published in: Journal of the Optical Society of America A 14:9 (1997) 2152-2161.

Separation of texture and shape in images of faces for image coding and synthesis

THOMAS VETTER AND NIKOLAUS F. TROJE

vetter[troje]@mpik-tueb.mpg.de

Max-Planck-Institut für Biologische Kybernetik,Spemannstr. 38, 72076 Tübingen, Germany

Abstract. Human faces differ in shape and texture. Image representations based on such a separation have been reported by several authors [for a review, see Beymer and Poggio, (1996)]. This paper investigates such a representation of human faces based on a separation of texture and two-dimensional shape information. Texture and shape were separated using pixel-by-pixel correspondence between the different images, which was established through algorithms known from optical flow computation. The paper demonstrates the improvement of the proposed representation over well established pixel-based techniques in terms of coding efficiency and in terms of the ability to generalize to new images of faces. The evaluation is performed by calculating different distance measures between the original image and its reconstruction and by measuring the time human subjects need to discriminate them.

Keywords: Image synthesis, face recognition, flexible templates

1. Introduction

Human language is organized in terms of hierarchical categories that comprise similar objects of the world into object classes. A natural description of an object typically assumes a priori knowledge about the object class per se and evaluates only its particularities and deviations with respect to a prototype. This comparison contains differences in surface properties as well as in the spatial arrangement of object features.

For human faces, most attributes used in such a description are continuous. Faces can have lighter or darker skin, larger or smaller eyes, a wider or a narrower mouth. Other image attributes that do not belong directly to the face itself, such as the illumination or the viewpoint from which the face is seen, are also physically continuous and are perceived and described as such. Some of the above attributes, e.g. the skin colour or the illumination, correspond to continuous intensity changes within parts of the image. Others, such as the size of the eyes or the shape of the mouth, correspond to changes in the spatial arrangement between the parts in the image. They have to be described as image distortions rather than intensity changes. In most cases, it is easy to classify an attribute as having changes in the surface properties only or changes in the spatial arrangement of the features only. We will refer to the information contained in surface properties as the texture of the face and the information contained in the spatial arrangement as the shape of the face (since we are working with images, we mean two-dimensional shape). Image representations separating shape and texture information have been introduced by several authors (for a review, see Beymer and Poggio, (1996)) They differ in principal from approaches that do not rely on information about the feature-by-feature correspondence and subse-

quently do not separate shape and texture information (for a review, see Valentin, (1994))

In this paper, we argue that a representation that separates the information contained in the image of a face into its *texture* and *shape* components has several advantages for modelling and for efficient coding. We will present an algorithm that separately represents texture and shape. We will then evaluate the quality of low dimensional reconstructions using this representation by comparing it to the quality of low-dimensional reconstructions in a pixel-based image space.

We will use the term *pixel-based representation* for representations that are based on a code in which a digitized image of size $s = nxm$ is described by a vector of length s by simply concatenating all the intensity values in the image. Low-dimensional representations can then be achieved by performing a Karhunen-Loeve expansion and using subspaces spanned by only the first principal components (Ahmed and Goldstein, 1975). Such representations were first introduced by Sirovich and Kirby (1987) and have been applied successfully to many different tasks, such as face recognition(Turk and Pentland,1991; O'Toole et al., 1993; Abdi et al., 1995) and gender classification (O'Toole et al., 1991). O'Toole, Deffenbacher, Valentin and Abdi (1994) have used such a representation to model the "other race effect", a well known psychophysical phenomena (Feingold, 1914).

As pointed out by several authors(Craw and Cameron, 1991; Vetter and Troje, 1995,Beymer and Poggio, 1996) pixel-based face representations have very unpleasant properties. An important property of a linear space is the existence of an addition and a scalar multiplication defining linear combinations of existing objects. All such linear combinations are objects of the space. In a pixel-based representation, this is typically not the case. One of the simplest linear combinations - the mean of two faces - will in general not result in a single intermediate face, but will appear as two superimposed images. Any linear combination of a larger set of faces will appear blurry. The set of faces is not closed under addition. These disadvantages can be reduced by carefully standardizing the faces in the images, for instance by providing for a common position of the eyes (e.g. Kirby and Sirovich, 1990) However, even after such a standardization step the matching error can still be very large, yielding fuzzy and imprecise images.

Aligning the eyes of two faces requires only translation and scaling operations. For an alignment of all features in a face, such linear image operations are not enough. Nonlinear deformations have to be used to change the spatial arrangement of the different features. The spatial arrangement, however, might be an important part of the character of a face and changing it would mean changing appearance and identity. Aligning faces feature-by-feature leads to a shape-free (Craw and Cameron, 1991) representation, which is deprived of an important part of the information contained in the face. On the other hand, a space spanned by shape-free faces (at least if enough features were aligned) is a proper linear space with the set of faces being convex in the sense that any point between two faces will result in a sharp face again. The mean of two "shape-free" faces will no longer be qualitatively distinguishable from the original shape-free faces.

The shape that has been eliminated in the "shape-free" representation can well be described in terms of the nonlinear deformation that maps a given face onto a common face. This common face serves as a prototype and defines the origin of the resulting "shape space". The "shape-free" representation together with the "shape" itself contains all the information that had originally been in the image. A representation of the shape in terms of the deformation field with respect to a common prototype is also convex. The mean of two deformation fields is a valid deformation field.

The crucial step in defining the deformation fields between different images is to establish feature-by-feature correspondence between them. For this reason, we call this kind of representation a correspondence-based representation in contrast to pixel-based representations, which are based on the unprocessed images. As mentioned above, we will use the terms texture and shape for the two parts of the correspondence-based representation.

In the past few years, different researchers have worked with correspondence-based representations of human faces,(Craw and Cameron, 1991; Vetter and Troje, 1995; Perrett et al. 1994; Vetter, 1996; Costen et al.,1996; Hancock et al., 1996). The features used for establishing correspondence

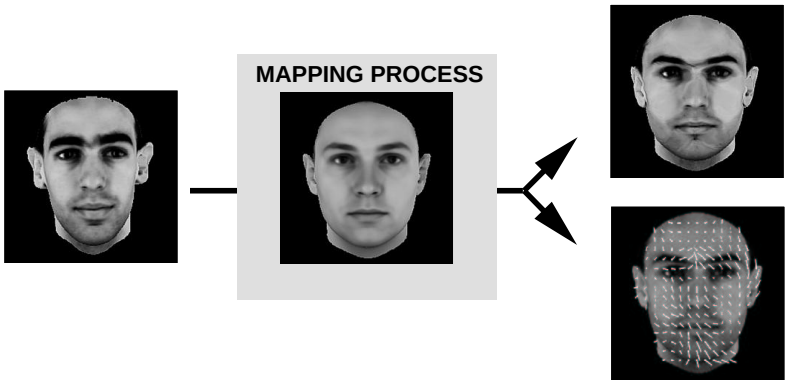

Fig. 1. An example of a face image (left) mapped onto the reference face (center) using pixel-by-pixel correspondence established through an optical flow algorithm is shown. This separates the 2D-shape information captured in the correspondence field (lower right) from the texture information captured in the texture mapped onto the reference face (upper right).

span the whole range between semantic meaningful features, such as the corners of the eyes and mouth, to pixel level features that are defined by the local grey level structure of the image. Most authors have established correspondence by hand-selecting a limited set of features in the faces. Beymer, Shashua and Poggio (1993) solved the correspondence problem by using an adapted optical flow algorithm that established correspondence on the single pixel level.

The purposes for using a correspondence-based face space vary. Beymer and Poggio, (1996) used this representation to train a regularization network to learn image deformations produced by expression and pose changes. Craw and Cameron (1991) concentrated on artificial face recognition and compared recognition systems, based on principal component analysis (PCA), that used either pixel-based representations or correspondence-based representations. Their results clearly showed the advantage of a correspondence based representation in a recognition task. Hancock showed that principal components derived from a correspondence-based representation better reflect psychophysical ratings of distinctiveness and memorability than do principal components derived from a pixel-based representation.

In this paper, we investigate the advantages of a correspondence-based representation for modelling, coding efficiency and its generalizability. How much do we gain in terms of reconstruction quality when using a correspondence-based representation instead of a pixel-based representation? We will evaluate the quality of low dimensional reconstructions by means of theoretical considerations and by means of a psychophysical experiment.

Coding efficiency and generalizability was evaluated in a cross validation experiment. Faces from a test set were coded with the faces from a training set. Reconstructions were thus obtained by projecting test faces into spaces spanned by different numbers of principal components obtained from a set of training faces. As Kirby and Sirovich (1990) pointed out, this allows us to test the coding abilities of the representation better than we could by evaluating only reconstructions of faces that were already used to calculate the principal components. In the latter case, using all the principal components will always result in a perfect reconstruction, irrespective of the number of faces used for the calculation. Only if new faces are used will the relation between the reconstruction quality and the number of faces used to span the space be able to be used to draw conclusions about the dimensionality of the set of faces. The quality of the reconstructions will be evaluated by calculating different distance measures between the original image and its reconstruction and by means of a psychophysical experiment.

The paper is organized as follows. First, a principal component analysis (PCA) is performed separately on the shape and texture information as well as on the images themselves. All technical details of the implementation used to separate shape and texture in images of human faces are described in the Appendix. Following the theoretical evaluation of the representations, we describe a psychophysical experiment that evaluates the quality of the reconstructions. Finally, the main properties and possible future extensions of the representation are discussed.

2. Principalcomponent analysis, reconstructions and reconstruction errors

2.1. Separation of texture and shape in images of faces

The central part of the approach is a representation of face images that consists of a separate texture vector and 2D-shape vector, each with components referring to equivalent feature points. In our approach, we treat any single pixel as a "feature" and establish pixel-by-pixel correspondence between a given image of a face and a reference image. All images of the training set are mapped onto a common reference face. The correspondences were computed automatically using a gradient based optical flow technique which has already been used successfully previously on face images (Beymer et al., 1993; Vetter and Poggio, 1996). Technical details can be found in Appendices B and C. Assuming a pixel-to-pixel correspondence to a reference face, a given example image can be represented as follows: Its 2D-shape is coded as the deformation field that has to be applied to the reference image in order to match the example image. This deformation field is defined at each single pixel. So the shape of a face image is represented by a vector $S = (\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2, ...\Delta x_n, \Delta y_n)^T$, that is by the $\Delta x, \Delta y$ displacement of each pixel with respect to the corresponding pixel in the reference face. The texture is coded as a difference map between the image intensities of the exemplar face and its corresponding intensities in the reference face. This normalized texture can be written as a vector $T = (\Delta I_1, \Delta I_2, ...\Delta I_n)^T$, which contains the image intensity differences ΔI of the n pixels of the image (Fig. 1).

2.2. Linear analysis of texture, shape and pixel-represented images

We performed a PCA separately on both the texture and the shape part of the correspondence-based representation. In addition, we calculated the principal components from the images themselves (i.e. on the pixel-based representation). The database of images contained 100 faces. For details, see Appendix A.

Principal components were obtained by calculating the eigenvectors of the covariance matrix of the data. Figure 2 and 3 show variations of the shapes and the textures along the first four principal components of the two subspaces. To the average face we added the respective normalized principal component with weights corresponding to two, four and six standard deviations in both directions. The center row in both Figures 2 and 3 always shows the same image: the average face consisting of the average texture and the average shape. Although six standard deviations widely exceed the range of naturally occurring images, even the most extreme faces still look rather normal. The first few principal components clearly mark important characteristics. The first component of the shape shows a size effect that seems to correspond with the perceived gender. The second component captures the size of the forehead. Component three accounts for rotations around the horizontal axis in the image plane. The fourth component shows the transition from a narrow head to a wide head and thus accounts for the "aspect ratio" of the face.

The first principal component of the textures clearly captures the variability in illumination that is still present in our data base. The light source is moving from below to above and changes in its intensity. The second texture component accounts for facial hair. Although the data base contained no people with beards, some males were better shaved than others. Extrapolating six standard deviations away from the mean recovers the beard. This correlates with the strength of the eyebrows. People with pronounced eyebrows also

have a more pronounced beard-growth. However, variations of the strength of the eyebrows also occur in women and carefully shaved men. They show up in the third and fourth principal component. The higher components (not shown) are not so clearly characterizable. A lot of them code for very small changes in the intensity distribution in the eyes.

Figure 4 illustrates the first four principal components as calculated from the images themselves. To make them comparable with the principal components derived from the correspondence-based representation, we also show the images corresponding to locations that were either two, four, or six standard deviations away from the mean. The images are more blurry, in particular in the center of the space. The faces in the periphery carry strange features, such as the white ghost mouth in component 2 or the white outline in component 3.

The first principal component in the pixel-based representation mainly captures illumination differences. The second component captures the position of the mouth and the ears. The third component captures the size of the face which, as in the first component of the shape subspace, goes along with a shift in perceived gender. Note that this size change occurs here in the form of adding or subtracting a ring around the face. If too much is added, this ring becomes unnaturally white. The fourth component seems to account for left/right illumination changes. However, this impression is misleading. The database did not contain any horizontal variability in the direction of the light source. The fourth principal component rather accounts for a slight rotation around a vertical axis. In the pixel-based representation, this means that there has to be a little bit added onto the left side of the face if it is turned to the right side and vice versa. If too much is added (six standard deviations away from the mean), this leads to a white edge that is than interpreted as being due to illumination.

$$testing\ error_k = \frac{1}{n\sigma^2}\sum_n (X_k - X)^2. \qquad (1)$$

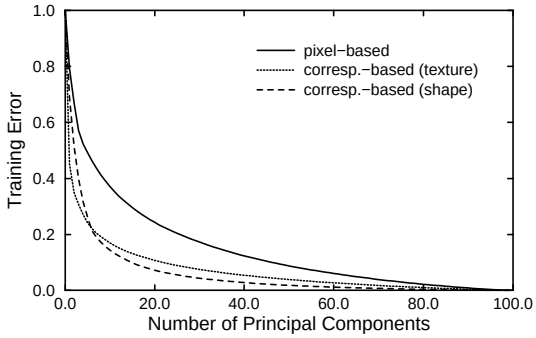

Fig. 2. Training error. In this diagram one minus the relative cumulative variance is plotted. The cumulative variance is equal to the mean of the squared Euclidian distance between the original face and reconstructions derived by truncating the principal component expansion. The calculation was performed for the two parts of the correspondence-based representation and for the pixel-based representation.

2.3. Reconstructions and reconstruction errors

PCA yields an orthogonal basis with the axes ordered by means of their overall variance. In Figure 5, we plotted for the three different PCAs described in the previous section one minus the relative cumulative variance covered by the first k principal components. The relative cumulative variances were calculated by successively summing up the first k eigenvalues ν_i and dividing them by the sum of all eigenvalues:

$$training\ error_k = 1 - \frac{\sum_k \nu_i}{\sum_n \nu_i}. \qquad (2)$$

Note that this term is equivalent to the expected value for the mean squared distance between the reconstruction X_k and the original image X divided by the overall variance σ^2 .

$$training\ error_k = 1 - \frac{\sum_k \nu_i}{\sum_n \nu_i} = \frac{1}{\sigma^2(n-1)}\sum_n (X_k - X)^2. \qquad (3)$$

It is thus an appropriate measure for the reconstruction error. Since it refers to the set of faces that was used to construct the principal component space from which the reconstructions were made we call this kind of error the *training error*.

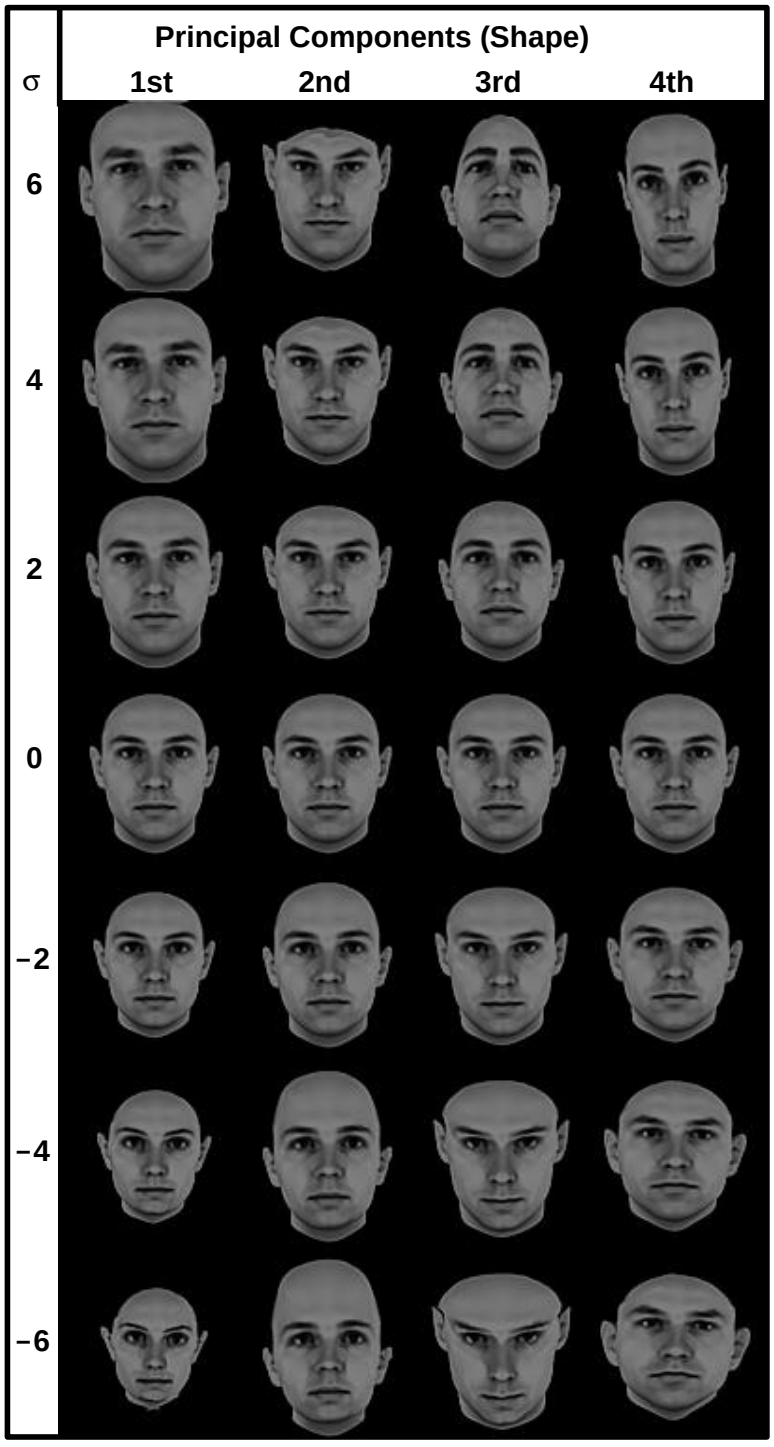

Fig. 3. Images along the four first principal components of the shape. The coefficients for the respective axis have values corresponding to two, four, and six standard deviations away from the mean face in both directions. All other coefficients (including the ones coding for the texture) are set at zero.

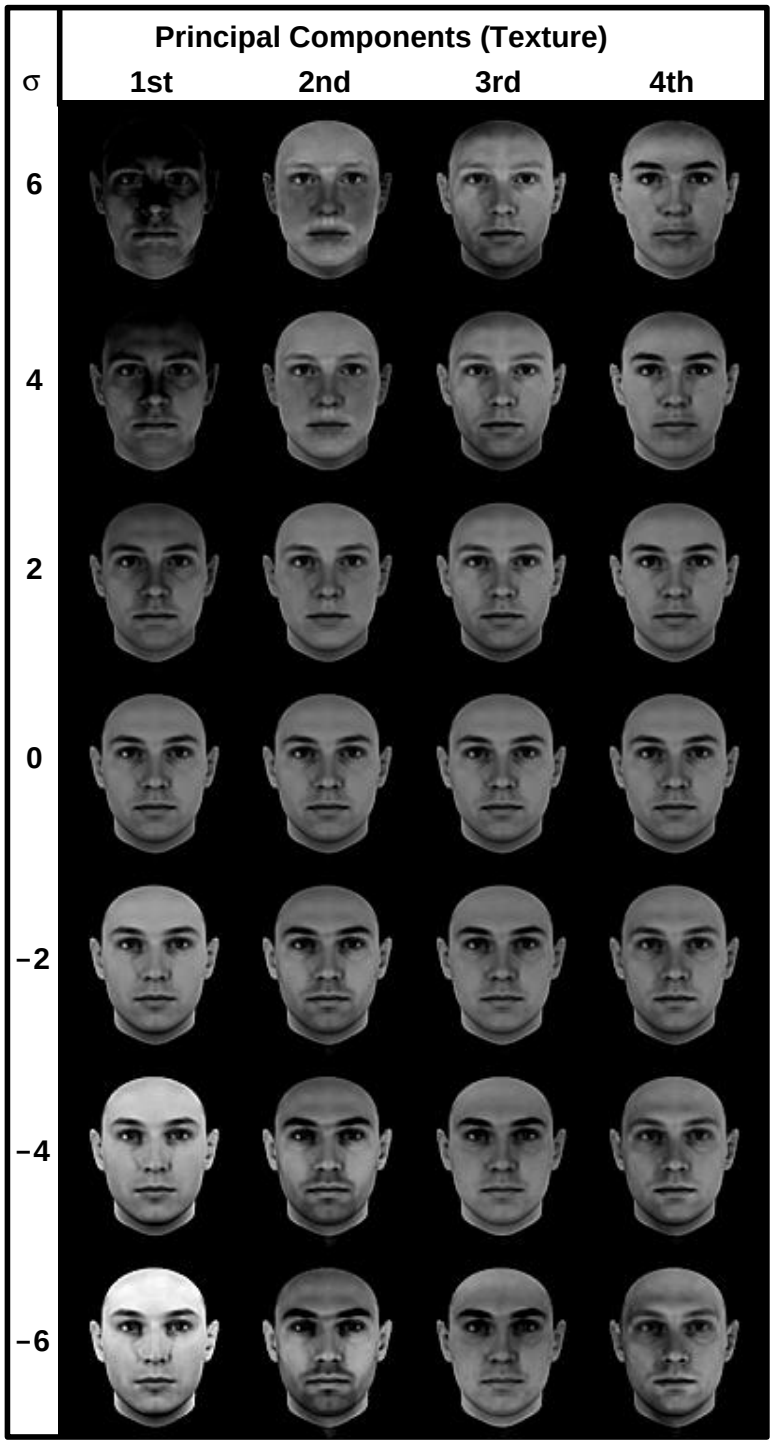

Fig. 4. Images along the four first principal components of the texture. The coefficients for the respective axis have values corresponding to two, four, and six standard deviations away from the mean face in both directions. All other coefficients (including the ones coding for the shape) are set at zero.

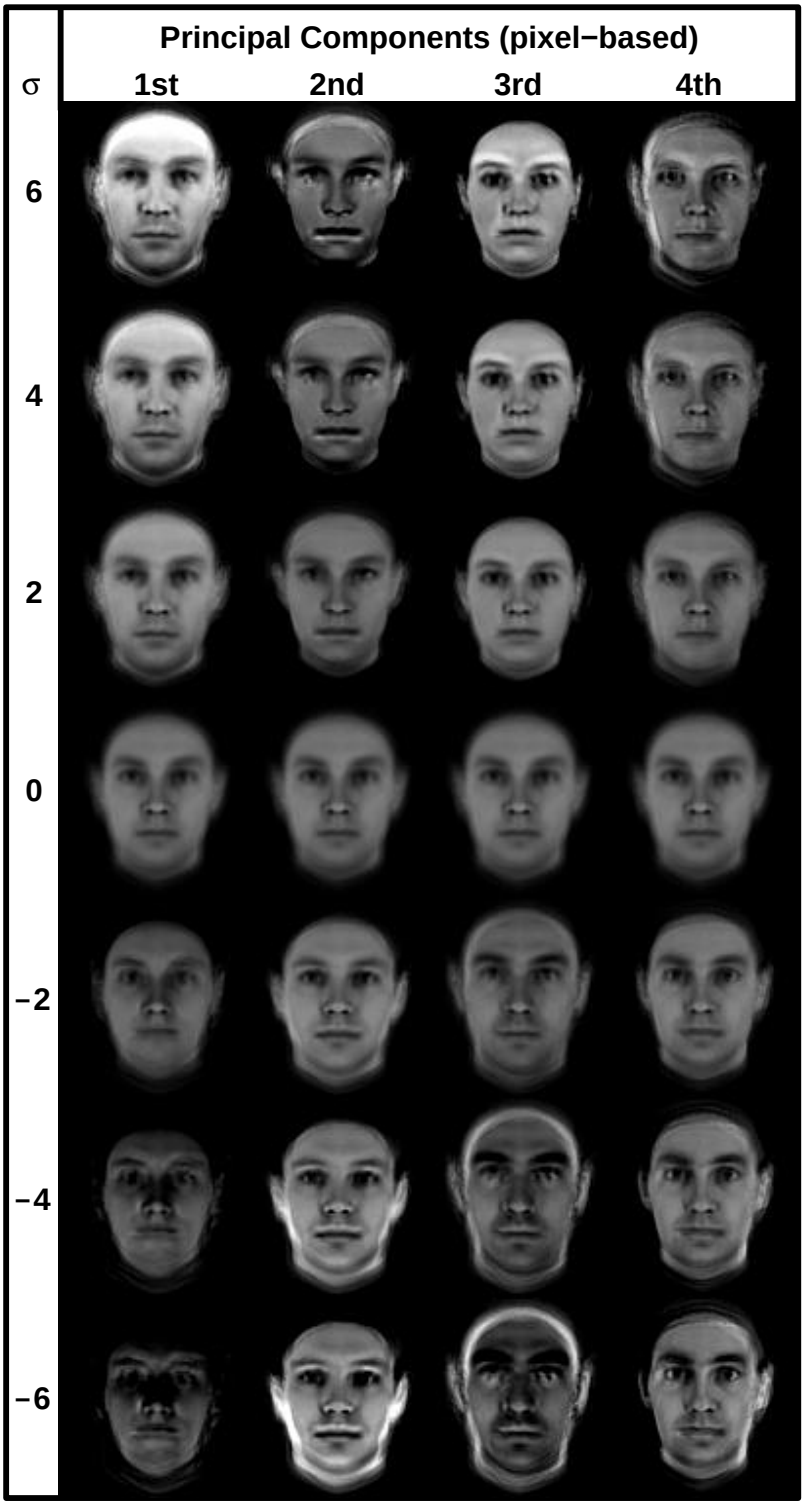

Fig. 5. Images along the first four principal components derived from the pixel-based representation. The coefficients for the respective axis have values corresponding to two, four, and six standard deviations away from the mean face in both directions. All other coefficients are set at zero.

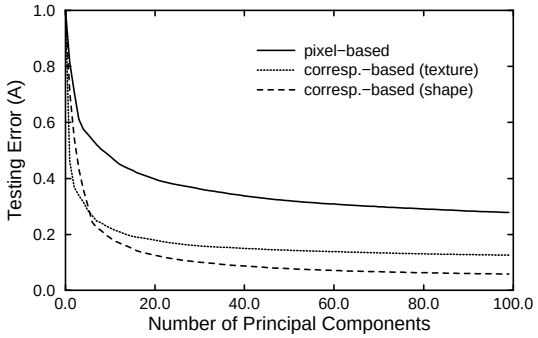

Fig. 6. Testing error (A). The relative mean squared Euclidian distance between the original and its reconstructions. In this case, the reconstruction was derived by projecting the data into spaces spanned by principal components computed from the set of remaining faces which did not contain the original face. The calculation was performed for the two parts of the correspondence-based representation and for the pixel-based representation.

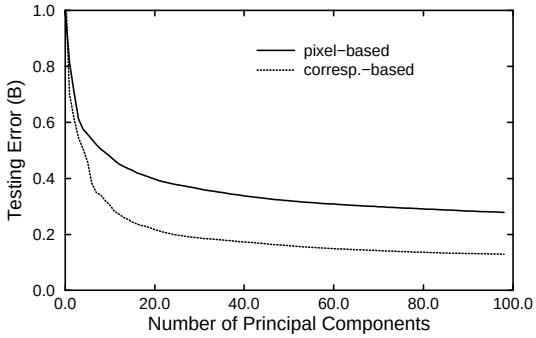

Fig. 7. Training error (B). As for the calculation of testing error A the faces were projected into principal component spaces derived from the remaining faces. The error for the pixel-based representation is the same as the one plotted in Figure 5. The error corresponding to the correspondence-based representation is measured by the squared Euclidian distance in the pixel space after combining the reconstructed shape with the reconstructed texture to yield an image (for details, see text).

For a training error of 10% (i.e. to recover 90% of the overall variance), the first 47 principal components are needed in the pixel-based representation, 22 principal components are needed in the texture representation and 15 are needed in the shape representation. Because the test face was contained in the set from which the principal components were derived the training error approaches zero when using all available principal components for the reconstruction. To evaluate the generalizability to new faces of the representation, we performed a different computation. Using a leave-one-out procedure, one face was taken out of the data base and PCA was performed on the remaining 99 faces yielding 98 principal components. Then, the single face was projected into various principal component subspaces ranging from dimensionality k=1 to 98 to yield the reconstruction X_k .

In Figure 6, the evaluation of this procedure is illustrated. The plot shows the generalization performance of the different representations in terms of the *testing error*. Very much like the training error, the testing error is defined by the mean squared difference between reconstruction and original image divided by the variance σ^2 of the whole data set:

The testing error using the pixel-based representation is never smaller than 28%, even if all 98 principal components are used for the reconstruction. A testing error of 28% is reached with only 5 principal components for the texture space and 5 principal component for the shape space. If all principal components are used, the testing error can be reduced to 6% for the shape and to 12% for the texture. A single image of a face can be used to code either one principal component in the pixel-based representation or one principal component of the shape subspace and one principal component of the texture subspace of the correspondence-based representation. Thus the information contained in five images is enough to code for 72% of the variance in a correspondence-based representation, whereas 98 images are needed in the pixel-based representation.

The reconstruction error in Figures 5 and 6 was measured in terms of the squared Euclidian distance between reconstruction and original in the respective representation. To make the three distances comparable, we normalized them with respect to the overall variance of the data base in the respective representation. Texture and shape

Fig. 8. An example of the different kind of reconstructions used. We chose an example for which the reconstructions in the correspondence-based space are relatively poor to illustrate what kind of errors still occur. a. Original face. b. Reconstructed texture combined with the original shape. Each texture was reconstructed by projecting it into the set of the 99 other textures. c. Reconstructed shape combined with the original texture. d. Reconstructed texture combined with reconstructed shape. e. Reconstruction using a the pixel-based representation.

part of the correspondence-based representation were treated separately.

To directly compare the reconstruction qualities achieved with the pixel-based and with the correspondence-based representation, we combined reconstructed texture and reconstructed shape (see Appendix D) to yield a reconstructed image. The distance between this reconstruction and the corresponding original image can be measured by means of the squared Euclidian distance in the pixel-based image space, and thus in the same space, and with the same metric as the reconstruction error of the pixel-based representations. Figure 7 shows the results of such a calculation. To reach a reconstruction error of 28% – the best that can be reached with 99 faces using a pixel-based representation – only 12 principal components have to be used in the correspondence based representation. If all principal components of the correspondence-based representation are used, a reconstruction error of 13% can be achieved.

Figure 8 gives an example of different kind of reconstructions. All are using the full set of available principal components in the respective representation. We picked out a bad example to illustrate the differences still present between the original and the different kind of reconstructions. The majority of the faces are reconstructed so well that the slight differences could not be seen in the small reproductions of Figure 8.

3. Psychophysical evaluation of the reconstructions

3.1. Purpose

The above distance measures are all based on the Euclidian distance in the different face spaces used. These distances, however, might only approximately reflect the perceptual distance used by the human face recognition system. Consider, for instance, the fact that human sensitivity to differences between faces is not at all homogeneous within the whole image. Changes in the region of the eyes are more likely to be detected than changes of the same size (with respect to any of our distance measures) in the region of the ears. Since it seems to be very difficult to formulate an image distance that exactly reflects human discrimination performance, we use human discrimination performance directly and evaluate the reconstruction quality by means of a psychophysical experiment. In the experiment, subjects were

simultaneously presented with three images on a computer screen. In the upper part of the screen, an original face from our data base was shown. Below this target face, two further images were shown. One of them was again the same target face, the other was a reconstruction. The subject was told that one of the two lower images is identical to the upper one and was instructed to find out which one it was. We measured the time they needed for this task and their accuracy.

3.2. Methods

Stimuli. The reconstructions used in this experiment were all made by projecting faces into spaces spanned by the principal components derived from all the other faces in our data base. We thus used the same "leave-one-out" procedure as described in the context of calculating the testing error (section 2.3). Four different kinds of reconstructions were used. To investigate the reconstruction quality within the texture subspace we combined reconstructed textures with the original shape. Similarly, we showed images with reconstructed shape in combination with the original texture. The third kind of reconstruction was made from a combination of reconstructed shape and reconstructed texture. Finally, we used reconstructions using the principal components derived from the pixel-based representation. In any of the four reconstruction modes, reconstructions using the first 5, 15 and all 98 principal components were shown. We chose these values because 5 and 15 principal components cover approximately one and two thirds, respectively, of the overall variance.

Design: A two-factor mixed block design was used. The first factor was a within-subject factor named QUALITY coding for the quality of the reconstruction. It had the levels REC05, REC15 and REC98, corresponding to reconstructions made by using either only 5, 15 or of all 98 principal components. The second factor was a between-subjects factor named MODE that had the four levels TEX, SHP, BTH, and PIX. TEX corresponds to trials using images with only the texture reconstructed, SHP to trials with only the shape reconstructed, BTH to trials with both reconstructed texture and shape, and PIX to trials using reconstructions in the pixel-based space (see also Fig. 9). Twenty four subjects were randomly divided into four groups, each assigned one of the levels of factor MODE. Each subject performed 3 blocks. Each block contained 100 trials using either REC05, REC15 or REC98 reconstructions. The order of the blocks was completely counterbalanced. There are six possible permutations and any of them was used once for one of the six subjects in each group. Each of the 100 faces was used exactly once in each block.

Procedure: Each stimulus presentation was preceded by a fixation cross that was presented for 1 sec. Then, three images were simultaneously presented on the computer screen. Together they covered a visual angle of 12 degrees. One of the two lower images was identical with the single upper image. The subject had to indicate which one by pressing either the left or the right arrow key on the keyboard. Subjects were instructed to respond "as accurately and as quickly as possible". The images were presented until the subject pressed one of the response keys.

3.3. Results

Figure 9 illustrates the results of this experiment. Accuracy is generally very high as expressed by the low error rates (mean: 5.9%) and differences due to factor MODE do not reach significance (two-factor ANOVA on the error rate,$F(3,20) = 1.49, p > 0.05$). In all four conditions of factor MODE, we find an increase in the error rate with the number of principal components used for the reconstruction (main effect of factor QUALITY: $F(2,40) = 14.05, p < 0.01$).

The response times are effected strongly by both factor MODE ($F(3,20) = 10.9, p < 0.01$) and factor QUALITY ($F(2,40) = 21.8, p < 0.01$). The mean response time needed to discriminate between an original image and its reconstruction in the pixel-based representation (condition PIX) is 606 msec. The mean response times in conditions TEX and SHP were 3488 msec and 3385 msec, respectively. In condition BTH the mean response time was 1872 msec. In all four conditions of factor MODE, response times increased with the number of principal components, although only very slightly in condition PIX. Note that the

time needed to identify the worst reconstruction in the correspondence-based representation (BTH, REC05) from the original is still almost twice the time needed for the best reconstruction in the pixel-based space (PIX, REC98).

4. Discussion

The results clearly demonstrate an improvement in the coding efficiency and generalization to new face images of the correspondence based image representation over pixel based techniques previously proposed (Sirovich and Kirby, 1987; Turk and Pentland, 1991). The correspondence, here computed automatically through an optical flow algorithm, allows the separation of two-dimensional shape and texture information in images of human faces. The image of a face is represented by its projection coefficients in separate linear vector spaces for shape and texture. The improvement was demonstrated theoretically as well as in a psychophysical experiment. The results of the different evaluations indicate the importance of the proposed representation for an efficient coding of face images. We have demonstrated the coding efficiency within a given set of images as well as the generalizability to new test images not in the data set from which the representations were originally obtained. In comparison to a pixel based image representation, the number of principal components necessary for the same image quality is strongly reduced. The expected error for coding a new image is less than half.

Human observers could discriminate a reconstruction derived from the pixel-based representation much faster from the original face than a reconstruction derived from the correspondence-based representation. The image quality, using a large number of basis faces, is sufficient for recognition tasks and demonstrates the importance of the representation for image synthesis and computer graphics applications. The results from the psychophysical experiments are important, since it is well known that the Euclidian distance used to optimize the reconstructions as well as to compute the principal components by itself does not in general reflect perceived image distance (Xu and Hauske, 1994). Clearly, the crucial step in the proposed technique is a dense correspondence field between images of faces seen from one view point. The optical flow technique used on our data set worked well; however, for images obtained under less controlled conditions a more sophisticated method for finding the correspondence might be necessary. New correspondence techniques based on active shape models (Cootes et al., 1995; Jones and Poggio, 1995) are more robust against local occlusions and larger distortions when applied to a known object class. Their shape parameters are optimized actively to model the target image. This technique incorporates object class specific knowledge directly into the correspondence computation step. Similarly, Hallinan (1995) demonstrated an improvement by using a low-dimensional, linear illumination model as prior knowledge.

In this paper, we used a PCA for a parameterization of the face space. Such a parameterization is optimal in the sense that the mean squared error introduced by truncating the expansion is at a minimum. The perceptual interpretation of single principal components, however, should not be overrated. A situation in which the direction of single axes is apparently arbitrary emerges when two or more Eigenvalues have about the same size. But even without this situation occurring, there is no reason to assume that the principal components are corresponding to semantically meaningful "features". The benefit of PCA is rather that it provides an orthogonal basis for the face space with the axes ordered by means of their contribution to the overall variance. The variance is based on the Euclidian distance in the face space and must not necessarily reflect perceptual distance.

Additionally, it is not clear yet how much redundant information is kept in the principal components of shape and texture. The proposed image representation, separating shape and texture informations, is not at all dependent on a PCA. A future reduction of this redundancy, based on a more extended example set of images, might lead to an even more efficient parameterization of images of faces. The main result of this paper proves that an image representation in terms of separated shape and texture is advantageous over a pixel-based image representation in any comparison we performed. These results complement other findings in which a separate texture and shape representation of three-dimensional objects in general

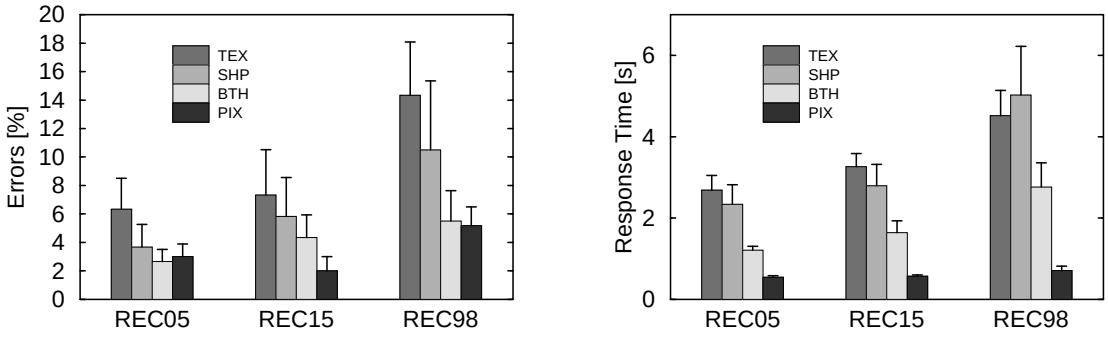

Fig. 9. Psychophysical evaluation of the different kinds of reconstructions. Error rates (a) and response times (b) are plotted. `TEX`: Reconstructed texture combined with original shape. `SHP`: Reconstructed shape combined with original texture. `BTH`: Reconstructed texture combined with reconstructed shape. `PIX`: Reconstruction in the pixel-based space. `REC05`: Reconstructions based on the first 5 principal components. `REC15`: Reconstructions based on the first 15 principal components. `REC98`: Reconstructions based on all 98 principal components.

was used for visual learning1 and enabled the synthesis of novel views from a single image (Beymer et al., 1993; Vetter and Poggio, 1996). Finally, based on our psychophysical experiments, we suggest that the correspondence based representation of faces is much closer to a human description of faces than a pixel-by-pixel comparison of images, ignoring the spatial correspondence of features.

Appendix A. Images

Images Images of 100 caucasian faces were available. The images were originally rendered for psychophysical experiments (Troje and Bülthoff, 1995) from a data base of three-dimensional human head models recorded with a laser scanner ($Cyberware^{TM}$). All faces were without make-up, accessories, and facial hair. Additionally, the head hair was removed digitally (but with manual editing), via a vertical cut behind the ears. The images were rendered from a view point 120 cm in front of the face and using ambient light only. There was still a little bit of shading present that resulted from the scanning procedure. The scanner uses approximately cylindrical illumination and the resulting texture map contains some shadows below the chin. From each face one frontal-view was rendered. The resolution of the grey-level images was 256-by-256 pixels with 8 bits per pixel. The images were aligned to a common position at the tip of the nose.

Appendix B. Image matching

The essential step in our approach is the computation of the correspondence between two images for every pixel location. That means we have to find for every pixel in the first image, e.g. a pixel located on the nose, the corresponding pixel location on the nose in the other image. Since we controlled for illumination, and since all faces are compared in the same orientation, a strong similarity of the images can be assumed and problems attributed to occlusions should be negligible. These conditions make an automatic mechanism for comparing the images of the different faces feasible (Beymer et al., 1993). Algorithms for finding correspondence between similar images are known from optical flow computation, in which points have to be tracked from one frame of a series of images to another. We used a coarse-to-fine gradient-based optical flow algorithm (Bergen et al., 1992) applied to the Laplacians of the images and following an implementation described in Bergen and Hingorani (1990). The Laplacian of the images were computed from the Gaussian pyramid adopting the algorithm proposed by Burt and Adelson (1983). Beginning with the lowest level of a resolution pyramid for every pixel (x_i) in the first image, the error term

$$E = \sum_i (I_{x_i} \Delta x_i + I_{y_i} \Delta y_i \ - \ \Delta I)^2.$$

(**B.** Image matching1)

is minimized for Δx and Δy with I_x and I_y being the spatial derivatives of the Laplacians and ΔI the difference of the Laplacians of the two images. The resulting vector field $(\Delta x, \Delta y)$ was then smoothed and the procedure was iterated through all levels of the resolution pyramid. The final resulting vector field was used as the correspondence pattern between the two images.

Appendix C. Reference face

For a consistent representation, the correspondence fields between each face and a single reference face had to be computed. In theory, any face could be used as a reference. However, small peculiarities of a face can influence the automated matching process strongly. We thus used a synthetic face as a reference face, namely the average face of our data base. This average could not be determined in one single step, but had to be calculated in an iterative procedure: Starting from an arbitrary face as the preliminary reference, the correspondence algorithm established correspondence between all other faces and this reference. The correspondence fields were only satisfying for a part of the faces. We selected these faces, calculated the mean face out of them and used it as the new reference. Repeating this procedure twice resulted in a face that was taken as the final reference face. Further iteration did not improve the correspondence fields.

Appendix D. Synthesis of an image

To generate the original image from its correspondence based representation, each pixel in the texture map had to be shifted to the new location given by the shape vector. The new location generally did not coincide with the equally spaced grid of pixels on the destination image. A common solution of this problem is known as forward warping (Wolberg, 1990). For every new pixel, we used the nearest three points to linearly approximate the pixel intensity. Not only can the images of the original faces be reproduced in this way, but also new, synthetic faces can be generated. New textures can be generated from any linear combination of already existing textures. The same can be done for the shapes. A new image can be generated combining any texture with any shape. This is possible because both are given in the coordinates of the reference image.

References

1. H. Abdi, D. Valentin, B. Edelman and A.J. O'Toole. More about the difference between men and women: Evidence from linear neural networks and the principal component approach. *Perception*, 24:539-562, 1995.
2. N. Ahmed and M. H. Goldstein. Orthogonal Transforms for Digital Signal Processing. Springer Verlag, New York, 1975.
3. J. A. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical-model-based motion estimation. In *Second European Conference on Computer Vision*, G. Sandini, ed., (Springer Verlag, Berlin, 1992), pp. 237-252.
4. J.R. Bergen and R.Hingorani. Hierarchical motion-based frame rate conversion. Technical report, David Sarnoff Research Center Princeton NJ 08540, 1990.
5. P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31:532-540, 1983.
6. D.Beymer and T.Poggio. Image representation for visual learning. *Science* 272:1905-1909, 1996.
7. D.Beymer, A.Shashua, and T.Poggio, Example-based image analysis and synthesis. A.I. Memo No. 143 1, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.
8. T.F. Cootes, C.J. Taylor, D.H. Cooper, and J.Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38-59, 1995.
9. N. Costen, I. Craw, G. Robertson and S. Akamatsu. Automatic face recognition: What representation. in *Computer Vision - ECCV'96 Lecture Notes in Computer Science 1064*, B. Buxton and R. Cippola, ed., (Springer, Berlin, 1996), pp. 504-513.
10. I. Craw and P. Cameron. Parameterizing images for recognition and reconstruction. in *British Machine Vision Conference*, P. Mowforth, ed., Springer Verlag, 1991 pp. 367-370.
11. C. A. Feingold. The influence of environment on identification of persons and things. *Journal of Criminal Law & Police Science*, 5:39-51, 1914.
12. M.Jones and T.Poggio. Model-based matching of line drawings by linear combination of prototypes. in *Proceedings of the 5th International Conference on Computer Vision*, IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 531-536.

13. P. W. Hallinan. A deformable model for the recognition of human faces under arbitrary illumination. Doctoral Dissertation. Harvard University, 1995.
14. P. J. B. Hancock, A. M. Burton, and V. Bruce. Face processing: Human perception and principal components. *Memory & Cognition*, 24:26-40, 1996.
15. M.Kirby and L.Sirovich. Application of the Karhunen-Loewe procedure for characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103-109, 1990.
16. D. I. Perrett, K. A. May and S. Yoshikawa. Facial shape and judgements of female attractiveness. *Nature* , 368:239-242, 1994.
17. L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4:519-554, 1987.
18. A.J. O'Toole, H. Abdi, K.A. Deffenbacher, and D.Valentine. Low-dimensional representation of faces in higher dimensions of the face space. *Jounal of the Optical Society of Amerika A*, 10:405-411, 1993.
19. A. J. O'Toole, H. Abdi, K. A. Deffenbacher and J. C. Barlett. Classifying faces by face and sex using an autoassociative memory trained for recognition. In *Proceedings of the thirteenth annual conference of the Cognitive Science Society*, K. J. Hammond and D. Gentner, eds. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1991), pp.847-851.
20. A. J. O'Toole, K. A. Deffenbacher, D. Valentin and H. Abdi. Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, 22:208-224, 1994.
21. N. Troje and H. H. Bülthoff. Face recognition under varying pose: The role of texture and shape. *Vision Research*, 36:1761-1771, 1995.
22. M. Turk and A. Pentland Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71-86, 1991.
23. D. Valentin, H. Abdi, A. J. O'Toole, and G. W. Cottrell. Connectionist models of face processing: A survey. *Pattern recognition*, 27:1209-1230, 1994.
24. T. Vetter and N. F. Troje. Separation of texture and two-dimensional shape in images of human faces. in *Mustererkennung 1995*, S. Posch, F. Kummert, and G. Sagerer, eds, Springer Verlag, 1995, pp. 118-125.
25. T. Vetter. Synthesis of novel views from a single face image. Technical Report No.26, Max-Planck-Institut fr biologische Kybernetik Tbingen, Germany, 1996.
26. T. Vetter and T. Poggio. Image synthesis from a single example image. in *Computer Vision - ECCV'96 Lecture Notes in Computer Science 1064*, B. Buxton and R. Cippola, ed., (Springer, Berlin, 1996), pp. 652-659.
27. Georg Wolberg, Image Warping. IEEE Computer Society Press, Los Alamitos CA, 1990.
28. W. Xu and G. Hauske. Picture quality evaluation based on error segmentation. *Visual Communications and Image Processing*, 2308:1-12, 1994.