View-based recognition under illumination changes using local features

Proceedings of CVPR'01

Christian Wallraven*
Max-Planck-Institute for Biological Cybernetics
72076 Tübingen, Germany
christian.wallraven@tuebingen.mpg.de

Heinrich Bülthoff
Max-Planck-Institute for Biological Cybernetics
72076 Tübingen, Germany
heinrich.buelthoff@tuebingen.mpg.de

Abstract

We present a view-based face recognition system which combines elements from both feature-based and appearance-based approaches to increase recognition performance under illumination changes. It uses corners and their local neighborhood at several scales to construct local features which form the representation of each face image. Matching feature sets takes into account both configurational and appearance-based similarity. We present recognition results on a highly realistic synthetic face-database demonstrating the system's ability to tolerate illumination changes. In addition, the proposed framework agrees well with current findings from psychophysics.

1 Introduction

Approaches for face recognition can be very broadly characterized into two categories ([3], for an overview on face recognition see also [16]): The first category is the feature-based approach, which uses specific features extracted from the image normally corresponding to distinct features of the face, such as eyes, nose, etc. These are then used in a next step to calculate configurational distances, such as the distance between the eyes and the wideness of the mouth (e.g., [3]). The second category is the appearance-based approach, which uses the raw or processed pixel data of the image as a whole (e.g., the eigenface technique [11]).

The feature based approach typically results in a low-dimensional description (on the order of 100 dimensions) of the face and thus presents a rather high level of abstraction. Furthermore, the extracted features and their relationships are in general rather insensitive to illumination variations, since they correspond to precisely defined points in the face. This approach, however, relies heavily on a good feature extraction technique, which is difficult to implement automatically and thus has often to be done manually.

The appearance-based approach on the other hand is easy to implement since processing usually can be done automatically (such as calculating the eigen-representation). The drawback of this kind of approach is that it needs many training images to yield invariant recognition under the large range of viewing conditions typically found in natural object recognition tasks, such as changes in viewpoint, illumination and expression.

In this paper, we propose a recognition approach, which combines elements of these two approaches into a view-based recognition system and which we tested using a face database of laser-scanned 3D heads. Furthermore, the system shows properties which agree well with current findings for face and object recognition from psychophysics. In addition, the system demonstrates that good recognition performance under illumination changes is possible without using any high-level a-priori knowledge such as illumination models [1] or generic 3D face models [2].

In Section 2, the database used for recognition experiments is described, Section 3 presents the image representation of our system and the matching algorithm which is used. In Section 4 results of recognition experiments on that database are given. Section 5 discusses the results with respect to psychophysical findings and points to future extensions to the framework.

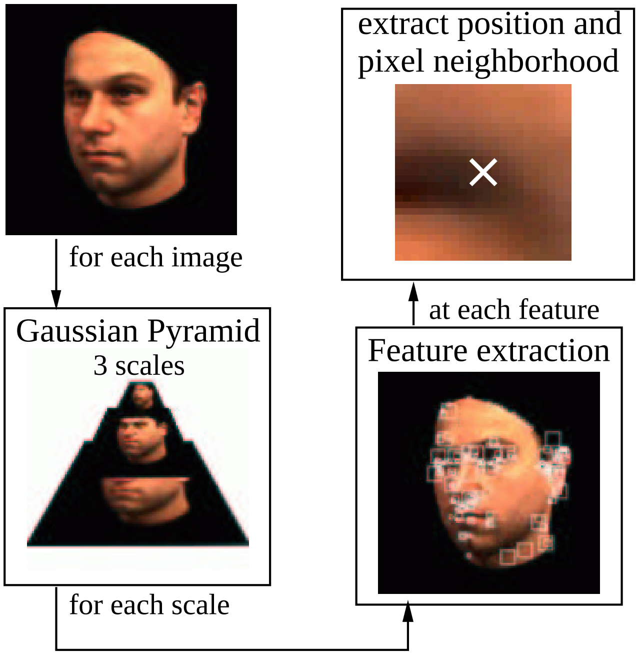


Figure 1. Representation of face images.

2 Test Database

For recognition experiments, we created a test database containing 30 individuals from the face database of the Max-Planck-Institute. The face database consists of highly realistic 3D laser-scans from 200 individual heads with both shape and texture data[1].

For testing, 256x256 pixel images of 30 faces were rendered on a black background. For each face, a sequence of 21 poses from -90 degrees (left profile view) to +90 degrees (right profile view) in 6 degree steps was generated (Fig.9) under frontal lighting. In addition each pose was rendered under illumination of a point light source with light incident from -80, -60, -30, +30, +60, +80 degrees with both azimuth and elevation angle, yielding a total of 24 lighting variations (Fig.10). In all images, ambient light of strength 20% was added to the scene in order to provide a small amount of illumination for the most extreme poses.

3 Description of the Framework

In the proposed framework, an image is represented by a number of feature-points at several scales together with their local pixel neighborhoods. The extraction of feature-points is done using corners as features since it was shown that while for example edges are very sensitive to illumination changes [4], corners were found to be stable in many viewing conditions in a recent study [6].

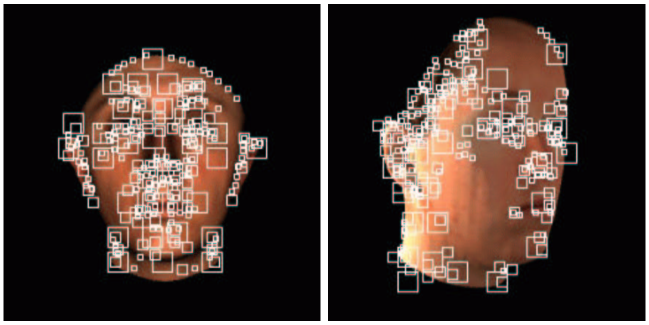


Figure 2. Feature extraction in two face images.

In addition, we use a large number of corners in the representation, ensuring that also finer details of the facial texture (such as freckles, wrinkles, etc.) are included. Finally, the pixel intensities in a small quadratic region $\mathcal{P}$ around each corner are added to the representation to capture the local appearance around each feature-point in the face (see Fig. 1).

3.1 Corner extraction

Extraction of corners is done by a standard algorithm [9] modified to integrate information about all three color channels. Corners are found by evaluating the structure in a small 9x9 pixel neighborhood $\mathcal{N}$ of each pixel by building the following matrix $\mathbf{H}$:

$$\mathbf{H} = \begin{pmatrix} \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial x} \right\rangle & \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial y} \right\rangle \\ \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial y} \right\rangle & \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial y}, \frac{\partial \vec{I}}{\partial y} \right\rangle \end{pmatrix}$$

with $<,>$ as dot-product and $\vec{I}$ as the vector of RGB-values such that an element of $\mathbf{H}$ is e.g. $H(1,2) = \sum \frac{\partial I_r}{\partial x}\frac{\partial I_r}{\partial y} + \frac{\partial I_g}{\partial x}\frac{\partial I_g}{\partial y} + \frac{\partial I_b}{\partial x}\frac{\partial I_b}{\partial y}$. The smaller of the two eigenvalues λ_2 of $\mathbf{H}$ then yields information about the structure of the neighborhood.

The extension to color processing in RGB space yielded further robustness with respect to repeatability of corners. Examples of feature extraction are shown in Fig.2, which shows the features as squares whose sizes correspond to the three scale levels used. An image yields typically around 200 features each containing 25 pixels over all scales, which corresponds to a size reduction of 97.5% with respect to the original image size.

[1] Examples from this database can be downloaded from `http://faces.kyb.tuebingen.mpg.de`

3.2 The matching algorithm

In order to match an input image against the stored representations, i.e. to find corresponding features in two images, we use an algorithm proposed by Scott et al. [8] and further developed by Pilu [5]. The algorithm constructs a similarity mapping for two feature sets, where each feature pair (i,j) is given a weight according to

- the image distance (in pixels) between the features
- the image similarity of the features.

Image similarity is calculated via the Normalized Cross Correlation (NCC) in the local image regions $\mathcal{P}_i, \mathcal{P}_j$ around each feature. The NCC is defined as:

$$\text{NCC} = \frac{\sum_{\mathcal{P}_i,\mathcal{P}_j}(I_i - I_{m,i}) \cdot (I_j - I_{m,j})}{\sum_{\mathcal{P}_i}(I_i - I_{m,i})^2 \cdot \sum_{\mathcal{P}_j}(I_j - I_{m,j})^2}$$

where $I_{i,j}$ are intensity values, $I_{m,i}, I_{m,j}$ mean intensity values in regions $\mathcal{P}_i, \mathcal{P}_j$ respectively. From this definition it follows that the NCC is invariant under a linear transformation

$$\mathcal{P} \to a \cdot \mathcal{P} + b$$

of image intensities *within* $\mathcal{P}$.

This property thus ensures increased stability under lighting variations across the whole face since globally *non-linear* intensity changes due to lighting variations can be approximated as *linear* within $\mathcal{P}$. Using local features thus makes this framework more powerful than appearance-based approaches working with whole images. In order for the linearity approximation to hold, however, one main assumption has to be satisfied in our framework:

- The position of the corner should not change considerably under lighting changes thereby introducing changes in the internal structure of $\mathcal{P}$.

We will examine this assumption in more detail in Section 4.

The matching between two feature sets is then done by first constructing a similarity matrix $\mathbf{A}$, where each entry $A(i,j)$ is given by

$$A(i,j) = e^{\frac{1}{2\sigma^2_{\text{dist}}}(-\text{dist}(\text{f}_\text{i},\text{f}_\text{j}))} \cdot e^{\frac{1}{2\sigma^2_{\text{NCC}}}(1-\text{NCC}(\text{f}_\text{i},\text{f}_\text{j}))}$$

where f_i is the position of feature i in one image, f_j of feature j in another image and i,j index all feature pairs in the two images. The function $\text{dist}(\text{f}_\text{i},\text{f}_\text{j})$ returns the Euclidean distance between the two features in pixels whereas NCC is defined as above. The parameter σ_{dist} is used to bias the results towards close matches in distance and σ_{NCC} is used to bias towards similar matches.

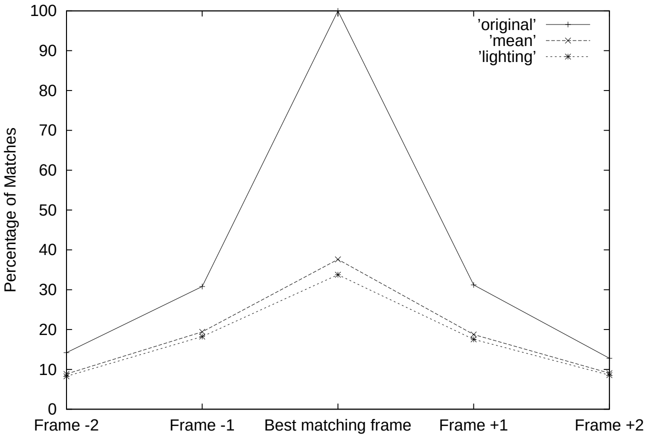


Figure 3. Percentage of matches around best matching frame.

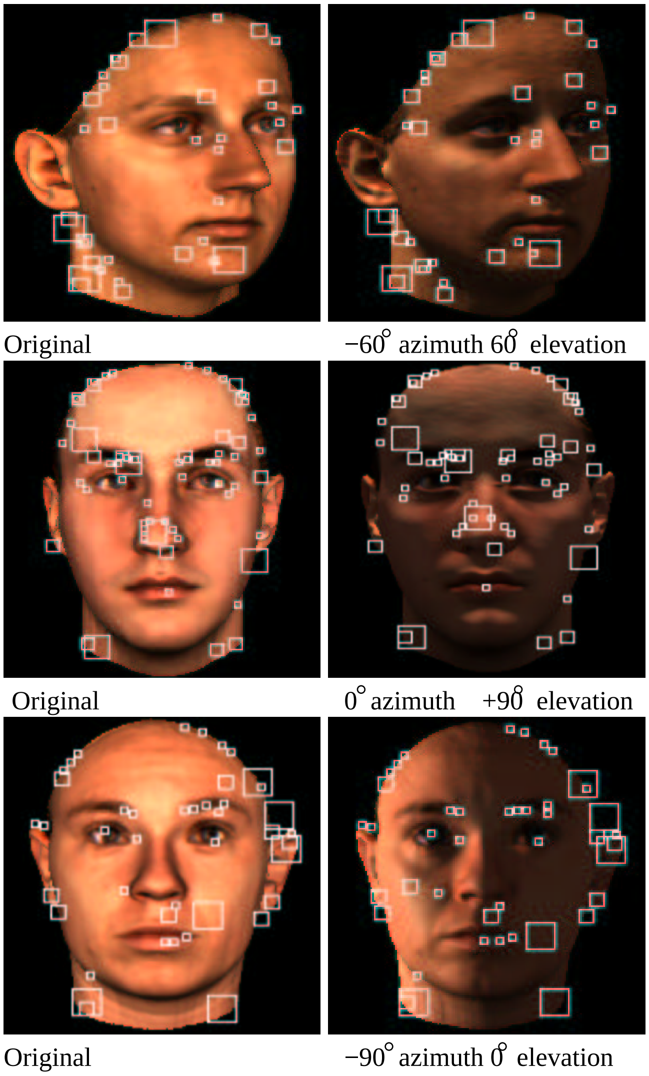
The Singular Value Decomposition of $\mathbf{A}$ then finds a one-to-one feature mapping between the two feature sets which takes into account feature similarity and feature distance. This mapping is refined in a last step by requiring that each match satisfies a minimal value of NCC given by th_{NCC}. The output parameter of the algorithm which is used for finding the best match in our implementation is then given by the *percentage of matches* between feature sets.

Corresponding feature sets are thus constructed by combining configuration (i.e. feature-based information about the layout) with image similarity (i.e. appearance-based information) and finding the optimal solution satisfying both constraints in a least square sense.

4 Recognition Experiments

In all recognition experiments recognition of a given image was done by extraction of the visual features and subsequent matching (Section 3) against all images stored in the database. For each image the matching algorithm yields the percentage of matches with respect to the test image. The image with the highest percentage of matches was then selected as the best match. Note, that due to the structure of the database the best match results in both recognition of individual faces *and* their pose estimation.

In order to have a trade-off between the false negative rate (percentage of matches for which a given face was not recognized by the system) and the false positive rate (percentage of matches for which a given face was more similar to another face) an additional threshold for acceptance of a match th_{acc} was introduced.


Original

−60° azimuth 60° elevation

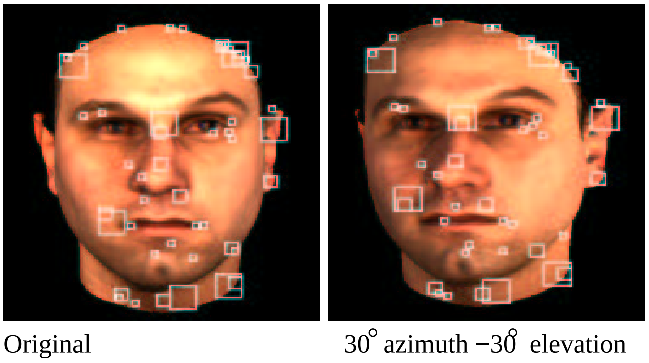
Original

0° azimuth +90° elevation

Original

−90° azimuth 0° elevation

Figure 4. Examples for matched images.


Original

30° azimuth −30° elevation

Figure 5. Examples for matched images across poses.

4.1 Experiment 1 - all poses

The first experiment consisted of recognizing a single pose under all lighting variations in all 30 faces. To restrict the displacement of features, σ_{dist} was set to low values of a few pixels and σ_{NCC} to a value of 0.85.

Using $th_{acc} = 20\%$ yielded a false negative rate for this experiment of **5.4%** and a false positive rate of **14.3%**. Considering the large illumination variations in the database these recognition results demonstrate the robustness of the system. In addition, the percentage of matches around the best matching frame decreased nearly monotonously. Figure 3 shows the percentage of matches around the recognized frame for images containing no lighting variations (the 'original' condition), for images under all lighting variations in the test-database ('lighting' condition) and the mean of the two conditions.

In order to assess the assumption about the stability of corners under lighting variations made in Section 3, some examples for recognized images are shown as close-ups in Fig. 4. The left column shows three original images under frontal lighting, while the right column shows the test images with corresponding features at all scale levels. The positions of the matched features between two images show only slight displacement, which is highest for extreme lighting conditions. While there are a few feature mismatches, the combination of tight matching constraints in both feature layout and feature similarity ensures a high matching fidelity.

Matching failures occurred mostly in conditions with extreme lighting (such as the ±80 ° elevation / ±80 ° azimuth conditions) where not enough matching features could be found. Since the learned representations consisted only of faces under frontal lighting this performance could be further improved by adding a few other lighting conditions to the training set.

4.2 Experiment 2 - reduced set of poses

For the second experiment, the number of poses for each face was reduced to only 7 evenly spaced poses. To allow for greater feature displacement for recognition across a larger viewing angle, σ_{dist} was set to higher values. Here, the false negative rate for $th_{acc} = 14\%$ was **24.3%** and the false positive rate was **32.1%** showing that this low number of views is at the limit of what the matching algorithm can perform in terms of perspective change while still maintaining a reasonable recognition rate (see Fig.5).

The main reason for the higher failure rates was that the number of matches across viewing angles decreased. Reducing th_{acc} to accommodate for this leads then to an increase in the false positive rate. Analysis of recognition rates of all images of poses *not* contained in the learned representation gave only a false positive rate of **14.2%**. This

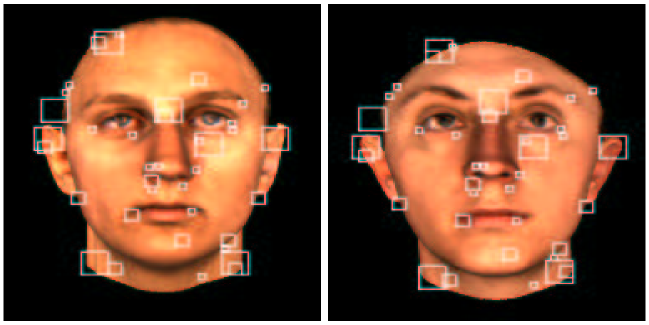


Figure 6. Two different faces and their matching features.

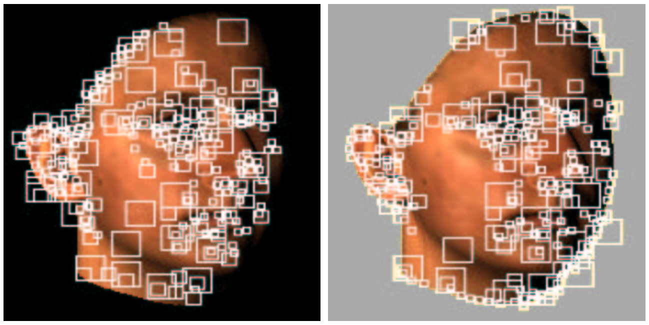


Figure 7. Effect of background color on feature extraction at the outline of the face.

shows that most false positives are due to different faces in the *same poses* as in the learned representations.

4.3 Pose recognition

An analysis of pose estimation performance in the previous experiments revealed that the algorithm recovered the correct pose of the face in **99.8%** of all recognized images. Furthermore, pose estimation was correct in **97.2%** of the rejected matches (false negatives), which fell below th_{acc}.

As already indicated in the previous section, pose recognition was even accurate for matching across exemplars thus leading to a relatively high number of false positives in both experiments. Figure 6 shows two faces in the same pose with their corresponding features. The layout of facial features and the facial texture in both images is highly similar thus leading to the relatively high number of matches (17% of the total number of features from the left face could be matched to the right face).

5 Psychophysical Considerations

First of all, the proposed framework is purely *view-based* in the sense that it does not extract 3D information or incorporates a-priori knowledge coming from detailed 3D models (such as [2]). This agrees well with current models of object recognition [13], where evidence from psychophysics and physiological research suggests a view-based approach.

In [15] we present an extension to this framework originating from research on temporal effects on object recognition [14], where the input to the system not only consists of single images but of sequences. These sequences are used to sequentially build a model from the visual input.

5.1 Combining appearance-based and feature-based approaches

Furthermore, it was shown in psychophysical studies that both appearance-based and configurational information play a role in face recognition [7]. In [7] subjects had to recognize rotated faces, where changes were made either in configurational layout of facial parts (e.g., manipulation of the inter-eye distance) or to the parts themselves (e.g., inserting a different mouth). The results from these experiments show that faces are not processed as a whole but are processed instead as consisting of facial features and their relations.

While our framework already combines appearance-based and feature-based elements on a low abstraction level, we are currently investigating to use the low-level representation as a starting point for derivation of more high-level features such as facial parts. One way to do this is to find smaller sets of features which can be matched across all exemplars in the database. This kind of representation would then allow not only recognition of exemplars but also allow categorization of classes. Categorization is a task which humans seem to perform effortlessly, but for which only recently successful computer vision systems have emerged [12]. The performance of our system with regard to pose estimation across exemplars represents already a step in that direction.

5.2 Illumination-induced apparent shift under lighting

Another psychophysical finding, which can be modeled in our framework, is the illumination-induced apparent shift in face orientation under azimuthal lighting variations [10]. It was found that subjects perceived the orientation of a face lit from one side to be shifted in the opposite direction (up to a maximum shift of 9 degrees). These experiments were done with the same kind of images we used in our test database. In addition, the effect nearly vanished when the background was rendered in white instead of black. The explanation put forward in [10] for this kind of effect was that people judge orientation by comparing visible parts left and right of the profile line in the image.

This effect can be understood within our framework if one considers a face lit from one side which makes the opposite side of the face darker. For such an image, the algorithm extracts features predominantly on the bright side of the face thus giving more weight to this side. When matching this feature set to the learned images in the database, the asymmetry is again encountered in a face which is rotated to the opposite side, thus giving the same tendency as in human observers. When the background is of a light color, however, more features around the now visible outline of the face will be extracted negating the asymmetry (see Fig.7).

In order to assess this effect, only those conditions in the first experiment in which the azimuthal position of the light source changed were evaluated again. For this the percentage of matches for the two neighboring frames of the best matching frame *in the opposite direction of the lighting change* was analyzed. The percentage of matches under azimuthal lighting changes for the frame next to the best matching frame deviated on average by **3.8%** from the baseline condition under frontal lighting. The deviation for the percentage of matches two frames away was **2.5%**, which shows that there is a significant skew towards the opposite lighting direction.

While the aim for the proposed framework was not to fully simulate human performance, it nevertheless uses the same low-level features available to the human visual system and shows performance characteristics similar to humans.

6 Conclusion and Outlook

We presented a view-based recognition system based on locally defined features. By combining both configurational and pixel-based similarity information we obtained good recognition results and reliable pose-estimation under illumination changes. A possible extension to this framework includes the introduction of facial parts which would provide increased recognition performance due to their higher level of abstraction and also allow for categorization. In addition we showed that our system is capable of modeling to some extent psychophysical findings about face recognition.

While recent systems achieved impressive recognition rates by incorporating more high-level knowledge about the image space under illumination [1], we want to emphasize that our low-level system combining elements from appearance-based and feature-based approaches is capable of achieving good performance *without* any a-priori knowledge about illumination.

References

[1] P. Belhumeur, A. Georghiades and D. Kriegman, ”From Few to Many: Illumination Cone Models for Face Recognition Under Variable Lighting and Pose”, *IEEE Trans. PAMI*, (23):6, 643-660, 2001.

[2] V. Blanz and T. Vetter. ”A morphable model for the synthesis of 3D faces”, *Proc. ACM SIGGRAPH 99*, 187-194, 1999.

[3] I. J. Cox, Joumana Ghosn and Peter N. Yianilos, ”Feature-based face recognition using mixture-distance”, *Proc. CVPR* ,209-216,1996.

[4] Y. Moses, ”Face recognition: generalization to novel images”, *PhD thesis, Weizmann Institute of Science*, 1993.

[5] M. Pilu, ”A direct method for stereo correspondence based on singular value decomposition”, *Proc. CVPR* , 261-266, 1997.

[6] C. Schmid, R. Mohr and C. Bauckhage, ”Evaluation of Interest Point Detectors”, *International Journal of Computer Vision*, 37(2), 151-172, 2000.

[7] A. Schwaninger, F. Mast and H. Hecht, ”Mental rotation of facial components and configurations”, *Proc. Psychonomic Society 41st Annual Meeting, New Orleans, USA*, 2000.

[8] G. Scott and H. Longuet-Higgins. ”An algorithm for associating the features of two patterns”, *Proc. Royal Society London, volume B244*, 21-26, 1991.

[9] C. Tomasi and T. Kanade, ”Detection and tracking of point features”, *Carnegie-Mellon Tech Report CMU-CS-91-132*, 1991.

[10] N.F. Troje and U. Siebeck, ”Illumination-induced apparent shift in orientation of human heads”, *Perception 27*, 671-680, 1998.

[11] M. Turk and A. Pentland, ”Eigenfaces for recognition”, *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[12] S. Ullman, Erez Sali and M. Vidal-Naquet, ”A Fragment-Based Approach to Object Representation and Classification”, *Proc. IWVF 2001*, 85-102, 2001.

[13] G. Wallis and H. Bülthoff, ”Learning to recognize objects”, *Trends In Cognitive Sciences 3*, 22-31, 1999.

[14] G. Wallis and H.H. Bülthoff, ”Effects of temporal association on recognition memory”, *Proceedings of the National Academy of Sciences of the United States of America 98*, 4800-4804, 2001.

[15] C. Wallraven and H. Bülthoff, ”Automatic acquisition of image-based representations for recognition from image sequences”, *Proc-. CVPR01*, 2001.

[16] W. Zhao, R. Chellappa, A. Rosenfeld and P.J. Phillips, ”Face Recognition: A Literature Survey”, *TR CVL, University of Maryland, ftp://ftp.cfar.umd.edu/TRs/FaceSurvey.ps.gz*, 2000.

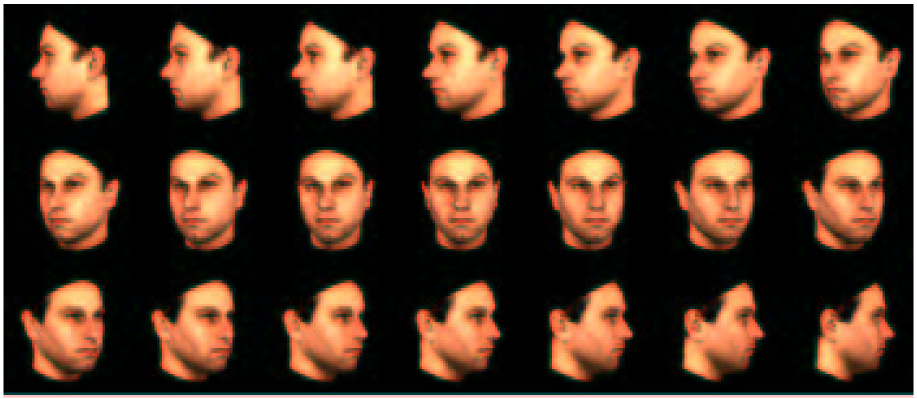


Figure 8. All 21 poses for one face under frontal lighting.

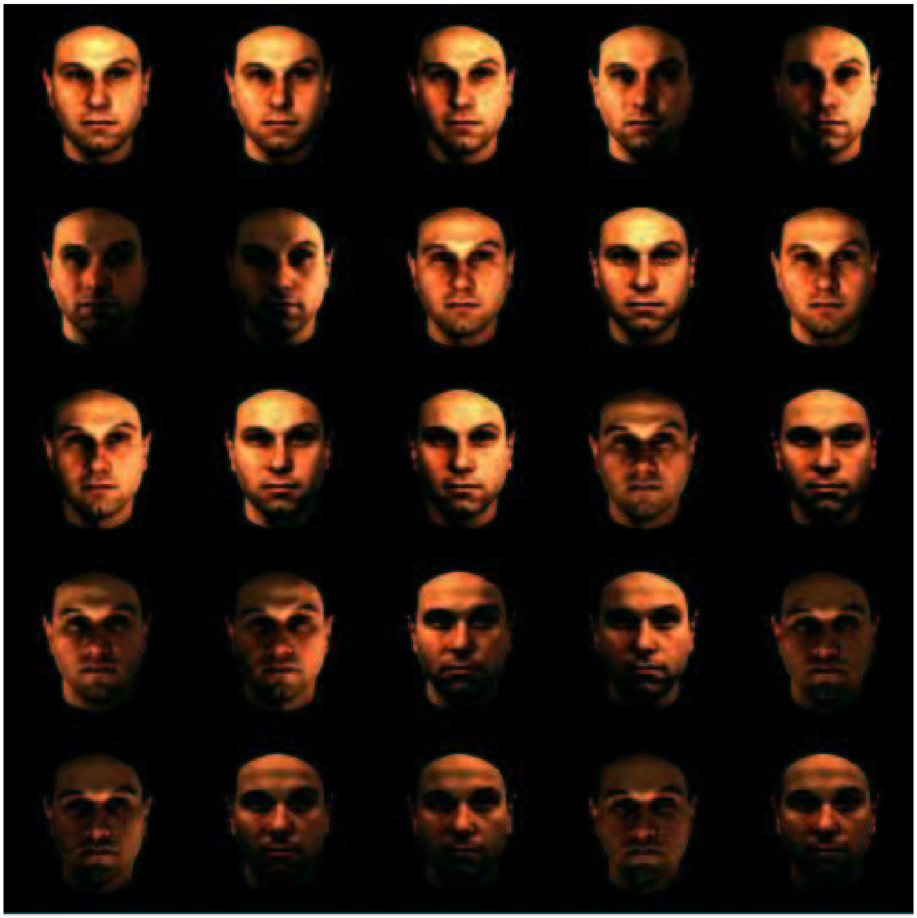


Figure 9. All 25 lighting variations for one face.