Max–Planck–Institut
für biologische Kybernetik
Arbeitsgruppe Bülthoff

Technical Report No. 9 December 1, 1994

Face Recognition across Large Viewpoint Changes

Alice J. O'Toole, Heinrich H. Bülthoff, Nikolaus F. Troje, Thomas Vetter

Abstract

We describe a computational model of face recognition that makes use of the overlapping texture and shape information visible in different views of faces. The model operates on view dependent data from three-dimensional laser scans of human heads, which provided three-dimensional surface data as well as surface image detail in the form of a texture map. View-dependent information from these surface and texture representations was registered onto separate three-dimensional head models. We used an auto-associative memory model as a pattern completion device to fill in parts of the head from a learned view when a test view with partially overlapping information was used as a memory key. We show that the overlapping visible regions of heads for both surface and texture data can support accurate recognition, even with pose differences of as much as 90 degrees (full face to profile view) between the learning and test view.

Alice O'Toole gratefully acknowledges support by the Alexander von Humboldt Stiftung and the hospitality of the Max-Planck Institut für biologische Kybernetik. We thank also Dan Kersten and Larry Maloney for very helpful discussions and Larry Maloney for comments on an earlier draft of this paper.

This document is available as `/pub/mpi-memos/TR-9.ps.Z` via anonymous ftp from `ftp.mpik-tueb.mpg.de` or by writing to the Max–Planck–Institut für biologische Kybernetik, Spemannstr. 38, 72076 Tübingen, Germany.

1 Introduction

A number of recent computational models of face recognition and analysis have relied primarily on face encodings derived from an image-based representation of a single view of a face [3, 4, 10, 11, 12]. The primary advantage of an image-based representation is that it eliminates the need to select and extract a specialized facial feature set for describing/representing faces. Additionally, with such representations, information about subtle shape and texture variations in the faces is retained and can be used for recognition. This kind of information is frequently discarded when preselected facial feature sets are used.

The primary limitation of image-based representations is that they are not optimally suited for recognizing faces across transformations that result in large changes in the image-based information. In other words, models that use these kinds of representations are able to recognize novel instances of faces only insofar as their image-based codings are similar to a learned/stored exemplar of the face. One important case in which this becomes highly problematic is when novel instances of stored faces differ in pose or viewpoint from the stored exemplar.

In recent years, clever elaborations of image-based codes have been developed to deal with this problem. For example, Lades, et al. [8] implemented a dynamic link architecture that operates on an elaborated image based code consisting of a series of orientation-selective *Gabor jet* filters. These jets sample the image at regular intervals on the vertices of a lattice and are "elastically" connected to their neighbors. The Gabor jet centers can deform to fit a novel instance of a face. A match cost is computed as a function of the quality of the filter match and a term penalizing lattice distortion. Lades et al. achieved excellent recognition performance for faces rotated 15 degrees from the original orientation. While it is likely that this model can be extended to handle larger viewpoint changes, it is unlikely that it will be extendable easily to viewpoint changes exceeding 25-30 degrees. This is due to the fact that substantial portions of the stored face over which the lattice samples were taken will not be visible in the novel face with viewpoint changes of this magnitude.

An alternative approach to the problem of view independent face recognition is to use multiple views of faces to generalize recognition performance between sampled views [1], [15]. This approach has been implemented successfully in template-based systems [14] by Beymer [1] and in autoassociative memory models by Valentin and Abdi [15]. In the template model, recognition of a novel view of a face occurs by locating a subset of facial features and by using their location/configuration to register the input face geometrically with the model views. The input face is then correlated with stored face images to produce a match. In the autoassociative memory, multiple views are stored and intermediary views can be reconstructed. For both cases, excellent recognition generalization over a wide range of poses was achieved. This approach is quite effective if several views of faces are available.

2 Rationale and Approach

It is well known that human observers are capable of recognizing familiar faces from any of a number of viewpoints. By *face recognition*, we mean the classification of a face as "familiar"/"known" versus "unfamilar"/"unknown". Clearly, this ability could be due to the fact that when a face is familiar to us, it is likely that we have encountered it previously in a variety of orientations. For unfamiliar faces, however, even those encountered only from a single viewpoint, while human performance across large pose changes is not perfect, it is still well above chance [7]. This latter fact indicates that there is information available in a single view of a face to make relatively accurate recognition judgments even across quite large changes in viewpoint (e.g., full face to profile). Intuitively, this is not surprising since much of the information that makes individual faces recognizable is visible from largely different poses/viewpoints. This kind of information includes both global features such as skin texture and tone, as well as very local features such as moles, blemishes and dimples.

In the present study, we describe a system that uses a view-dependent coding from only a single pose of a face, to recognize the face when it has been rotated by 45 and 90 degrees. To do this, we have made one important assumption. Specifically, we assume that a person or computational model can determine, with reasonable accuracy, the view from which a face is imaged, e.g., can determine whether any given view is taken from the front (i.e., full-face), side (i.e., profile), or some other intermediary position. With this information, it is possible to map the view-dependent information onto a three-dimensionally invariant code common to all heads. In short, when we know that we are looking at a profile, we can map the information it contains onto the "profile section"

of a memory representation of heads.

In any given pair of views of a head, up to nearly 180 degrees of rotation between the two, some common parts of the surface will be visible. In fact, for full face and profile views (90 degree rotation), quite a large area of surface is visible in both views. Assuming that a reasonably accurate mapping can be made to a standardized head code, the question then becomes, "How useful is the overlapping information between any given pair of views of a head for recognizing the head (i.e., distinguishing a "known" head from other "unknown" heads)?

The method we employ to answer this question is a simple extension of the eigenvector-based analysis that has been used frequently in computational models of image-based face recognition in recent years [3, 4, 10, 11, 12]. As has been pointed out [10, 13], this method is consistent with much older work by Kohonen [5] using autoassociative memories for face recognition. Of direct interest for the present work, Kohonen illustrated that an autoassociative memory can serve as a pattern completion device when noisy or partially ablated faces are used as memory keys.

The present study is a very straightforward extension of this approach as follows. We treat a view of a head (e.g., full-face, three-quarter, or profile) as partial information about the head. The remaining information about the head is not completely unspecified since we have general information about the shape of human heads. Thus, a single view of a face is coded in the present study as a complete head, part of which contains the information visible from the encountered view. The remaining parts are filled in with average head values taken from a set of faces. When a face is encountered from a single view, and re-encountered from a novel view, the re-encountered view (coded as above) acts as a partially ablated version of a learned pattern. This partially ablated pattern can be used as a retrieval cue for the autoassociative memory, which acts to complete the pattern based on its similarity to a learned pattern. Thus, the question becomes, "To what degree is overlapping information sufficient to retrieve the learned view, and hence, distinguish previously encountered faces from novel cases under various degrees of pose change?"

3 Simulations

We carried out two simulations contrasting the relative utility of texture versus surface-based data in discriminating learned from novel heads. The methods are carried out separately on each kind of face code. For convenience and brevity, we describe the methods for the range data, noting any differences in methods required to deal with the texture data.

3.1 Methods

Apparatus. Simulations were performed on a database of laser scanned three-dimensional head models that were collected using a Cyberware Laser Scanner^{TM}.

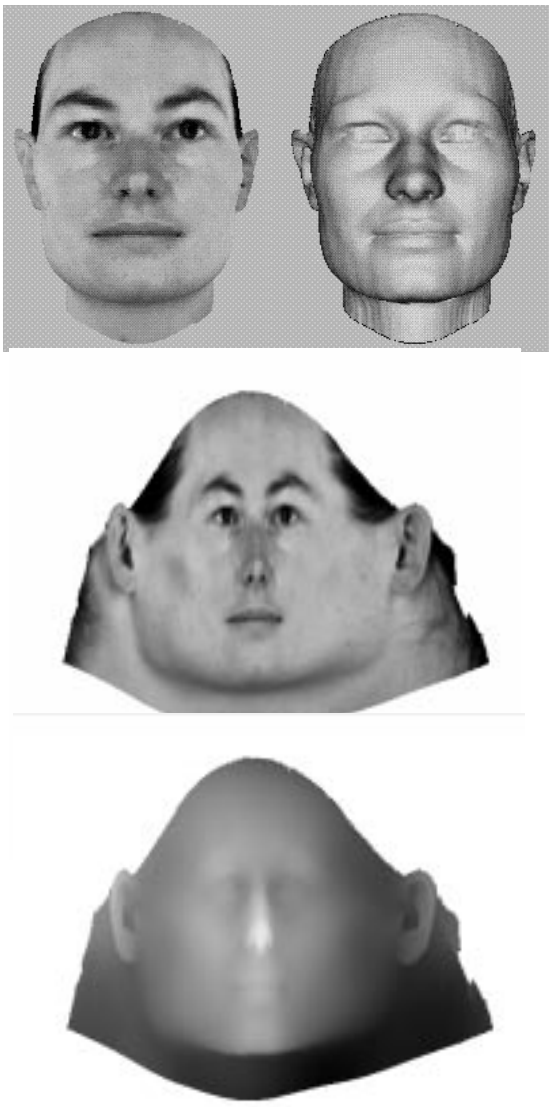

Figure 1: Top left: Subject represented using both the surface data and texture map. Top right: the same subject represented using only the surface data. Middle: the texture data only, "unrolled" to see the entire head. Bottom: the range data only, "unrolled" to see the entire head.

The representation of the scanned heads con-

sisted of two parts: (1) a three-dimensional surface map, which we will refer to as *range data*; and (2) a *texture map* containing color values at all points of the three-dimensional surface. Figure 1 illustrates the difference between texture and range data. The top left image of Figure 1 shows a full-face view of the texture map pasted onto the range data. The top right image is a full-face view of the range data displayed here by modeling the illumination of the three-dimensional surface (range data) of the head with a standard Lambertian shading model. This image resembles a bust of a person made out of uniform material (constant albedo). The middle image of Figure 1 shows the texture map for the entire head. This can be thought of as a kind of "peeled off" skin. A similar "unrolled" representation of the range data appears in the bottom image. These representations comprised the data for our simulations. Gray levels are used to display "depth" values for the surface. We discuss the nature of the depth value coding for the head model in detail shortly.

If the texture data are wrapped around the surface (range data of the same person), any view of the person can be computed with standard computer graphics techniques. Figure 2 shows the three-quarter and profile view of the person shown in full face view in Figure 1.

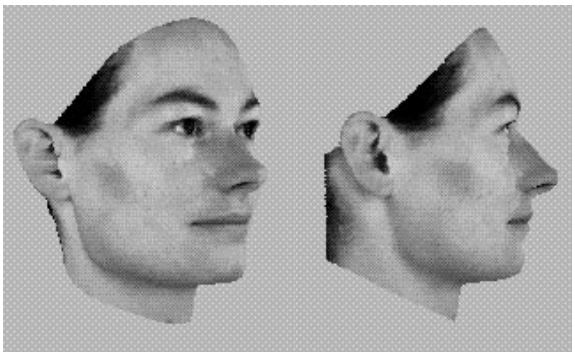

Figure 2: Left: Subject (shown in Fig. 1) rotated to three-quarter view. Right: the same subject shown in profile view.

Stimuli. Sixty-eight volunteers from the Max Planck Institute for Biological Cybernetics in Tübingen, Germany and the surrounding Tübingen area were scanned. To keep the face area free from hair, volunteers were asked to wear a bathing cap, which was adjusted to hide the hair as much as possible. Care was taken to align heights of volunteers' heads with a central point that rested on the top of the head surface during the scan.

The range data consisted of the lengths of radii from a vertical axis centered in the middle of the subject's head to the head surface. Specifically, the vertical axis formed the center of an imaginary cylinder. Each head comprised a 512-by-512 grid of radius lengths sampled at equally-spaced angles about the vertical axis and at equally-space heights along the long axis of the head. This grid is represented schematically in the "unrolled" range and texture maps in Figure 1.

The texture data consisted of a standard *rgb* image that maps point for point onto the range data grid. For purposes of the present simulations, *rgb* values were reduced to gray levels using a standard weighted linear combination of the red, green, and blue values ($gray = .30 \times r + .59 \times g + .11 \times b$).

Since the quality of the laser scan data in the region of the hair was unsatisfactory, and since we were not interested in the back of the head, further processing of the heads was carried out as follows. First, the region covered by the bathing cap was removed completely. Second, a vertical cut was made behind the ears. Third, with a horizontal cut, we removed the shoulders. Finally, prior to the simulations, we aligned all of the head data vertically to a constant eye height, by locating (manually) the y coordinate of the eye level of each head individually and digitally translating the head surfaces accordingly. Simple arithmetic averages of the range and texture maps over all 68 heads were used as data for filling in parts of the head not visible in a given view.

Views of heads for each pose group were created by ablating parts of the surface map that would not be visible from that view and replacing these parts of the range or texture map with values taken from the average range or texture map, respectively. While there are different ways to define the face views, we approximated them simply as follows. First, the radius values of the cylindrical coordinate system were converted into three-dimensional Cartesian coordinates. Next, the head was rotated about the vertical axis 0, 45, and 90 degrees, for the full-face, three-quarter, or profile views, respectively. Finally, the outer edges/contours of the rotated face were located by finding the minimum and maximum x coordinates in each row of the scan. "Hidden" sample points were eliminated by replacing radii in each row with indices greater or lesser than these outer edge coordinates with values taken from the average head. This algorithm for finding the views is not perfect, since it will miss some internally blocked regions of the face, which have the extreme x coordinates

in the row more peripherally located. (e.g., the inner ear regions in the full face, for which the extreme x coordinates for the row are located on the pinnae of the ear). However, in general, quite good approximations of the views can be made. The top of Figure 3 shows a three-quarter view of the range data taken from a head. The bottom two images of this figure show: (a.) left: the same head turned to a full-face view so that the missing parts can be seen easily; and (b) right: the same head and view with the missing parts filled in by values taken from the average range map.

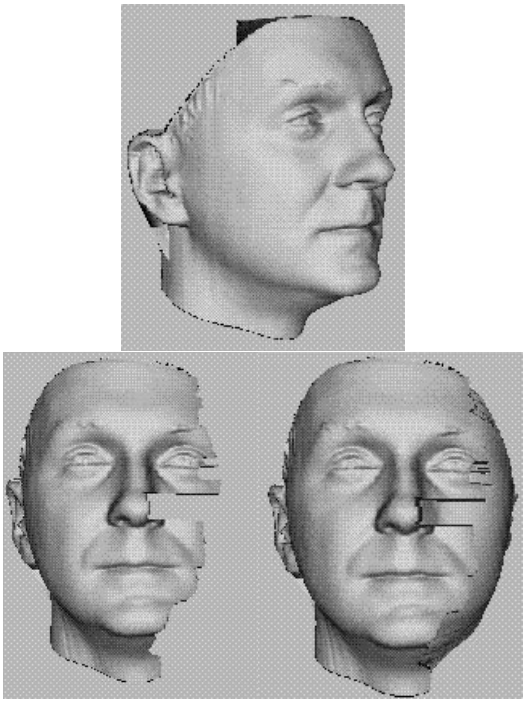

Figure 3: Top: Range data from a face viewed from three-quarter pose. Left: Head rotated to see missing parts. The weakness of the view algorithm can be seen (i.e., some parts behind the bridge of the nose should be invisible but are retained due to the existence of more peripheral x values on the cheeks). Right: Rotated head with missing parts filled in with values from the mean head.

Learning Procedure. A cross-product matrix was created from the range data for 51 of the original 68 heads (26 male and 25 female). Henceforth, we refer to the heads used to create this matrix as the *learning set*. Heads in the learning set comprised three pose groups: full-face, three-quarter, and profile, with almost equal numbers of male and female heads (i.e., 17) per group. The remaining 17 heads from the original 68 were reserved for testing purposes. Each face view was coded as a vector consisting of the concatenation of the rows of the range map. The mean of the learning set was subtracted from all face vectors, and the vectors were normalized in length, such that $f_i^T f_i = 1$. The cross-product matrix was computed as:

$$\mathbf{A} = \sum_{i=1}^{n} f_i f_i^T \tag{1}$$

where f_i is the i^{th} face and where n is the number of faces[1].

Recognition Testing. Using an eigenvector-based representation computed from the learning set of head stimuli, a face view f_i from this set can be expressed, without error, as a weighted sum of the eigenvectors of the matrix $\mathbf{A}$ as follows:

$$f_i = \sum_{j=1}^{r} (f_i \cdot u_j) u_j \tag{2}$$

where $(f_i \cdot u_j)$ is the dot product between the i^{th} face and the j^{th} eigenvector and where r is the rank of the matrix. An estimate of a novel view of a learned face or an unlearned face can be made by applying the same operation (i.e., Eq. 2) to the novel view or unlearned face. In this case, however, the left side of Eq. 2 does not produce a perfect reconstruction of the face, but rather, an estimate, which we will refer to as $\hat{f}_i$. The quality of this estimate can be evaluated by taking a measure of the similarity between the original and reconstructed vectors, which we measured as the cosine between f_i and $\hat{f}_i$ as follows:

$$\frac{(f_i \cdot \hat{f}_i)}{\|f_i\| \|\hat{f}_i\|} \tag{3}$$

Perfect reconstructions of f_i yield cosines of 1.

In summary, recognition testing is applied to three kinds of inputs: (1) learned views of learned faces, which are retrieved without error; (2) novel views of learned faces; and (3) unlearned faces.

The design of the present study involved the manipulation of two independent variables: learning pose (full-face, three-quarter, and profile) and testing pose (full-face, three-quarter, and profile). The dependent variable was a measure of the model's ability to discriminate learned versus novel faces in each of the nine combinations of the two independent variables. Face recognition, the dependent variable in this design, involves a decision about whether a face is "known" or "unknown". With eigenvector-based representations of faces this task has been simulated in

[1] In fact, the normalization and centering procedures makes this matrix a correlation matrix

several ways (cf., [11] for an alternative method to the one employed here). In the present study, we have used signal detection theory [6] to model face recognition (cf., [10]). In general, the idea is that the "signal" is comprised of known faces, which must be discriminated reliably from "noise" or unknown faces. In the present case, "known" faces include all views of any face learned, regardless of the viewpoint from which the face was learned. "Unknown" faces were those not learned from any viewpoint. The measure on which this known/unknown discrimination was made is the quality of the face estimation or reconstruction, measured as the cosine between the original and reconstructed face. This can be considered as a sort of "resonant familiarity". The model is said to be able to recognize faces when, on the average, the cosines for the learned faces exceed the cosines for the novel faces. The measure taken is referred to as *d'* and is simply the distance, in *z*-score units between the means of the cosine distributions for learned and unlearned faces.

Recognition testing was implemented in the present study as follows. All views of all 68 faces (both novel and learned) were estimated using Eq. 2. For each learn-test condition, a mean cosine was computed across the learned faces. For each test condition for the unlearned faces a mean cosine was computed. A *d'* was computed for each learn-test condition by setting a criterion cosine as the mean of the means for the learned faces and novel faces for the appropriate test condition. Thus, for example, for the full-learn/profile-test condition the profile test condition for novel faces was used as the noise distribution.

One final methodological point is worth noting before presenting results. Since we had a relatively small number of heads, we ran four simulations *counterbalancing* the heads over conditions. The counterbalance was implemented such that over the set of 4 simulations, every head appeared in each of the conditions exactly once (i.e., learned full, learned three-quarter, learned profile, novel). This was done to minimize the possibility of a sampling fluke occurring with particularly distinctive heads clustered in any given condition.

3.2 Results

All results are from the four counterbalance simulations for the texture and range data, which we present in two ways. First, we plot the quality of reconstructions, measured as cosines between original and reconstructed vectors for all learning-test transfer conditions and for the novel heads in each pose condition (see Figure 4). From these data, several points are worth noting. First, for both range and texture data, the quality of reconstructions was, in general, better for the learned faces than for the unlearned faces. This was true, regardless of the match/mismatch between the learn and test conditions. Thus, novel views of the learned faces can effectively retrieve enough of the learned view to produce reconstructions better than those seen for the unlearned faces. Second, smaller changes in viewpoint between the learn and test views resulted in better reconstructions than larger changes. This indicates, not too surprisingly, that better retrieval of the learned view was possible when larger regions of the learned and tested view overlapped.

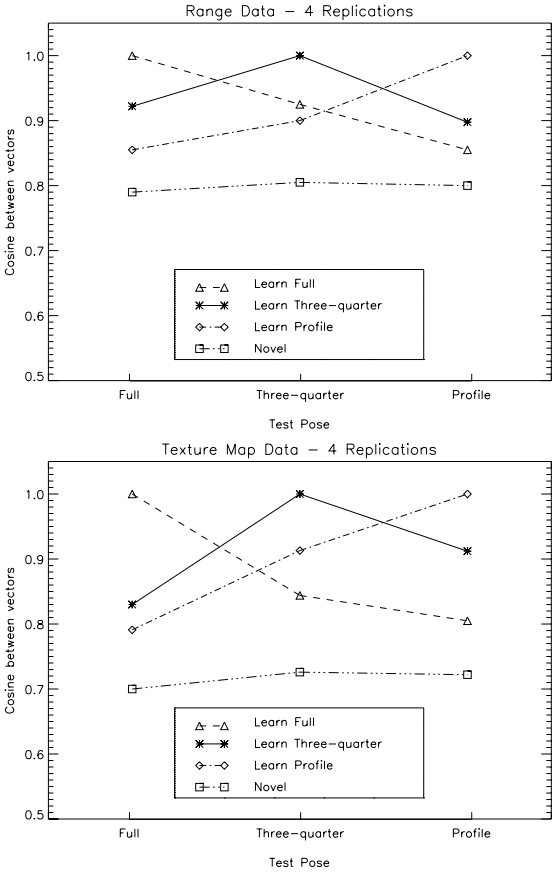

Figure 4: Quality of reconstruction for range and texture maps, measured as cosine, for learned faces as a function of learn and test view, and for unlearned faces as a function of test view.

The cosine data alone, however, do not give an indication about how reliably the learned faces (regardless of test view) can be *discriminated* from the unlearned faces. For this, we applied signal detection theory, which extends the above anal-

Range Data

	Training view								
Testing View	Full			Three-quarter			Profile		
	Hit	FA	*d'*	Hit	FA	*d'*	Hit	FA	*d'*
Full	1.00	0.00	3.78	0.97	0.05	3.54	0.94	0.36	1.92
Three-quarter	0.97	0.08	3.30	1.00	0.00	3.78	0.96	0.23	2.59
Profile	0.79	0.36	1.17	0.99	0.20	2.74	1.00	0.00	3.78

Texture Data

	Training view								
Testing View	Full			Three-quarter			Profile		
	Hit	FA	*d'*	Hit	FA	*d'*	Hit	FA	*d'*
Full	1.00	0.00	3.78	0.91	0.32	1.82	0.75	0.32	1.15
Three-quarter	0.88	0.23	1.92	1.00	0.00	3.78	0.99	0.17	2.86
Profile	0.79	0.32	1.28	1.00	0.16	2.89	1.00	0.00	3.78

Table 1: Recognition performance measured as *d'* for the range and texture map data.

ysis by setting a criterion cosine for determining the learned/novel status of faces in each condition. This technique is commonly applied in the psychological literature to measure human recognition memory for faces.

These data appear in Table 1, for the surface maps (top), and for the texture maps (bottom). In each table cell, three values are given: (1) the *hit rate*: the proportion of times a learned face was correctly labeled learned; (2) the *false alarm rate*: the proportion of times an unlearned face was incorrectly called learned; and (3) the *d'*: the discrimination index. As can be seen, performance ranges from moderately good to excellent in the different conditions.

Perfect discrimination of learned and novel faces was observed when the learned and test faces were of the same view[2]. Again, when the learned and test views were not the same, the general pattern of results indicated better performance with smaller pose changes

4 Conclusions

The present study illustrates that information for face recognition across pose change is available in the overlapping visible surface or texture maps between pairs of views. This information is reliable for pose changes of as much as 90 degrees and can be retrieved using a simple linear autoassociative model, when there is sufficient overlap between the learned and test faces. As noted, we rely here on the assumption that it is possible to make a reasonably accurate assessment of the pose of a face. While to our knowledge, there is no psychophysical data to support this conclusion, we think that it is certainly reasonable to believe that humans can make reasonably accurate estimates of pose. There is also computational work indicating that the pose of a face is detectable by relatively simple models [1, 15].

The present study provides only an exploratory look at the utility of this approach for measuring the quality of information available in surface/texture maps for making pose transfers. We do not wish to claim that the present data represent the "last word" on the subject. In fact, the particular way in which we have implemented this model has implications for the precise outcome of the data. For example, we have "filled-in" nonvisible parts of the heads with values from the average head. With the additional assumption that heads are generally symmetric, it is likely that symmetric fills of the head, computed from each individual head, would benefit pose transfer performance. Additionally, such a representation might show better pose transfer abilities between symmetric pose changes than between other smaller non-symmetric changes in pose.

We view the importance of representational assumptions to the outcome of the transfer data here as a positive aspect of the model. This characteristic makes the model a very useful tool for testing quite specific psychophysical hypotheses about human representations of faces and quite specific computational hypotheses concerning the optimal-

[2]In cases where hit or false alarm rates indicate perfect performance, *d'* is effectively infinite. Therefore, we applied the standard correction for perfect hit or false alarm rates, cf., [9], leading in the present case to *d'*s of 3.78 indicating no errors.

ity of these representations. As such, psychophysical data can be collected on the recognizability of these heads over various pose changes, and comparisons can be made between the human data and the model performance as a function of these representational assumptions.

Additionally, the optimality of different representations for different tasks (e.g., recognition versus sex classification) can be examined. This is because the eigenvector-based representation allows for a detailed analysis of the utility of individual eigenvectors for different tasks. With face images, this analysis has shown that *different* low dimensional representations of faces are optimal for recognition versus categorizations (e.g., sex and race classifications [10]). This analysis is likely to prove fruitful for the present stimuli since we are able to separate texture versus surface based information.

References

[1] D.J. Beymer, "Face recognition under varying pose", *A.I. Memo, MIT* No. 1461, 1993.

[2] V. Bruce, *Recognizing Faces*, Lawrence Erlbaum, 1988.

[3] G.W. Cottrell, and M.K. Fleming, "Face recognition using unsupervised feature extraction.", *Proceedings of the International Neural Networks Conference*, Kluwer Dordrecht, pp. 322-335, 1991.

[4] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski, "SexNet A neural network that identifies sex from human faces", In R. Lippmann, J. Moody, and D. S. Touretsky (Eds.) *Advances in Neural Information Processing Systems 3*, Morgan Kaufmann, 1991.

[5] T. Kohonen, *Associative Memory: A System Theoretic Approach*, Springer-Verlag, 1977.

[6] D.M. Green, and J.A. Swets *Signal Detection Theory and Psychophysics*, New York: Wiley, Reprinted by Krieger, Huntington: NY, 1966.

[7] F.L. Krouse, "Effects of pose, pose change, and delay on face recognition", *Journal of Applied Psychology*, Vol.66, pp. 651-654, 1981.

[8] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture", *IEEE Transactions on Computers*, Vol. 42, pp. 300-311, 1993.

[9] N.A. Macmillan, and C.D. Creelman *Detection Theory: A User's Guide*, Cambridge University Press, 1991.

[10] A.J. O'Toole, H. Abdi, K.A. Deffenbacher and D. Valentin, "A low-dimensional representation of faces in higher dimensions of the space", *Journal of the Optical Society of America A*, Vol.10, pp. 405-411, 1993.

[11] M. Turk, and A. Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, Vol.3, pp. 71-86, 1991.

[12] L. Sirovich, and M. Kirby, "Low-dimensional procedure for the characterization of human faces", *Journal of the Optical Society of America A*, Vol.4, pp. 519-554, 1987.

[13] D. Valentin, H. Abdi, A.J. O'Toole, and G.W. Cottrell, "Connectionist models of face processing: A survey", *Pattern Recognition*, Vol. 27, pp. 1209-1230, 1994.

[14] R. Brunelli and T. Poggio, "Face recognition: Features versus templates", *IEEE Trans. PAMI*, Vol. 15(10), pp. 1042-1051, 1993.

[15] D. Valentin, H. Abdi, "How come when you turn your head I still recognize you", *Paper presented at the Annual Meeting of the Society of Mathematical Psychology*, Seattle, Washington, 1994.