show that a 50% morph between typical and atypical parent faces tends to be judged as more similar to the atypical parent, suggesting that the basin of attraction for atypical faces is indeed larger. This paper specifically investigated the atypicality bias in an attractor network model as well as in a model based on principal component analysis.

The attractor network model makes several assumptions about face space. These assumptions were supported on a set of face images. For an image representation based on the outputs of Gabor filters, typical faces were indeed closer to the population mean than atypical faces, and the density of face images was indeed higher near typical faces than atypical faces. Next, we demonstrated that a reconstruction model based on principal component analysis did not account for the atypicality bias. The PCA model showed the opposite effect, in which the PCA representations of 50% morph images were closer the typical rather than atypical parent. Principal component analysis is a form of feedforward Hebbian learning [Oja, 1989, Singh, 1991], whereas the attractor network model is a recurrent Hebbian system.

We next presented an attractor network model, and demonstrated that the atypicality bias observed in human face perception is also exhibited in the attractor network model. Atypical faces had a larger basin of attraction in this model than typical faces. A small percentage (19%) of the morphed face images elicited the same sustained response pattern as the atypical parent face through the 50% morph point. Most morphed face images (69%) settled to an activity pattern that differed from both parents, yet was closer to the atypical parent.
This is consistent with a percept of greater facial similarity between the morph and the atypical parent, without the perception that the morph face shared the same identity. The attractor field hypothesis predicts categorical perception of face images within the basin of attraction of a given stored face, in which a range of input patterns elicit the same stored representation. While some of the faces showed categorical effects through 50/50 morphs, most of the face images showed a categorical effect of a slightly different nature, in which a range of face images between 35% and 65% morphs elicited a consistent representation which was closer to but not identical to the atypical parent, as shown in Figure 10.

The attractor field hypothesis is consistent with Krumhansl’s observations on the effects of exemplar density on similarity judgments, in which higher stimulus density is associated with lower perceived similarity given equal physical differences [Krumhansl, 1978]. The attractor network offers one possible model for how stimulus density affects perceived similarity, in which higher density regions of the stimulus space have smaller basins of attraction. A related account based on information theory is explored in [Bartlett, ress]. Note that the perceptual magnet effect [Kuhl, 1991] would make the opposite prediction. The magnet effect would predict that faces judged as particularly good exemplars would draw neighboring faces closer in face space. Hence a 50% morph would be judged to be more similar to the typical than atypical parent. This is consistent with the PCA model of experiment 2. Thus learning the stimulus correlations in a feedforward system leads to behavior consistent with the perceptual magnet effect, whereas the attractor network leads to behavior consistent with the density hypothesis. Both PCA and the attractor network are decorrelating systems. The primary difference between the two is the sustained activity in the recurrent connections of the attractor network.
References


