Eigenfaces for Familiarity

Matthew N. Dailey (MDAILEY@CS.UCSD.EDU)
Garrison W. Cottrell (GARY@CS.UCSD.EDU)
Computer Science and Engineering Department
University of California, San Diego
9500 Gilman Dr., La Jolla CA 92093-0114 USA

Thomas A. Busey (BUSEY@INDIANA.EDU)
Department of Psychology
Indiana University
Bloomington, IN 47405 USA

Abstract

A previous experiment tested subjects' new/old judgments of previously-studied faces, distractors, and morphs between pairs of studied parents. We examine the extent to which models based on principal component analysis (eigenfaces) can predict human recognition of studied faces and false alarms to the distractors and morphs. We also compare eigenface models to the predictions of previous models based on the positions of faces in a multidimensional "face space" derived from a multidimensional scaling (MDS) of human similarity ratings. We find that the error in reconstructing a test face from its position in an "eigenface space" provides a good overall prediction of human familiarity ratings. However, the model has difficulty accounting for the fact that humans false alarm to morphs with similar parents more frequently than they false alarm to morphs with dissimilar parents. We ascribe this to the limitations of the simple reconstruction error-based model. We then outline preliminary work to improve the fine-grained fit within the eigenface-based modeling framework, and discuss the results' implications for exemplar- and face space-based models of face processing.

Introduction

The errors that subjects make during face recognition tasks may hold the key to improving our understanding of the representations and mechanisms underlying face processing and visual perception. One effective way of evoking such errors is testing subjects' recognition of studied faces that have been combined in some way (e.g. Solso and McCarthy, 1981; Reinitz, Lammers, and Cochran 1992). In a recent series of behavioral and modeling experiments, Busey and Tunnicliff (submitted) have examined the effects of facial blending and distinctiveness on subjects' tendencies to make recognition errors.

Their experiments used facial images of bald males and morphs between pairs of these images (see Figure 1) as stimuli. In an earlier study, Busey (in press) had subjects rate the similarity of all pairs in a large set of faces and morphs, then performed a multidimensional scaling (MDS) of these similarity ratings to create a 6-dimensional "face space" (Valentine and Endo, 1992). Busey interpreted the resulting dimensions as representing 1) age, 2) race, 3) facial adiposity, 4) extent of facial hair, 5) head aspect ratio, and 6) hair color (shading). He also found that in general, a carefully-created morph lies near the average of its two "parents" in MDS space, with a systematic bias toward youth (dimension 1) and adiposity (dimension 3).

In "Experiment 3" (Busey and Tunnicliff, submitted), 179 subjects studied 36 target faces, 8 similar pairs of parent faces (16 images), and 8 dissimilar pairs of parent faces (16 images), for a total of 68 studied images.[1] All facial images were of bald men, and none of the parents used for morphing had facial hair. The subjects were asked to study the images and remember them for a recognition test. In the test phase of the experiment, the 16 morphs between parent pairs and 20 new distractors were added to the pool of stimuli. At test time, each subject was presented with 8 of the morphs, one parent of each of the unseen morphs (8 parents), the 36 targets, and the 20 distractors. They were asked to make old/new judgments of each of the test stimuli. The results of this experiment were extremely interesting. For many of the morphs, subjects responded "old" to the unstudied morph more often than to its studied parents. On a finer scale, subjects were more likely to respond "old" to the similar morphs than to those morphs' parents, and less likely to respond "old" to the dissimilar morphs than to those morphs' parents (see "Experiment 3 Data" in Table 1). One theoretical explanation is that the similar parents are so similar to their "child" morph that memories of both contribute toward an "old" (false alarm) response to the morph.

Busey and Tunnicliff (submitted) used the 6-dimensional MDS scores for each of the stimuli in a series of models tuned to predict the probability of the humans responding "old" to each image (hereafter referred to as P(old)). They applied two alternative versions of GCM, the Generalized Context Model (Nosofsky, 1986), and "SimSample," a model based on the assumption that a test face results in sampling a similar face from memory and a response of "old" if the similarity between the probe and sample are above some threshold. The best-fitting GCM models are surprisingly poor predictors of the data, but the SimSample models perform much better. To achieve a strong quantitative fit to the mean human responses for each of the six stimulus categories (targets, distractors, similar parents, similar morphs, dissimilar parents, and dissimilar morphs), however, Busey and Tunnicliff had to introduce a prototype mechanism into the SimSample framework. The prototypes, situated at the locations of the morphs in MDS space and weighted by the similarity of their parents, are assumed to be the result of a similarity-dependent blend-

[1] The similarity of each parent pair was determined by human subjects in a pilot study.

Figure 1: Three normalized morphs from the database.

ing or abstraction mechanism.

In this paper, we apply an alternative flavor of model to the Experiment 3 data, based on the quality with which a test image can be reconstructed after it is "compressed" by projecting it onto a subset of the principal component eigenvectors (eigenfaces) of the studied face set. Eigenfaces, which are the orthogonal axes along which the study data vary the most, can be computed with a simple procedure, principal components analysis (PCA). Models of this type assume that rather than storing the studied exemplars explicitly, subjects construct a low-dimensional manifold containing (with some error) the representations of the studied stimuli. The probability a subject responds "new" to a test stimulus, then, is a monotonic function of the model's reconstruction error (distance to the manifold). In this view, reconstruction error models the "novelty" of a test stimulus with respect to the study set, and its inverse, reconstruction quality, models the "familiarity" of a test stimulus with respect to the study set (Kohonen et al., 1977; O'Toole, Millward, and Anderson, 1988). Metcalfe, Cottrell, and Mencl (1992) have shown that a nearly identical mechanism (reconstruction error in the autoencoding backpropagation network, which essentially performs PCA on the study set) is a good candidate model for the explicit memory task of cued word recall. Past successes like these motivated us to determine the PCA reconstruction error model's ability to account for the subjects' new/old judgments in Experiment 3.[2]

An eigenface-based familiarity model is an appealing alternative to a model based on facial positions in MDS face space for several reasons:

- *"Unsupervised" representations:* Eigenface-based representations are not dependent on human judgments. For MDS face space approaches, all the stimuli must be simultaneously subjected to an experiment with human observers to determine where the new faces lie in face space. Adding new faces to the study set in a PCA model simply requires a new PC analysis; adding new faces to the test set requires no effort.

- *Underlying mechanisms:* MDS face space exemplar techniques only model the mechanisms underlying face recognition indirectly. But PCA models are *processing* models; they actually implement the process of obtaining a familiarity rating directly from the stimulus. First, several neural network architectures using Hebbian learning are capable of learning the principal components of a training set (Diamantaras and Kung, 1996). Second, many of the cells in monkey temporal cortex are sensitive to various aspects of face stimuli (Perrett et al., 1992). Thus the PCA model is a biologically plausible candidate abstraction of the mechanisms involved in familiarity judgments.

- *Free parameters:* PCA models have few free parameters, and these parameters can often be set in a principled manner independent of the specific effects being sought. For instance, we can fit the range of eigenfaces used for projection to the human responses for the studied stimuli only, and examine how the model generalizes to the novel test data (a one-parameter fit).

We found that a simple model based on PCA reconstruction error provides a fairly good fit to the overall P(old) ratings from Experiment 3, effectively separating targets from distractors. The model often exhibits a "morph familiarity inversion effect," in which a morph is judged as more familiar than one or both of its parents. However, it cannot account for the fine structure of the human familiarity ratings; in particular, it predicts more frequent false alarms to the morphs with dissimilar parents than to the morphs with similar parents, just the opposite of the pattern in the human data. In the discussion section, we outline the simple reconstruction error model's fundamental limitations in accounting for the similar/dissimilar parent effect and outline our preliminary attempts to improve on our models' fit within the eigenface modeling framework. We then discuss our results' implications for face space- and exemplar-based models of face processing.

Experimental Methods

This section details the methods we applied to modeling the human data from Busey and Tunnicliff (submitted) Experiment 3. We normalized the face images from the experiment to make them amenable to computational analysis and then performed a principal components analysis on the set of images the subjects studied. We then interpreted the ability of the model to reconstruct a studied or novel image as a measure of the image's familiarity.

[2]Since the term "recognition" is overloaded with easily confused meanings, we will use the term "familiarity" (with respect to the study set) to refer to the probability a subject responds "old" to a test stimulus. We do not make any assumptions as to whether a subject "recognizes" or "identifies" individuals at test time.

Figure 2: First five eigenfaces of the studied image set.

Face Data

Figure 1 shows three of the morphs between "parent" pairs in the database (the original images are copyrighted and cannot be published). The original images were 104 digitized 560x662 grayscale images of bald men, with consistent lighting and background, and fairly consistent position. The subjects varied in race and facial hair.

Normalization

We used eye templates to automatically locate the left and right eyes of each face image and then translated, rotated, scaled, and cropped all images so that the individuals' eye positions were in the same location. We did not, however, normalize the position of the mouth in each image. Since the ratio of a head's width to its height was a significant factor in rating the similarity between two of these stimuli (Busey, in press), we instead scaled the images by the same amount in both the horizontal and vertical directions. The images were scaled to 115x143 pixels to make the principal component analysis tractable on a workstation. The morph images in Figure 1 are examples of the result of this process.

Principal Components Analysis

Principal components analysis (PCA) is a technique that extracts the orthogonal axes along which a data set varies the most by computing the eigenvectors and eigenvalues of the data's covariance matrix. When applied to facial images, these eigenvectors are often called "eigenfaces." Using Turk and Pentland's (1991) efficient technique, we computed the eigenvectors of the covariance matrix of the 68 images used for study in Busey and Tunnicliff's Experiment 3. Figure 2 shows the five most significant eigenfaces for this dataset.

PCA Reconstruction Error as Novelty

To model the old/new judgments in the testing phase of Experiment 3, we projected the 68 studied images, the 18 distractors, and the 16 morphs between studied faces from image space onto the subspace defined by the first k eigenfaces of the study image set:

$$\mathbf{p}_i = \mathbf{A}\mathbf{x}_i \quad (1)$$

where $\mathbf{x}_i$ is the i-th input image represented as a 16,445-element column vector and $\mathbf{A}$ is a matrix formed from the k unit-length row eigenvectors with the highest eigenvalues. We then computed each image's reconstruction by projecting $\mathbf{p}_i$ back to image space:

$$\hat{\mathbf{x}}_i = \mathbf{A}^T\mathbf{p}_i$$

and then computed its reconstruction error:

$$err_i = ||\mathbf{x}_i - \hat{\mathbf{x}}_i||$$

We treated k as a free parameter, and for each of its possible values from 1 to 68, we computed the err_i measure for all 104 test images used in Experiment 3. As k approaches 68, the reconstruction error for the studied images approaches 0. But with intermediate values of k, we can interpret the reconstruction error err_i as a measure of novelty, the inverse of familiarity (Kohonen et al., 1977; Metcalfe et al., 1992; Pomerleau, 1993).

Generally speaking, an unstudied image $\mathbf{x}_i$ that is not similar to the studied images will have a large err_i. On the other hand, an unstudied image that is similar to study images will have a lower reconstruction error. Thus we expect the reconstruction quality to provide a good model of the probability a subject will respond "old" to a previously-studied face or false alarm to an unstudied face that is similar to some of the studied faces. By the same token, however, the reconstruction error model may not account as well for high hit rates for the most distinctive studied faces. The next section shows how well this simple k-PC model can account for the Experiment 3 data.

Results

Reconstruction error

As expected, average reconstruction error for the test images decreases with the number of eigenfaces used in the projection. For the studied images, reconstruction error decreases to 0 when 68 eigenfaces are used. For the unstudied images, the average reconstruction error decreases much more slowly: with 68 eigenfaces, it is approximately 60% of the error obtained using one eigenface.

We chose the value of free parameter k, the number of eigenvectors for projection, as the value for which the PCA model's rankings of the familiarity of the studied images (as measured by the ranking induced by negative err_i) best fit the human subjects' ranking of the familiarity of the studied images. Figure 3 illustrates how well the reconstruction error ranking correlates with the human subjects' familiarity ranking of the studied images, using Kendall's τ (a) (Kendall and

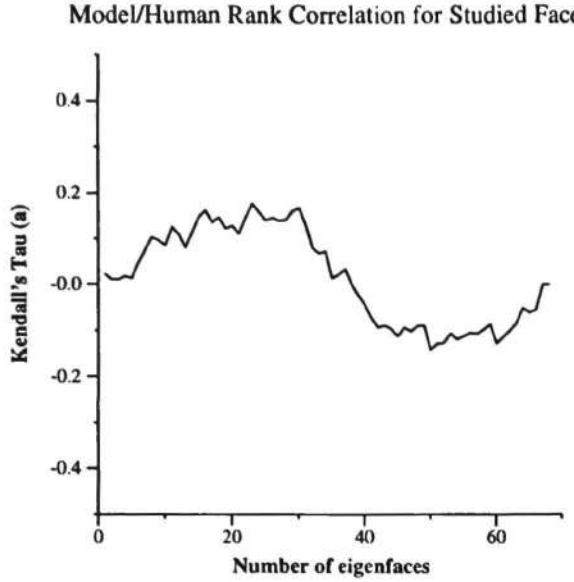

Figure 3: Study set rank correlation between model's predicted familiarity rankings and the human familiarity rankings, as a function of the number of eigenfaces used for projection.

Gibbons, 1990) as the rank correlation measure. The best correlation, with τ (a) = 0.177, occurred at $k = 23$.

We then determined how well reconstruction error in the 23 PC model predicts the probability of a human subject responding "old" to a test face in Busey and Tunnicliff's Experiment 3. The results for the individual test images are presented in Figure 4. We transformed the reconstruction error into a prediction of the human subjects' probability of responding "old" using a function of the form

$$pred_i = c_1(e^{-c_2 \times err_i} - c_3),$$

with parameters c_1, c_2, and c_3 fit to minimize the root mean squared error (RMSE) over all of the test images. (This function simply provided a better mapping from reconstruction error to familiarity than did a linear function.) The resulting RMSE was 0.169 and $r^2 = 0.315$. This fit is a large improvement over Busey and Tunnicliff's fit of the Generalized Context Model (GCM), which had a RMSE of 0.271. However, the reconstruction error model does not outperform either of their SimSample models, which had RMSE 0.148 and 0.141. At the same time, though, it is a surprisingly good fit considering the low number of free parameters and the fact that it is based directly on the image pixel data, without relying on human similarity ratings.

Although we did find that the model often predicted a higher P(old) for a morph than one or both of its parents, unfortunately, it does not capture the most interesting aspect of the human familiarity ratings: that human subjects tend to false alarm to morphs with similar parents, but not to morphs with dissimilar parents. In fact, on average, the model predicts the opposite. Table 1 compares the 23-PC model with the human ratings and Busey and Tunnicliff's SimSample fits. The next few subsections describe our attempts to build better models based on PCA techniques.

Higher dimensional reconstruction error

We have made several preliminary attempts to improve the reconstruction-based model's fit, with limited success. We found that projection and reconstruction using an *intermediate* range of principal components (e.g. projection and reconstruction using PCs 3–23) can improve the model's fit to the morph and parent data. One might expect this based on O'Toole, Deffenbacher, Valentin, and Abdi's (1994) observation that the most significant eigenfaces typically encode intergroup differences such as race and gender rather than subtle within-group differences that make faces distinctive or typical. However, at these parameter settings, the model still predicts the most false alarms to the morphs with dissimilar parents, rather than those with similar parents.

Nonlinear autoencoder networks

We have also attempted to improve our model fitting by experimenting with nonlinear autoencoding neural networks. It is well-known that in an autoencoding backpropagation network with a k-node linear "bottleneck" hidden layer, the trained network's hidden unit weight vectors will span the k-dimensional principal component subspace corresponding to the covariance matrix of its training set, effectively implementing PCA (Baldi and Hornik, 1989; Cottrell and Munro, 1988). However, as Japkowicz, Hanson, and Gluck (submitted) have observed, autoencoders with nonlinear hidden layers tend to fit more complicated nonlinear reconstruction error surfaces to their training data. Thus reconstruction quality in such networks can sometimes provide better models of novelty and familiarity, depending on the application domain. In principle, then, a nonlinear autoencoder could possibly account for the low false alarm rate for the dissimilar-parent morphs and the high false alarm rate for the similar-parent morphs due to a larger reconstruction error for the dissimilar-parent morphs. To test this hypothesis, we built several nonlinear autoencoder reconstruction error models on the study face set. Although their fit to the overall human familiarity data can compare favorably with the PCA models (e.g. Kendall's τ (a) of 0.28 for a 20-hidden node network, compared to 0.31 for the 23-PC model), their familiarity rankings for the morphs and parents are typically uncorrelated or even negatively correlated with the humans. Thus the nonlinear autoencoders do not merit further consideration in this domain.

A PCA projection exemplar model

In another attempt to account for the high false alarm rate for morphs with similar parents, we fit a simple exemplar model to the data using projections of test images onto the space formed by a range of principal component eigenvectors. The model is identical to the Generalized Context Model (Nosofsky, 1986), except we do not incorporate variable attentional weights. It assumes that the study faces' projections (the $\mathbf{p}_i$ in Equation 1) are placed in memory, and that the familiarity of a test face is the sum of the similarity of that face to each of the stored exemplars. That is, familiarity is defined as:

$$f_i = \sum_{j \in studyset} \eta_{i,j}$$

where $\eta_{i,j}$ is the similarity of face i to face j:

$$\eta_{i,j} = e^{-cd_{i,j}}$$

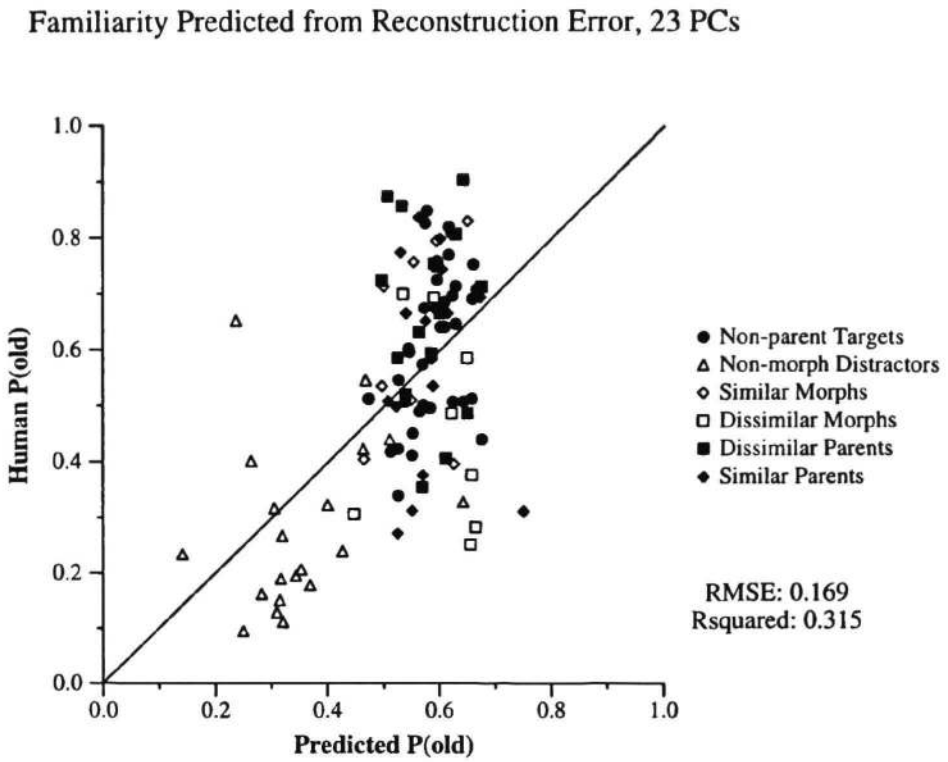

Figure 4: Fit of the 23-PC reconstruction error model to human subjects' probability of responding "old" to a test face.

and $d_{i,j}$ is the Euclidean distance from image i's projection to image j's projection. By hand-tuning this model, we found parameter settings that roughly fit the humans' high false alarm rate to the morphs with similar parents and low false alarm rate to the morphs with dissimilar parents, but the rank order of the response averages was incorrect, and it failed to fit the overall familiarity ratings as well as the reconstruction error model. This model has an RMSE of 0.193 and an r^2 of 0.110. For comparison with the other models, the results are reported in the last column of Table 1.

Discussion

The apparent failure of PCA reconstruction error to predict the high false alarm rate to similar-parent morphs and the low false alarm rate to dissimilar-parent morphs is simple to interpret in retrospect. When a morph is projected into PC space, its projection will be near the average of the projections of its parents. When its parents are dissimilar, the morph's projection will very likely be closer to the "center" of the PC space (more typical) than either parent (on average 37% closer for this data set), so the dissimilar-parent morphs will often be better reconstructed than their parents. When the morph's parents are similar, however, the morph is not much more typical than either parent (only 19% closer to the origin), and since the parents were in the study set, the model will most likely reconstruct the parents better than the morph.

Despite this limitation of the PCA reconstruction error model, our results show that it can provide a fairly good fit to the overall human familiarity ratings. In particular, it outperforms Nosofsky's GCM applied to the MDS face space by a large margin (though it does not outperform the Identification variant of the GCM). This is surprising considering that the MDS representation was derived directly from human similarity ratings, presumably exploiting the higher-level perceptual mechanisms of the humans making the similarity judgments. It is also surprising that the PCA exemplar-similarity model, which is equivalent to the GCM without variable attentional weights, performed no worse than the MDS GCM model. Taken together, these two results indicate that the position of a face in MDS space may not be the best representation for modeling familiarity judgments in new/old experiments. A possible explanation is that the human subjects in the similarity experiments "post process" low-level representations of the stimuli to the point that the similarity ratings no longer adequately reflect the underlying data they were derived from. Of course, it is difficult to make this claim given that Busey and Tunnicliff's SimSample models make better familiarity predictions than either of the PCA models. We will examine the issue more closely in future work.

Future Work

The modeling results we have obtained thus far point in two main directions for further research. First, since it has a history of success in modeling old/new judgments of faces, we cannot reject PCA reconstruction error as a predictor of novelty prematurely. One potential problem in the current model is that the eigenface decomposition was only performed on the studied faces. This is certainly a common approach in modeling memory tasks, but it ignores possible biases due to the subjects' prior experience. Perhaps the true effect of studying a set of novel faces is to somehow bias a preexisting set of memories toward the new faces rather than to store them in isolation. We can potentially model this process within the PCA reconstruction error framework by perform-

Condition	Experiment 3 Data	SimSample Predicted P(old)	SimSample + Prop. Prot. Predicted P(old)	23-PC Recon. Error Predicted P(old)	PC Projection Similarity
Dissimilar Parents	0.665	0.628	0.625	0.581	0.539
Similar Morphs	0.619	0.521	0.632	0.553	0.552
Targets	0.611	0.623	0.604	0.589	0.568
Similar Parents	0.578	0.604	0.585	0.580	0.543
Dissimilar Morphs	0.462	0.413	0.470	0.601	0.529
Distractors	0.280	0.323	0.303	0.348	0.486
RMSE		0.148	0.141	0.169	0.193

Table 1: Model fits to human data for each of the six stimulus types used in Busey and Tunnicliff's Experiment 3. Their best MDS exemplar-based model, SimSample, fits the overall human P(old) fairly well, but cannot account for high false alarm rate to morphs with similar parents. Their best model, "SimSample + Proportional Prototypes," fits the category means and overall human P(old) quite well, but requires the addition of prototypes at the morphs' positions in MDS space, weighted by their parents' similarity.

ing the eigenface decomposition on a larger set of faces that includes faces not used in Experiment 3 as well as the studied faces.

The other main thrust of our current and future research is to develop improved exemplar models and representations for comparison with the reconstruction error model and Busey and Tunnicliff's (submitted) exemplar models. We have constructed an exemplar-based model in MDS space that provides a much improved fit to the human data by modulating the exemplar similarity function's width and height by a face's distinctiveness. We are currently experimenting with image-based representations in the context of this new model.

Acknowledgments

We would like to thank Chris Vogt and anonymous reviewers for valuable comments on a previous version of this paper.

References

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2:53–58.

Busey, T. A. (in press). Where are morphed faces in multidimensional face space? *Psychological Science*.

Busey, T. A. and Tunnicliff, J. (submitted). Accounts of blending, distinctiveness and typicality in face recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Cottrell, G. W. and Munro, P. (1988). Principal components analysis of images via back propagation. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, Cambridge, MA. SPIE.

Diamantaras, K. and Kung, S. (1996). *Principal Component Neural Networks*. Wiley, New York.

Japkowicz, N., Hanson, S., and Gluck, M. (submitted). Nonlinear autoencoding is not equivalent to PCA.

Kendall, M. and Gibbons, J. (1990). *Rank Correlation Methods*. Edward Arnold, London, 5 edition.

Kohonen, T., Lehtio, P., Oja, E., Kortekangas, A., and Makisara, K. (1977). Demonstration of pattern processing properties of the optimal associative mappings. In *Proc Intl. Conf. on Cybernetics and Society*, Washington, D.C.

Metcalfe, J., Cottrell, G. W., and Mencl, W. E. (1992). Cognitive binding: A computational-modeling analysis of the distinction between implicit and explicit memory systems. *Journal of Cognitive Neuroscience*, 4:289–298.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 116(1):39–57.

O'Toole, A., Deffenbacher, K., Valentin, D., and Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, 22(2):208–224.

O'Toole, A., Millward, R., and Anderson, J. (1988). A physical system approach to recognition memory for spatially transformed faces. *Neural Networks*, 1:179–199.

Perrett, D., Hietanen, J., Oram, M., and Benson, P. (1992). Organisation and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London*, 335:23–30.

Pomerleau, D. A. (1993). Input reconstruction reliability estimation. In *Advances in Neural Information Processing Systems 5*, pages 279–286. San Mateo: Morgan Kaufmann.

Reinitz, M., Lammers, W., and Cochran, B. (1992). Memory-conjunction errors: Miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition*, 20(1):1–11.

Solso, R. L. and McCarthy, J. E. (1981). Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, 72(4):499–503.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *The Journal of Cognitive Neuroscience*, 3:71–86.

Valentine, T. and Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *The Quarterly Journal of Experimental Psychology*, 44A(4):671–703.