Max-Planck-Institut
für biologische Kybernetik
Arbeitsgruppe Bülthoff

Spemannstraße 38 • 72076 Tübingen • Germany

Technical Report No. 41 October 16, 1996

Representations of human faces

Nikolaus F. Troje and Thomas Vetter

Abstract

Several models for parameterized face representations have been proposed in the last years. A simple coding scheme treats the image of a face as a long vector with each entry coding for the intensity of one single pixel in the image (e.g. Sirovich & Kirby 1987). Although simple and straightforward, such pixel-based representations have several disadvantages. We propose a representation for images of faces that separates texture and 2D shape by exploiting pixel-by-pixel correspondence between the images. The advantages of this representation compared to pixel-based representations are demonstrated by means of the quality of low-dimensional reconstructions derived from principal component analysis and by means of the performance that a simple linear classifier can achieve for sex classification.

This document is available as /pub/mpi-memos/TR-041.ps via anonymous ftp from **ftp.mpik-tueb.mpg.de** or from the World Wide Web, **http://www.mpik-tueb.mpg.de/projects/TechReport/head.html**

1 Introduction

Few object classes have been examined as extensively as the class of human faces. Investigations have been carried out in several different scientific disciplines. Psychologists and human-ethologists are interested in the way our perceptual system deals with faces when performing tasks such as recognizing individual persons or rating gender, age, attractiveness or facial expression. Computer scientists work with human faces in different areas. In machine vision, much effort has been put into constructing artificial face recognition systems that are able to generalize between different appearances of the same face. In computer graphics, faces play an important role for modelling and animation purposes. Faces also make interesting objects for the study of efficient coding schemes, as this information is relevant for video conferencing and telecommunication.

Most of the problems that have to be solved in face recognition are shared by other visual object recognition tasks. In the following paragraphs, we will discuss these problems in more general terms, speaking about "object recognition" rather than about "face recognition". We will come back to human faces, however, when discussing concrete examples for different representations.

The input information for our brain are the retinal images. Artificial systems usually also have to rely on two-dimensional images. Object recognition can be described as the process of finding an appropriate measure for the distance between stored representations and an incoming image. If the task is object identification, then such a measure should provide a relatively small distance between views of the same objects, regardless of orientation, illumination and other scene attributes not related to the object's identity. It also should provide a relatively large distance between views of different objects even if they share common properties, such as illumination or orientation. If the task is to estimate the orientation, the size, or the colour of an object, then a distance measure is needed that clusters images of different objects with the same orientation, size or colour, irrespective of their identity.

The search for an efficient distance measure depends strongly on the choice of an appropriate representation of the images. The simplest and most straightforward representation of an image is a representation that will be called *pixel-based representation* throughout this paper. In this representation, the image is coded by simply concatenating all the intensity values of a number of sample points into a large vector. The sample points can correspond to the regular grid of pixels provided by a digitized image on the computer screen, or they can correspond to the photoreceptor array in our retina. A 256x256 pixel image thus results in a 65536 dimensional vector located in a vector space of equal dimensionality.

A distance measure in such a simple pixel-based image space must have a fairly complex structure to provide for object identification or classification. Imagine only the locations of two views of the same object in such a space that differ only by a slight translation of the object in the image. Although a human observer could hardly perceive the difference between the two images, their locations in pixel space would be very distinct from each other.

A space that much better fits the requirements of object (and face) recognition is a space in which the objects are coded by means of complex high-level features. If the features are chosen such that they are diagnostic for one attribute (such as identity) but invariant to others (such as illumination or orientation), it is easy to construct simple metrics that cluster views of identical objects irrespective of the viewing conditions. Identification as well as classification with respect to other object attributes can then be carried out by using simple linear classifiers.

Contrasting the pixel-based representation with a representation based on high-level features illustrates the trade-off between the complexity of the representation and the complexity of an appropriate distance measure. Using a simple representation of the image of an object requires sophisticated distance measures and complex classifiers, whereas with a complex representation, simpler distance measures and linear classifiers may be sufficient.

The transformation from the pixel space into a feature space involves complex operations, often including *a priori* information about the object class the system is dealing with. A crucial point of feature-based representations is how to define and how to extract relevant features from the images. The same set of indexed features should be available in all images in order to be able to compare them. This means that correspondence between the features of different images has to be established. As a consequence, a high-level feature space tends to be model specific. Searching for a nose, for instance, only makes sense if the algorithm is confronted with a face.

A disadvantage of high-level feature spaces may be a loss of information due to the feature extraction process. A representation coding features such as the size of eyes, nose and mouth, distances and distance ratios between these features, etc., may serve for identification and classification tasks, but it might be difficult to reconstruct the original image from this information.

In the past few years, different researchers have developed feature-based representations of human faces (Beymer, Shashua, & Poggio, 1993; Beymer & Poggio, 1996; Costen et al., 1996; Craw & Cameron, 1991; Hancock et al. 1996; Perrett, May, & Yoshikawa, 1994; Vetter, 1996; Vetter & Troje, 1995). The features used for establishing correspondence span the whole range between semantically 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. Establishing correspondence between the images has been done by either hand-selecting a limited set of features or by using adapted optical flow algorithms that define correspondence on the single pixel level.

In this paper, we will present a particular way of establishing a representation of human faces that we have developed (Vetter & Troje, 1995). This representation is a feature-based representation, but nevertheless retains all of the information contained in the original image. It can thus be used not only for recognition purposes but also for modelling new faces. Since this representation is based on a pixel-by-pixel correspondence between two images, we call it a *correspondence-based* representation.

We will compare this representation with a simple pixel-based representation and evaluate it by means of three different issues that we consider to be important criteria for a representation flexible enough to serve many of the purposes occurring when processing human faces. These criteria are:

- The set of faces should be convex. If two vectors in the corresponding space are natural faces, then any vector on the line connecting these vectors should also correspond to a proper face.
- The representation should provide an efficient coding scheme. Redundancies should be reduced.
- Important attributes (identity, sex, age, facial expression, orientation) should be easily separable.

That convexity is fulfilled for the correspondence-based representation will become directly evident in the next section, in which we develop our representation step by step starting from a simple pixel-based representation. In section 3, we address the question of coding efficiency by evaluating low-dimensional reconstructions based on principal component analysis. In section 4, we use sex classification as an example of a classification task. The generalization performance of a simple linear classifier using the different representations as input is investigated.

2 Developing a correspondence-based representation

As mentioned above, it would be desirable for a number of different purposes to develop a representation of faces that makes it possible to treat them as objects in a linear vector space. Such a “face space” is the basis for developing metrics that correspond to differences in identity, gender, age, etc.

An image of a face (as any other image) can be coded in terms of a vector that has as many coordinates as the image has pixels. Each coordinate codes the intensity of one particular pixel in the image (Figure 1), so that the vector contains all of the information in the image. The space spanned by such image vectors, however, has some very unpleasant properties. An important property of a linear space is the existence of an addition and a scalar multiplication which define 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 rather 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. As can be seen from Figure 2, the mean of two faces in this representation looks better than it did with no alignment. Nevertheless, there are still plenty of errors in the image. The eyes look good now, but the mean face contains two mouths and most other parts do not match either. To match the mouths while still keeping the eyes matched, a scaling operation is needed in addition

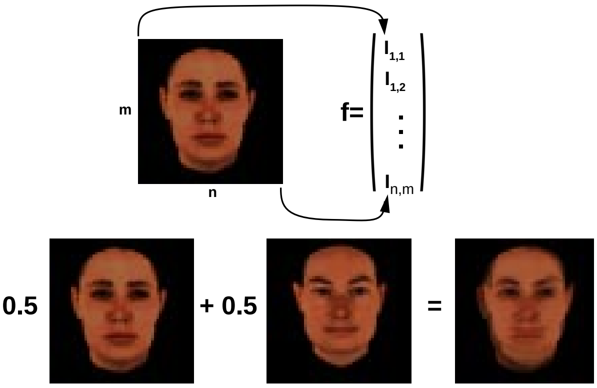

Figure 1: Pixel-based representation. The image of a face is coded as a vector by concatenating all the pixel values. The mean of two faces in this representation does not yield a "mean face" but rather a superposition of two single faces.

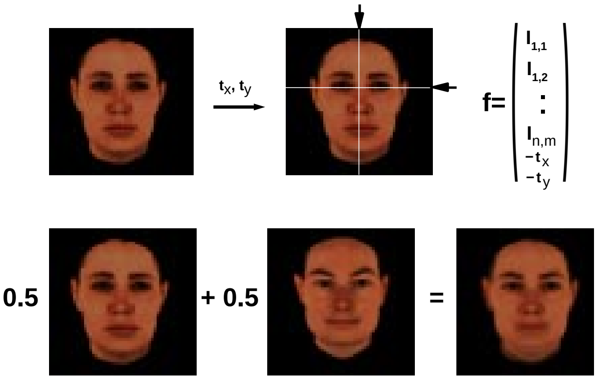

Figure 2: Here, the images are first aligned with respect to a common position of the symmetry plane and a commen height of the eyes. Then they are coded as in the pixel-based representation. The two factors describing the necessary translation are also added. The mean of two such representations still contains most features twice.

to the translation. This scaling must be done independently in the horizontal and vertical directions, leading not only to a change in size, but also to a distortion of the face.

Figure 3 shows the representation after this improved alignment. The first part of the vector encodes the image resulting from the alignment process. The last four coefficients account for the translation and scaling operations needed for the alignment. Note that we did not enter the translation and scaling factors themselves but their inverse values. The original image can thus be reconstructed from the vector representation by drawing the image encoded in the first part of the vector and then performing translation and scaling operations according to the last part of the vector. The simple mean of the two sample faces using this latter representation is much better now.

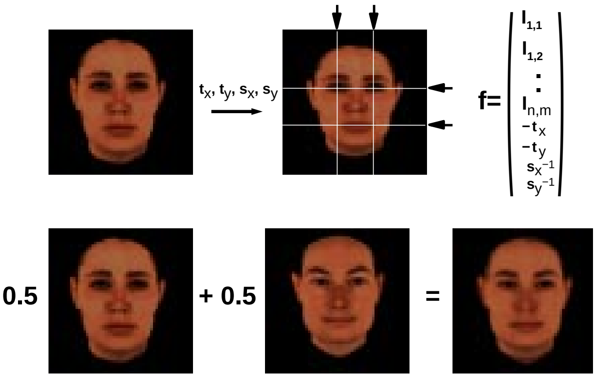

Figure 3: In this representation, the images are first aligned using two translations and two scaling operations. The image resulting from this alignment is coded together with the parameters describing the alignment. Although hardly visible in these small reproductions, the mean face still contains errors due to misalignment.

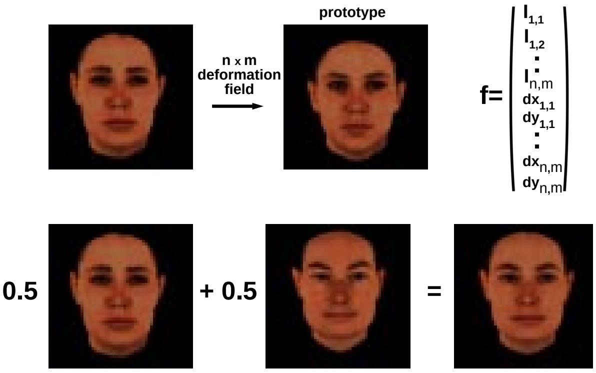

Figure 4: Correspondence-based representation. The images are deformed to match a common prototype. The vector contains the image after that deformation and the deformation field itself. The mean of two faces in this representation is one single face consisting of the mean texture on the mean shape.

Not only are the eyes aligned, but, at least roughly, the mouths are aligned as well. However, a more careful look at the images still reveals significant errors. Since the shapes of the two mouths were very different, a closer inspection shows that there are still two superimposed mouths rather than one. The noses are not aligned and other features including the outline of the face are stillnotmatched.

Continuing with this approach leads to the correspondence-based representation described in more detail by Vetter and Troje (1995). Rather than allowing only simple image operations such as translation or scaling, any image deformation can be used to align a sample face with a second face that serves as a common prototype. Figure 4 illustrates the resulting representation. The first part of the vector again codes the image resulting after aligning the face to the prototype. The second part of the vector codes the deformation that has to be applied to this image in order to recover the original image. The deformation is not encoded in a parameterized form. Rather, it simply describes for each pixel *i* in the image the displacement vector (dx_i, dy_i) necessary to match the corresponding pixel in the original image. The rule that decodes the image from this representation simply reads as follows: Draw the image given by the first part of the vector and apply the deformation field given by the second part of the vector. We will refer to the first part of the vector as the *texture* of the face and to the second part as the *shape* of the face.

The correspondence-based representation is completely smooth and convex in the sense that a linear combination of two or more faces cannot be identified as being synthetic. The mean of two faces results in a face with the mean texture applied to the mean shape. In computer graphics this hybrid face is often referred to as the *morph* between the two faces.

In fact, any inner[1] linear combination of existing textures reveals a new valid texture and any inner linear combination of existing shapes reveals a new valid shape. Furthermore, any valid texture can be combined with any valid shape to reveal a new face. The subspaces coding for texture and for shape can thus be treated independently.

A critical element of this approach is establishing pixel-by-pixel correspondence between the sample face and the common prototype. We used a coarse-to-fine gradient-based optical flow algorithm (Adelson & Bergen, 1986) applied to the Laplacians of the images 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). For more details, see Vetter and Troje (1995).

3 The quality of low-dimensional reconstructions

3.1 Images

The images of the faces were generated from a data base of 100 three-dimensional head models obtained by using a 3D laser scanner. All head models were sampled from persons between 20 and 40 years, without make-up, facial hair or accessories such as earrings or glasses. Half of them were male, and the other half were female. Head hair was digitally erased from the models. For details about the acquisition and the preprocessing of the models, see Troje and Bülthoff (1996).

The images showed the faces from a frontal view. The orientations of the head models in 3D space were aligned to each other by minimizing the sum-squared distances between corresponding locations of a set of selected features such as the pupils, the tip of the nose, and the corners of the mouth. Images were black and white and had a size of 256x256 pixels with a resolution of 8 bits.

3.2 Principal component analysis

Principal component analysis (PCA) is a tool that has been widely used to reduce the dimensionality of a given data set. PCA is based on the Karhunen-Loeve expansion -- a linear transformation resulting in an orthogonal basis with the axes ordered according to their contribution to the overall variance of the data set. Truncating the expansion yields low-dimensional representations of the data with a minimized mean squared error (Ahmed & Goldstein, 1975).

PCA was first used with images of faces by Sirovich and Kirby (1987) and has been applied successfully to different tasks, such as face recognition (Turk & Pentland, 1991; O'Toole, Abdi, Deffenbacher, & Valentin, 1993; Abdi, Valentin, Edelman, & O'Toole, 1995) and gender classification (O'Toole, Abdi, Deffenbacher, & Barlett, 1991). In all of these investigations, PCA was applied directly to the pixel-based representation of images, which were only aligned by means of simple transformations (translation, Sirovich and Kirby also used scaling) that do not change the character of the face.

Vetter and Troje (1995) applied PCA to the correspondence-based representation of faces. For the present investigation, we used the same technique, applying PCA separately to the subspaces that code for the texture and for the shape of the faces, respectively. In addition, we ran PCA on the images themselves.

3.3 Theoretical evaluation of the reconstructions

PCA yields an orthogonal basis with the axes ordered according to their overall variance. The principal components equal the eigenvectors of the covariance matrix of the data. The corresponding eigenvalues are equal to the variances

[1] That means, a linear combination in which all coefficients sum up to one.

along each component. The decrease of the variances associated with the principal components indicates the applicability of PCA for dimensionality reduction.

In Figure 5a, we plotted one minus the relative cumulative variance accounted for by the first *k* principal components for the three different PCAs. The relative cumulative variances were calculated by successively summing up the first *k* eigenvalues υ_i and dividing them by the sum of all eigenvalues:

$$\text{training error}_k = 1 - \frac{\sum_k \upsilon_i}{\sum_n \upsilon_i} \qquad (1)$$

This term is equivalent to the expected value for the mean squared distance between a reconstruction X_k and the original image X divided by the overall variance σ^2. By X_k we denote the reconstruction yielded using only the first *k* principal components.

$$\begin{aligned}\text{training error}_k &= 1 - \frac{\sum_k \upsilon_i}{\sum_n \upsilon_i} \\ &= \frac{1}{\sigma^2(n-1)}\sum_n (X_k - X)^2\end{aligned} \qquad (2)$$

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

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 how well the representation generalizes to new faces, we performed a leave-one-out procedure in which 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 . This was done for every face in the data base.

In Figure 5b, the quality of the reconstructions resulting from this procedure is illustrated. The plot shows the generalization performance of the different representations in terms of the *testing error.* 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:

$$\text{testing error}_k = \frac{1}{n\sigma^2}\sum_n (X_k - X)^2 \qquad (3)$$

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 components 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 errors in Figures 5a and 5b were 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 parts 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 to yield a reconstructed image. This was done by applying the deformation field, coded in the reconstructed shape to the images coded in the reconstructed texture. 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 5c shows the results

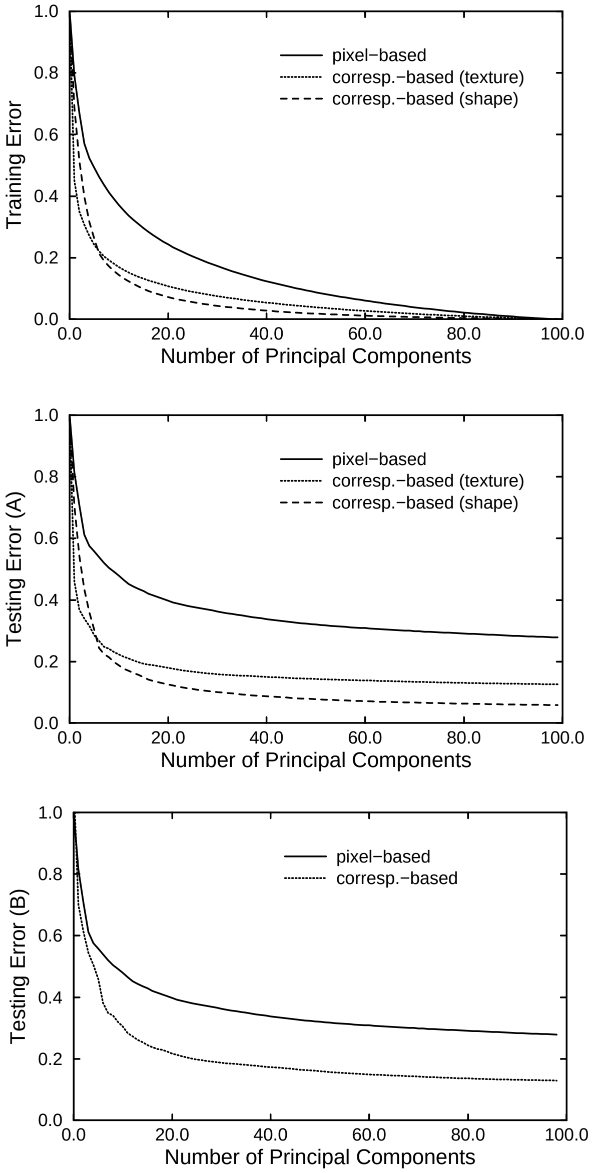

Fig. 5: (a) Training error. In this diagram, one minus the relative cumulative variance has been plotted. The cumulative variance is equal to the mean of the squared Euclidian distance between the original face and reconstructions derived by truncating the principal component expansion. The calculation was performed for the two parts of the correspondence-based representation and for the pixel-based representation.

(b) 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.

(c) 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).

of this calculation. To achieve 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.

3.4 Psychophysical evaluation of the reconstructions

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 original target face, the other was a reconstruction of it. The subjects indicated which of the two lower

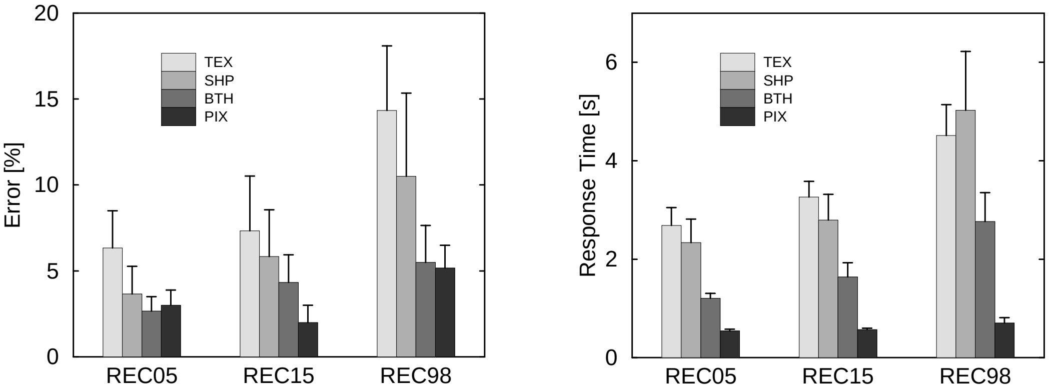

Fig. 6: Psychophysical evaluation of the different kinds of reconstructions. Error rates (a) and response times (b) are plotted. TEX: Reconstructed texture combined with original shape. SHP: Reconstructed shape combined with original texture. BTH: Reconstructed texture combined with reconstructed shape. 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.

images was identical to the upper one. The time they needed for this task makes an issue about the reconstruction quality.

Methods

The reconstructions tested 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 (see previous section). 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 each 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.

A two-factor mixed block design was used. The first factor was a within-subject factor named QUALITY that coded 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.

Twenty four subjects were randomly divided into four groups, each assigned to one of the levels of the 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 each 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.

Each stimulus presentation was preceded by a fixation cross that was presented for 1 sec. Then, the three images were simultaneously presented on the computer screen. Together they covered a visual angle of 12 degrees. The subject indicated which of the two bottom images was identical with the image on the top 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. We measured the subjects error rate as well as the time they needed to perform the task.

Results

Figure 6 illustrates the results of this experiment. Accuracy was generally very high as expressed by the low error rates (mean: 5.9%) and differences due to the factor MODE did not reach significance (two-factor ANOVA on the error rate, $F_{3,20} = 1.49$, $p > 0.05$). We found an increase in the error rate with the number of principal components used for the reconstruction (main effect of the factor QUALITY: $F_{2,40} = 14.05$, $p < 0.01$) and no interaction between the two factors.

The response times were effected strongly by both the factor MODE ($F_{3,20} = 10.9$, $p < 0.01$) and the factor QUALITY ($F_{2,40} = 21.8$, $p < 0.01$). The interaction between the factors was marginally significant ($F_{6,40} = 2.6$, $p < 0.05$). The mean response time needed to discriminate between an original image and its reconstruction in the pixel-based representation (condition PIX) was 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 the 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 was still almost twice the time needed for the best reconstruction in the pixel-based space (PIX, REC98).

3.1 Reconstruction quality and coding efficiency

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 (Kirby & Sirovich, 1990; Turk & Pentland, 1991). The correspondence, here computed automatically using 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 computationally as well as in a psychophysical experiment.

The results of the different evaluations indicate the utility of the proposed representation as 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 contained in the data set from which the representations were originally obtained. In comparison to a pixel-based image representation, the number of principal components needed for the same image quality is strongly reduced.

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 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 & Hauske, 1994).

4 Sex classification

4.1 Purpose

According to the criteria developed in section 1, we would expect an efficient and flexible representation of faces to cluster together groups of images that share common attributes. Images showing the same individual should be closer to each other than images of different faces, according to some simple metric. Also, images of faces of the same age, gender or race should cluster according to other metrics.

In this section, we investigate how well a simple linear classifier can distinguish between male and female faces, using either the pixel-based or the correspondence-based representation as an input. To examine how robust the respective representations are against miss-alignment of the faces, we generated different image sets differing in the degree of their mutual alignment.

4.2 Material and Methods

An extended data base consisting of 200 three-dimensional head models was used for these simulations. Half of them were male and half of them were female. Preprocessing of the models was performed as described in section 3.1. The initial alignment was also performed as described previously and frontal view images were rendered. In addition, we rendered two other sets of images by systematically misaligning the heads. For the first set, we applied small translations adding Gaussian noise with a standard deviation of 0.5 cm (corresponding to 5 pixels in the image) to the position of the head in the image plane. For the second set, we misaligned the faces by applying small rotations in 3D space before rendering the images. We added Gaussian noise with a standard deviation of 3 degrees to the orientation of the head around the vertical axis and around the horizontal axis perpendicular to the line of sight.

Correspondence-based representations were computed from the aligned face set and from the set that had been misaligned by rotations in 3D. The correspondence-based representation of the image set misaligned by small translations was derived directly by adding the constant translation vector to the flow field.

This procedure yielded nine different data sets: Each of the three sets of images existed now in terms of the pixel-based representation and in terms of the texture and the shape part of the correspondence-based representation.

Sex classification was performed on each of the nine data sets in the following way:

The 200 faces were randomly divided into two groups (A and B), each containing 50 males and 50 females. Two simulations were run. In the first, group A served as the training set and group B as the test set. In a second simulation the two groups were exchanged using group B for training and group A for testing. Each simulation began with the calculation of a principal component analysis on the training set. Then both training and test sets were projected on the first 50 principal components, yielding 50 coefficients for each face. We also tested the performance of the classifier when using all 99 principal components, but the results were never better than with 50 components.

The 50 coefficients were used as input for a linear classifier. The classifier itself was formulated as a linear system:

$$\mathbf{a}^{train} = \omega\mathbf{P}^{train} + \omega_0 \qquad (4)$$

$\mathbf{P}^{train}$ is a matrix containing the coefficients of the *i*th face of the training set in the *i*th column. $\mathbf{a}^{train}$ is a row vector containing the desired outputs ($\mathbf{a}_i^{train} = 1$ if $\mathbf{P}_i^{train}$ male, $\mathbf{a}_i^{train} = -1$ if $\mathbf{P}_i^{train}$ female). ω is the row vector containing the weights corresponding to the first 50 principal components and ω_0 accounts for a constant bias. The coefficients ω_i were optimized by using singular value decomposition in order to minimize the sum-squared error between the desired and the actual output. Note that this is equivalent to training a simple perceptron with linear transfer function.

After training, the test set was projected on the vector ω:

$$\hat{\mathbf{a}} = \omega\mathbf{P}^{test} + \omega_0 \qquad (5)$$

The output $\hat{\mathbf{a}}$ was compared with the desired output $\mathbf{a}^{test}$ to yield the error rate. An error was recorded if $\mathrm{sgn}(\mathbf{a}_i^{test}) \neq \mathrm{sgn}(\hat{\mathbf{a}})$.

In addition, we ran three further simulations to classify the three different image sets by using the coefficients corresponding to the first 25 principal components of the shape subspace together with the first 25 coefficients for the texture subspace as input data.

For all the simulations, the mean error of the two reciprocal simulations (exchanging training and test sets) is reported.

4.3 Results

In Figure 7, the generalization errors resulting from the classification experiments are presented. Using the images showing the faces previously aligned in 3D, classification on the pixel-based representation yielded a relatively low error rate of 4%. Using only the texture subspace of the correspondence-based representation, the error was somewhat higher (5.5%). With only the shape subspace, the error rate was 3%. The best classification (error rate 2%) was obtained when combining the coefficients corresponding to the first 25 principal components of the texture subspace with the coefficients corresponding to the first 25 principal components of the shape subspace.

Using images of faces that were misaligned, the classification performance for the pixel-based representation dropped significantly. In the first example in which a misalignment was introduced

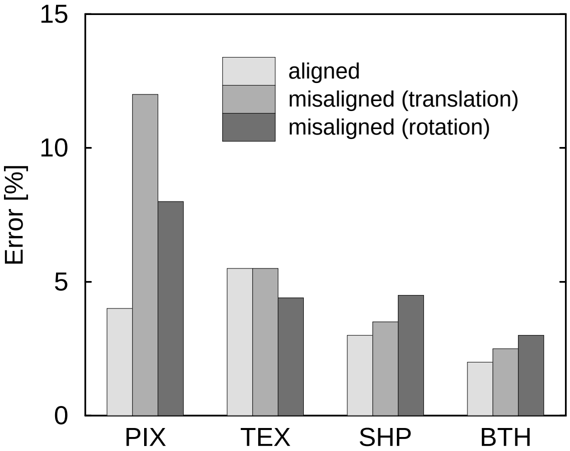

Figure 7: Generalization errors of the sex classification experiments on three different image sets (see text). As input for the classifier, the coefficients corresponding to the first 50 principal components derived from the pixel-based representation (PIX), the “texture” part (TEX) and the “shape” part (SHP) of the correspondence-based representation were used. Additionally, an input vector consisting of the first 25 principal components of “texture” and “shape” was used (BTH).

by applying a small translation to the images, the error rate was 12%. In the other example in which the misalignment was due to small rotations in depth, an error rate of 8% was obtained. The correspondence-based representation was much less effected by misalignment. The error rates for the classification using only the texture stayed constant, the ones for the classification using the shape increased only slightly. If a combination of the first principal components of texture and shape was used, the error rates also only slightly increased.

4.4 Discussion

The advantage of the correspondence-based representation is striking when using images of faces that are misaligned. The good performance on the classification for the pixel-based aligned images, however, shows that the full pixel-by-pixel correspondence is not needed for sex classification. What is needed is only enough information to perform an alignment of the heads in space. Sex classification is probably a relatively easy task compared with other classification tasks such as the classification of facial expression or the identification of a person. For these latter tasks, the advantage of the correspondence-based representation is expected to be even more pronounced and an optimal rigid alignment in 3D is probably not sufficient.

5 General Discussion

We contrasted the properties of a correspondence-based representation of images of human faces with pixel-based techniques. The motivation behind developing the correspondence-based representation was the lack of convexity of the pixel-based representation. The correspondence-based representation copes with this problem by employing pixel-by-pixel correspondence to perfectly match the images. This results in a representation separating texture and shape information.

We compared low-dimensional reconstructions derived from correspondence-based and pixel-based representations to demonstrate the advantage of the correspondence-based representation for efficient coding and modelling. Finally, we tested the different representations in a simple classification task. We trained a linear network to classify the sex of faces in a training set and tested for generalization performance using a separate testing set of faces.

Clearly, the crucial step in the proposed technique is a dense correspondence field between the images of the faces. 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 & 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. These techniques thus incorporate knowledge specific to the object class directly into the correspondence computation.

The main result of this paper is that an image representation in terms of separated shape and texture is superior to a pixel-based image representation for performing many useful tasks. Our results complement other findings in which a separate texture and shape representation of three-dimensional objects in general was used for visual learning (Beymer & Poggio, 1996), enabling the synthesis of novel views from a single image (Vetter & 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, which disregards the spatial correspondence of features.

References

Abdi, H., Valentin, D., Edelman, B. and O’Toole, A.J. (1995) “More about the difference between men and women: Evidence from linear neural networks and the principal component approach”, *Perception 24*:539-562.

Adelson, E.H. and J.R. Bergen (1986) “The extraction of spatiotemporal energy in human and machine vision”, *Proc. IEEE Workshop on Visual Motion*, Carlston, pp. 151-156.

Ahmed, N. and Goldstein, M.H. (1975) *Orthogonal Transforms for Digital Signal Processing*, New York: Springer.

Bergen, J.R. and Hingorani, R. (1990) “Hierarchical motion-based frame rate conversion”, *Technical report, David Sarnoff Research Center Princeton NJ 08540*.

Beymer, D. and Poggio, T. (1996) “Image representation for visual learning”, *Science 272*:1905-

1909.

Beymer, D., Shashua, A. and Poggio, T (1993) “Example-based image analysis and synthesis”, *Artificial Intell. Lab., Massachusetts Inst. Technol., Cambridge, Rep. A.I.M. 1431.*

Burt, P.J. and Adelson, E.H. (1983) “The Laplacian pyramid as a compact image code”, *IEEE Transactions on Communications 31*:532-540.

Cootes, T.F., Taylor, C.J., Cooper, D.H. and Graham, J. (1995) “Active shape models - their training and application”, *Computer Vision and Image Understanding 61*:38-59.

Costen, N., Craw, I., Robertson, G. and Akamatsu, S. (1996) “Automatic face recognition: What representation”, in: B. Buxton and R. Cippola, eds., *Computer Vision - ECCV’96, Lecture Notes in Computer Science 1064*, Cambridge UK: Springer, pp. 504-513.

Craw, I. and Cameron, P. (1991) “Parameterizing images for recognition and reconstruction”, *Proc. British Machine Vision Conference*, pp. 367-370.

Hancock, P.J.B., Burton, A.M. and Bruce, V. (1996) “Face processing: Human perception and principal components analysis”, *Memory and Cognition 24*:26-40.

Jones, M. and Poggio, T. (1995) “Model-based matching of line drawings by linear combination of prototypes”, in: *Proceedings of the 5th International Conference on Computer Vision*, pp. 531-536.

Kirby, M. and Sirovich, L. (1990) “Application of the Karhunen-Loewe procedure for characterization of human faces”, *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*:103-109.

Perrett, D.I., May, K.A. and Yoshikawa, S. (1994) “Facial shape and judgements of female attractiveness”, *Nature 368*:239-242.

O’Toole, A.J., Abdi, H., Deffenbacher, K.A. and Valentine, D. (1993) “Low-dimensional representation of faces in higher dimensions of the face space”, *Journal of the Optical Society of America A 10:*405-411.

O’Toole, A.J., Abdi, H., Deffenbacher, K.A. and Barlett, J.C. (1991) “Classifying faces by face and sex using an autoassociative memory trained for recognition”, in: K.J. Hammond and D. Gentner, eds., *Proceedings of the thirteenth annual conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 847-851.

Sirovich L. and Kirby, M. (1987) “Low-dimensional procedure for the characterization of human faces”, *Journal of the Optical Society of America A 4*:519-554.

Troje, N. and Bülthoff, H.H. (1995) “Face recognition under varying pose: The role of texture and shape”, *Vision Research 36*:1761-1771.

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

Vetter, T. and Poggio, T. (1996) “Image synthesis from a single example image”. In B. Buxton and R. Cippola, eds., *Computer Vision - ECCV’96, Lecture Notes in Computer Science 1064*, Cambridge UK: Springer, pp. 652-659.

Vetter, T. and Troje, N. (1995) “Separation of texture and two-dimensional shape in images of human faces”, in: S. Posch, F. Kummert, and G. Sagerer, eds., *Mustererkennung 1995*, New York: Springer, pp. 118-125.

Vetter, T. (1996) “Synthesis of novel views from a single face image”, *Max-Planck-Institut für biologische Kybernetik, Tübingen, Germany, Technical Report 26.*

Xu W. and Hauske, G. (1994) “Picture quality evaluation based on error segmentation”, *Proc. SPIE, Visual Communications and Image Processing 2308*:1-12.