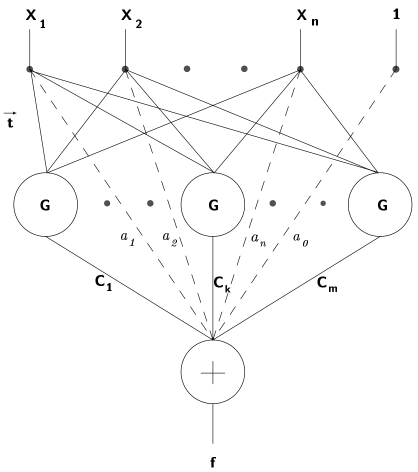

Fig. 20. A network diagram of the Hyper Basis Functions technique

Fig. 21. Images used to determine the dependency of correlation from illumination

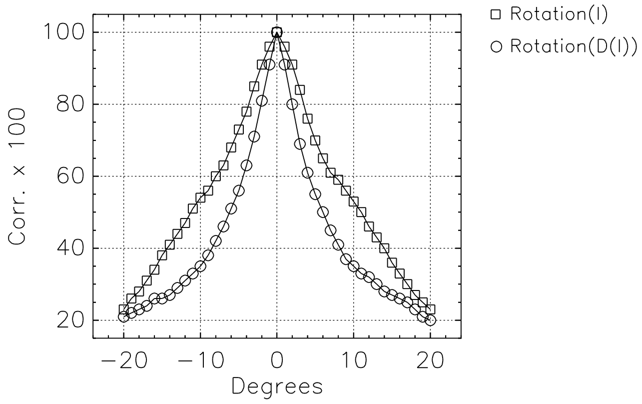

Fig. 23. Correlation value versus image rotation with axis perpendicular to the image plane

Fig. 24. Images used to compute the correlation dependency on rotation (approximately every 5 degrees)

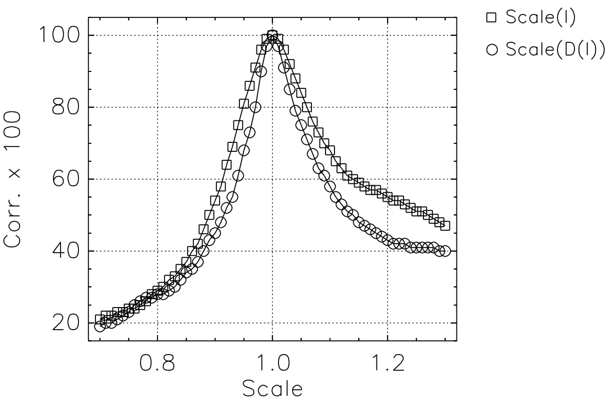

Fig. 22. Correlation value versus image scaling

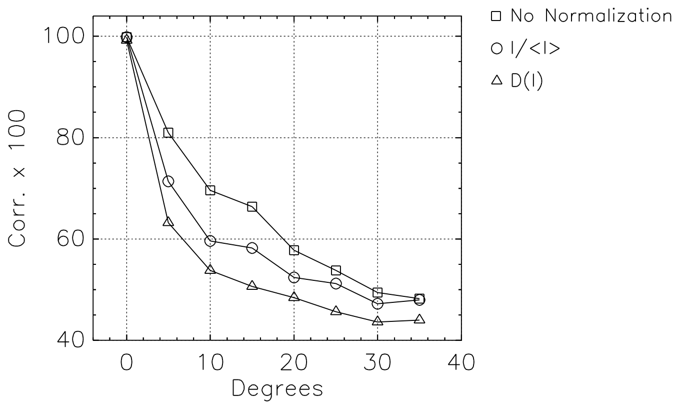

Fig. 25. Correlation versus rotation in the image plane

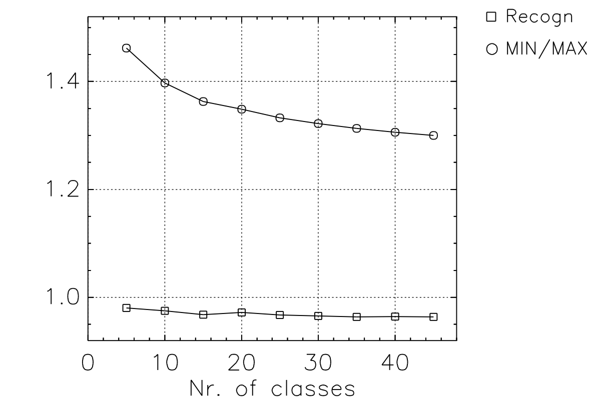

Fig. 16. Recognition performance and MIN/MAX ratio versus the number of classes

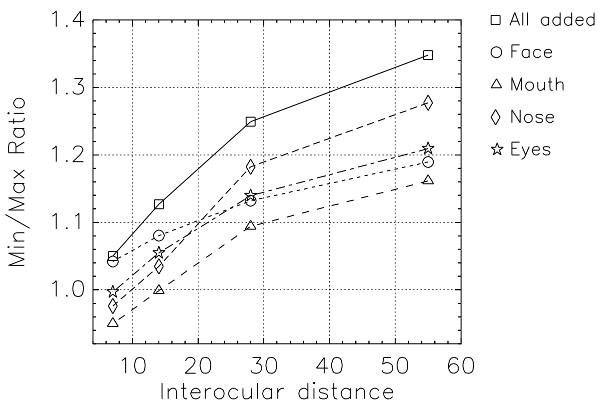

Fig. 18. Average MIN/MAX ratio for each single feature

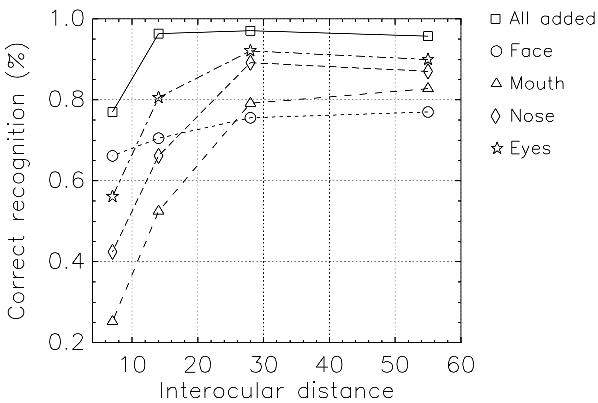

Fig. 17. Average performance for recognition based on each single feature

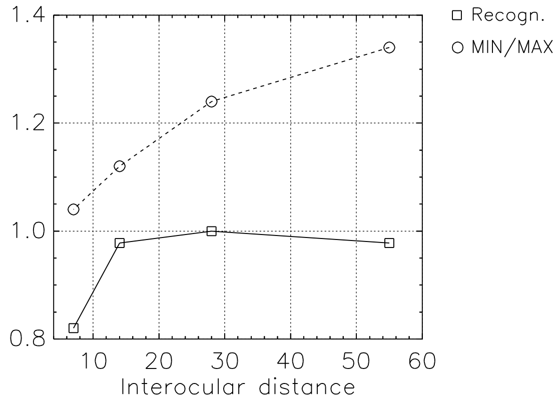

Fig. 19. Recognition performance and MIN/MAX ratio when using two templates per person in the reference database

References

[1] R. J. Baron. Mechanisms of human facial recognition. *International Journal of Man Machine Studies*, 15:137–178, 1981.

[2] Martin Bichsel. *Strategies of Robust Object Recognition for the Identification of Human Faces.* PhD thesis, Eidgenossischen Technischen Hochschule, Zurich, 1991.

[3] W. W. Bledsoe. Man-machine facial recognition. Technical Report Rep. PRI:22, Panoramic Research Inc, Paolo Alto, Cal., 1966.

[4] R. Brunelli. Edge projections for facial feature extraction. Technical Report 9009-12, I.R.S.T, 1990.

[5] R. Brunelli. Face recognition: Dynamic programming for the detection of face outline. Technical Report 9104-06, I.R.S.T, 1991.

[6] R. Brunelli and T. Poggio. Hyperbf networks for real object recognition. In Morgan-Kauffman, editor, *Proc. 12th IJCAI, Sidney*, 1991.

[7] R. Brunelli and T. Poggio. Hyperbf Networks for Gender Classification. In *Proc. DARPA Image Understanding Workshop*, 1992.

[8] J. Buhmann, J. Lange, and C. von der Malsburg. Distortion invariant object recognition by matching hierarchically labeled graphs. In *Proceedings of IJCNN'89*, pages 151–159, 1989.

[9] D. J. Burr. Elastic matching of line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(6):708–713, 1981.

[10] P. J. Burt. Smart sensing within a pyramid vision machine. *Proceedings of the IEEE*, 76(8):1006–1015, 1988.

[11] H. Chan and W. W. Bledsoe. A man-machine facial recognition system: some preliminary results. Technical report, Panoramic Research Inc., Cal, 1965.

[12] G. Cottrell and M. Fleming. Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference*, 1990.

[13] I. Craw, H. Ellis, and J.R. Lishman. Automatic extraction of face features. *Pattern Recognition Letters*, 5:183–187, Feb. 87.

[14] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1972.

[15] A. J. Goldstein, L. D. Harmon, and A. B. Lesk. Identification of human faces. In *Proc. IEEE, Vol. 59*, page 748, 1971.

[16] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems 3*, pages 572–577, 1991.

[17] Zi-Quan Hong. Algebraic feature extraction of image for recognition. *Pattern Recognition*, 24(3):211–219, 1991.

[18] T. Kanade. Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University, Dept. of Information Science, 1973.

[19] Y. Kaya and K. Kobayashi. A basic study on human face recognition. In S. Watanabe, editor, *Frontiers of Pattern Recognition*, page 265. 1972.

[20] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[21] Y. Lee. Handwritten digit recognition using k nearest-neighbor, radial basis functions and backpropagation neural networks. *Neural Computation*, 3(3), 1991.

[22] O. Nakamura, S. Mathur, and T. Minami. Identification of human faces based on isodensity maps. *Pattern Recognition*, 24(3):263–272, 1991.

[23] T. Poggio. *A Theory of How the Brain Might Work*, volume LV of *Cold Spring Harbor Symposia on Quantitative Biology*, pages 899–909. Cold Spring Harbor Laboratory Press, 1990.

[24] T. Poggio. 3d object recognition and prototypes: One 2d view may be sufficient. Technical Report 9107-02, I.R.S.T, 1991.

[25] T. Poggio and S. Edelman. A Network that Learns to Recognize Three-Dimensional objects. *Nature*, 343(6225):1–3, 1990.

[26] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Massachusetts Institute of Technology, 1989.

[27] T. Poggio and F. Girosi. Networks for Approximation and Learning. In *Proc. of the IEEE, Vol. 78*, pages 1481–1497, 1990.

[28] T. Poggio and L. Stringa. A Project for an Intelligent System: Vision and Learning. *Internation Journal of Quantum Chemistry*, 42:727–739, 1992.

[29] J. Sergent. Structural processing of faces. In A.W. Young and H.D. Ellis, editors, *Handbook of Research on Face Processing.* North-Holland, Amsterdam, 1989.

[30] M. Dalla Serra and R. Brunelli. On the Use of the Karhunen-Loeve Expansion for Face Recognition. Technical Report 9206-04, I.R.S.T, 1992.

[31] Luigi Stringa. Automatic Face Recognition using Directional Derivatives. Technical Report 9205-04, I.R.S.T, 1991.

[32] Luigi Stringa. Eyes Detection for Face Recognition. Technical Report 9203-07, I.R.S.T, 1991.

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

[34] H. Voorhees. Finding texture boundaries in images. Technical Report AI-TR 968, M.I.T. Artificial Intelligence Laboratory, 1987.

[35] A. W. Young and H. D. Ellis, editors. *Handbook of Research on Face Processing.* NORTH-HOLLAND, 1989.

[36] Alan L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.

Fig. 13. Different image normalizations (from left to right: I, $I/<I>$, $|\nabla I|_{L_1}$ and $\partial_{xx} I + \partial_{yy} I$)

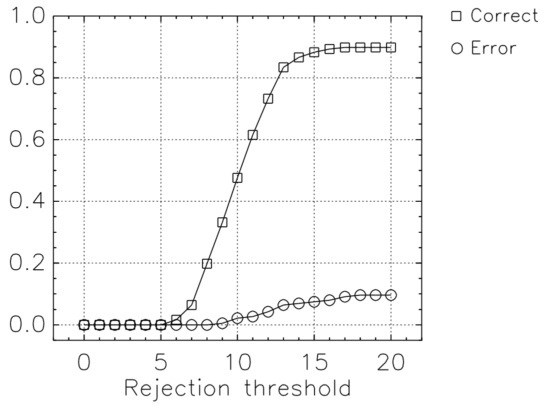

Fig. 11. Analysis of the classifier as a function of the rejection threshold

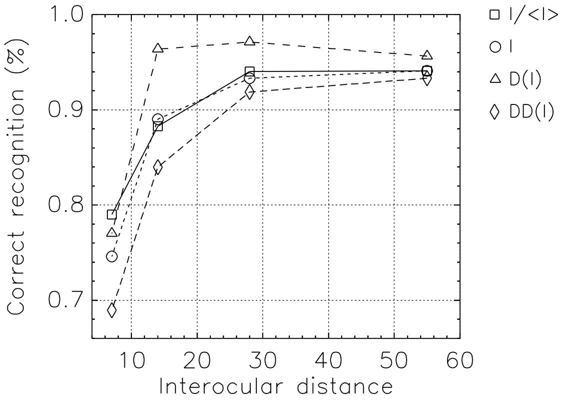

Fig. 14. The recognition performance for different preprocessings as a function of the inter-eyes distance. See text for preprocessing labels

Fig. 12. The different regions used in the template matching strategy

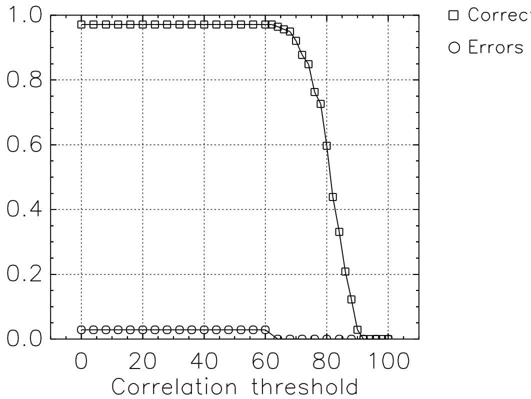

Fig. 15. Recognition performance with error information as a function of the rejection threshold

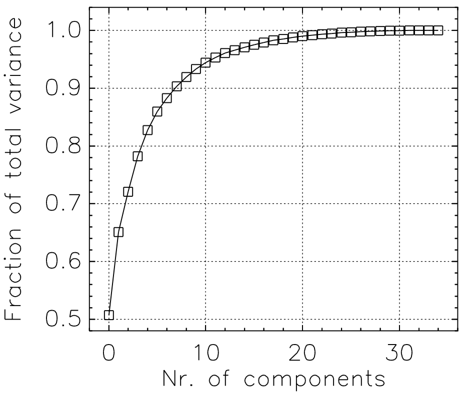

Fig. 7. Feature vectors are expanded in the eigenvectors of the covariance matrix of the available data

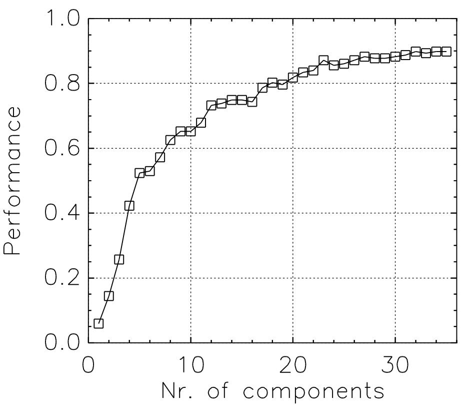

Fig. 8. Performances obtained with an increasing number of principal components

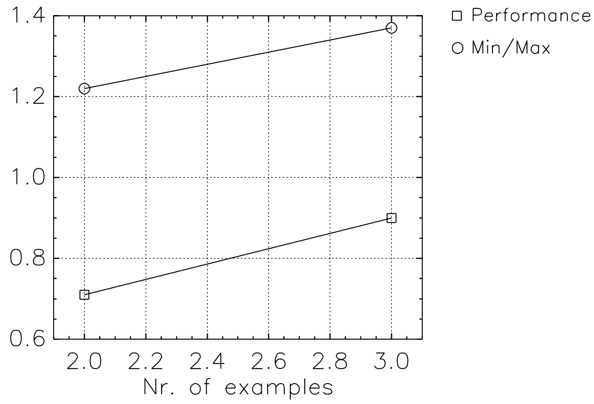

Fig. 9. Use of multiple examples per person improves classification

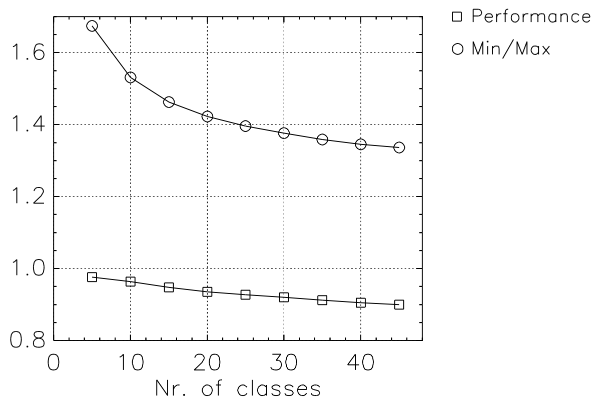

Fig. 10. Performance as a function of the number of classes to be discriminated

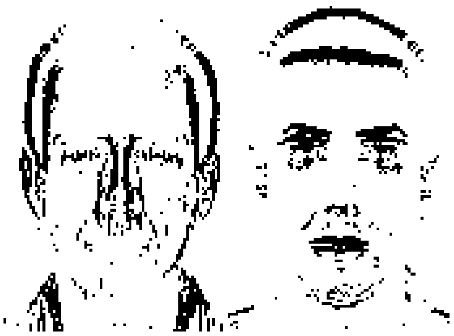

Fig. 1. Horizontal and vertical edge dominance maps

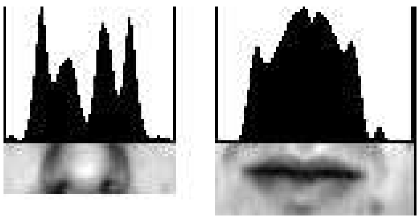

Fig. 2. Typical edge projections data

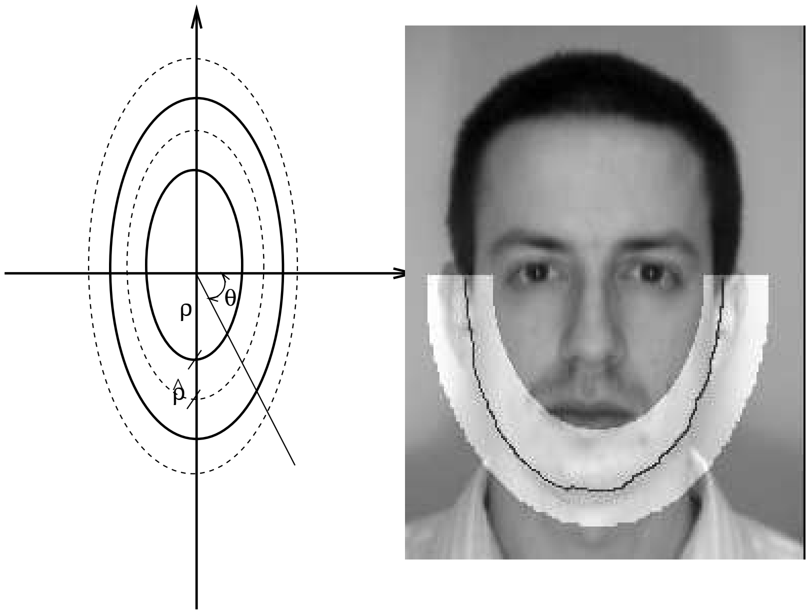

Fig. 3. LEFT: correspondences between the (x,y) coordinate system and the elliptical system; RIGHT: the elliptical annulus delimiting the search.

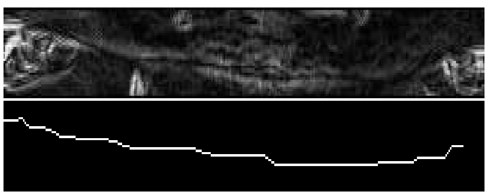

Fig. 4. The edge intensity map (top) and the followed path in the elliptical system (bottom).

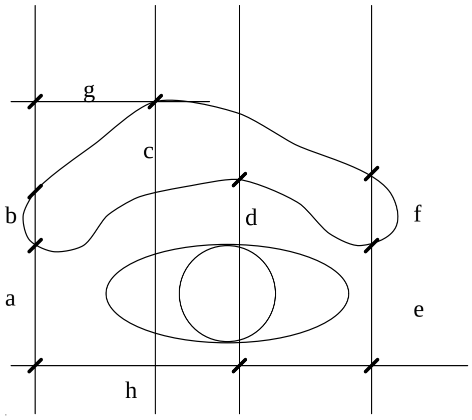

Fig. 5. The parameters used to give a coarse description of the eyebrow arch

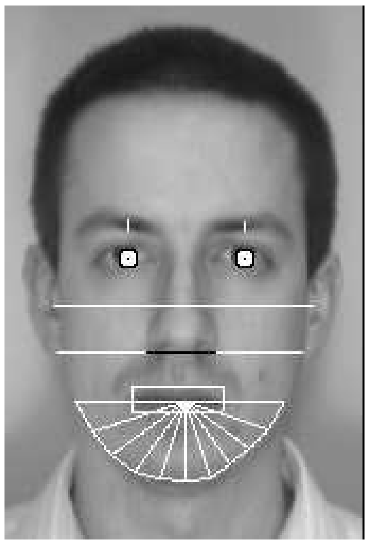

Fig. 6. Geometrical features (white) used in the face recognition experiments

note here that a HyperBF network having as inputs the gradient magnitudes at each pixel and as centers appropriate templates (different centers for different shifts) would be very similar to our template matching scheme (see Fig. 20). It would be somewhat more sophisticated, since it would correspond to a linear classifier on gaussian functions of the correlation coefficients instead of a simple max operation on the correlation coefficients themselves.[8]

In summary, our correlation-based approach seems to offer satisfactory results for recognition from frontal views. The results obtained so far should be verified on larger and more significant data bases. A more difficult problem that we did not consider here is the reliable rejection of images of faces not contained in the data basis. On the other hand, there are possible ways to improve the robustness of our scheme, for instance, by expanding the set of templates. The computational complexity of our scheme is high but not insurmountable, especially if special purpose hardware is developed.

How can the scheme be extended to deal with non-frontal views? If several views of each person are available for different viewpoints, it should be possible to use almost the same scheme, at the expense of a considerably greater computational complexity [9].

It may be the case, however, that only one frontal view per person is available to generate the person's templates. Clearly one single view of a 3D object (if shading is neglected) does not contain sufficient 3D information. If, however, the object belongs to a class of similar objects (prototypes) of which many views are available, it seems possible to make reasonable extrapolations and to guess correctly other views of the specific object from just one 2D view of it. Humans are certainly able to recognize faces turned way 20-30 degrees from frontal from just one frontal view, presumably because they exploit our extensive knowledge of the typical 3D structure of faces.

One of us has recently discussed ([24]) different ways for solving the following problem: *from one 2D view of a 3D object generate other views, exploiting knowledge of views of other, "prototypical" objects of the same class.* We are currently investigating the possibility of using three dimensional information to support recognition from non frontal views. In particular the possibility of using a three dimensional model (explicitly, given a volumetric description of an average face, or implicitly, as derived from a set of two dimensional views, see [24]) to generate the appearance of a face from different viewpoints seems to be promising.

It is worth mentioning that the results obtained (performance as a function of image preprocessing, of compared feature and image resolution) may provide some insights into human mechanisms of facial recognition, especially when considered against the background of available physiological data on neurons in area IT of the monkey cortex, that respond specifically to images of faces (see [23]).

Face identification is by no means exhaustive of face perception related tasks. People are able to discriminate sex, age, and expressions from faces. Gender classification, for example, has been recently investigated using either geometrical features ([7]) or templates ([16], L. Stringa, personal communication, T. Sejnowski). The comparison between the results obtained using geometrical features [7] and templates (Stringa) on almost the same database suggests that also in this task a template based approach has superior performance, but may not be as consistent with properties of human vision in gender classification.

Appendix

A. Correlation dependency on illumination, rotation and scale

The practicality of template matching is a function of its robustness against deformation and noise in the image, relative to the templates. Of course, a suitable set of templates can cope with expected deformations but there are obvious tradeoffs between sensitivity to deformations and number of needed templates. Typical image deformations to be expected in our case are illumination variation, scale variations, and deviations from frontal views. The following table and graphs address the dependency of the correlation value on these deformations.

The dependency of the correlation value on illumination has been measured by the ratio of average correlation, computed on the images of Fig. 21, to the average value as computed on matching faces of the database.

PreProcessing	Corr. Reduction
No Norm.	1.50
$I/<I>$	1.20
$\|\partial_x I\| + \|\partial_y I\|$	1.18
$\partial_{xx} I + \partial_{yy} I$	1.19

The preprocessing based on the computation of gradient magnitude gives the greater MIN/MAX ratio and exhibits the lowest illumination dependence. It is therefore the most invariant against illumination variations.

The dependency on scale and rotation (in the image plane) has been computed by deforming with an affine transformation a single testing image (see Figs. 22,23).

Finally, the dependency of the correlation value on rotations around the central vertical axis lying in the image plage has been determined using a set of images at (approximately) given rotation degrees. Results are shown in Fig. 25.

Acknowledgements

The authors thanks Dr. Luigi Stringa for helpful suggestions and stimulating discussions.

[8] It would also offer the possibility, in principle, to learn optimal centers – i.e. templates.

[9] Use of a HyperBF classifier with the ability of interpolating between views may be particularly advantageous.

The use of an intensity normalization, besides that of the normalized correlation coefficient, proved to be marginally effective (a consistently higher MIN/MAX ratio at the different scales and a minor improvement in the recognition rate at the more resolved scale).

An interesting question is the dependency of recognition on the resolution of the available image. To investigate this dependency, recognition was attempted on a multiresolution representation of the available pictures (a gaussian pyramid of the image suitably preprocessed). The range of resolution was $1 \div 8$ (the gaussian pyramids having 4 levels).

As the performance plots reveal, recognition is stable on a range $1 \div 4$, implying that correlation-based recognition is possible at a good performance level using templates (i.e. the windows we mentioned earlier) as small as 36×36 pixels[6]. A consequence of such a result is that the common objection that recognition through template matching is too expensive computationally does not really apply in this case [7].

How discriminating are the single facial features? A pragmatic answer is to assign the discrimination power based on the recognition performance using each single feature. The experimental analysis shows that the features we used can be sorted by decreasing performance as:

1. eyes
2. nose
3. mouth
4. whole face template

The recognition rate achieved with a single feature (eyes, nose or mouth) is remarkable and consistent with the human ability of recognizing familiar people from a single facial characteristic. The similarity score of the different features can be integrated to obtain a global score. The integration can be done in several ways:

- choose the score of the most similar feature;
- the feature scores are added;
- the feature scores are added, but each feature is given a different weight (the same for all people);
- the features are added using a person dependent weight;
- the features scores are used as inputs to a classifier such as a Nearest Neighbor or a HyperBF network (see [26], [27])

The strategy adopted in the reported experiment is the second one: the features score are simply added. The integration of more features has a beneficial effect on recognition and on the robustness of the classification (see the plots of the R_{mM} value in Fig. 17).

Interestingly, the whole face is the least discriminating template. This is partly due to the difficulty of perfectly normalizing the scale of the pictures. It is also expected, since the whole face is the template most sensitive to slight deformations due to deviations from frontal views.

In this approach both geometrical and holistic feature information is used at the same time. Geometrical information plays a role when the mask stored in the database is used to locate the corresponding zone on the unknown image, while holistic information is taken into account by the pixel-by-pixel comparison of the correlation procedure. Performance can be increased by using templates from more than one image per person. A last experiment, in which we used templates from two images per person in the reference database, has shown perfect recognition (at an intermediate resolution and on this data base) and increased MIN/MAX performance at all resolutions (see Fig. 19), when using as a matching score the maximum of the cumulative scores from the two available database images.

4. Conclusion

We have investigated performance of automatic techniques for face recognition from images of frontal views. Two different approaches have been compared in terms of two simple, new algorithms that we have developed and implemented. The two approaches are:

- identification through a vector of geometrical features;
- identification through a *template matching* strategy.

Our use of template matching is superior in recognition performance on our data base. It is also simpler. The feature-based strategy, however, may allow a higher recognition speed and smaller memory requirements (information can be stored at one byte per feature, requiring only 35 bytes per person in our experiments). The result is clearly specific to our task and to our implementations. Additional features may be used and it may be possible to extract them more precisely. We don't believe, however, that it is reasonable to separate the feature extraction stage, assuming for instance features extracted manually, in the evaluation of the approach: features are only as good as they can be computed.

It must be stressed that the template-matching scheme we have implemented, though simple, is quite different from straight grey-level correlation on the whole face. It uses preprocessing of the image that transforms it into a map of the magnitude of the gradient, quite close therefore to an edge map. A key to its success is how it exploits several different and smaller templates for the eyes, mouth and nose, respectively, in a way somewhat similar to using feature detectors. Most importantly, its first, critical step is the same as in the feature-based approach: detection of the eyes (through correlation matching eye prototypes) and associated automatic normalization of scale and orientation of the image. From this point of view, one may argue that our template matching algorithm contains some elements of feature-based approaches: features (the eyes) are used to normalize and template matching is done separately on a set of distinct features (eyes, nose, mouth). It is indeed possible that successful object recognition architectures need to combine aspects of feature-based approaches with template matching techniques.

Recent work in our laboratory has studied recognition schemes based on K-L decomposition, similar to a system recently described by Turk and Pentland ([33], see also [20]). Since principal components are linear combinations of the templates in the data basis, the technique cannot achieve better results than correlation but it may be capable of achieving a comparable performance with a smaller computational effort. Our results ([30]) indicate a performance of about 96 percent with a fraction of the computational effort needed by our correlation approach (which had a superior performance on the same testing set). Notice that we expect the K-L technique of Dallaserra and Brunelli to have better performance than Turk's, because of their use of several small templates, more stable against image distortions.

In a recent paper [6] we have used the HyperBF network in conjunction with feature vectors to recognize a 3D object from any view: the inputs to the network were the parameters of the features (such as their position in the image). It is interesting to

[6] Preliminary psychophysical recognition experiments have shown remarkable agreement with the scale dependence exhibited by template-matching recognition.

[7] The time needed to compare two images, using eye, nose and mouth templates, at an interocular distance of 27 pixels, is approximately 25msec on a SPARCStation IPX. Comparison of an unknown image to the whole database can take advantage of special-purpose hardware or distributed processing. An efficient strategy for template matching has been recently proposed in [31].

distribution, that is the covariance matrix Σ, by using all the examples in the learning set

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} \Sigma_i \tag{6}$$

where Σ_i is estimate of the covariance matrix obtained from the data of the i-th person. Once the covariance matrix is estimated, the probability of a given measurement can be directly associated to suitably defined distance:

$$\Delta_j(\mathbf{x}) = (\mathbf{x} - \mathbf{m}_j)^T \Sigma^{-1} (\mathbf{x} - \mathbf{m}_j) \tag{7}$$

where $\mathbf{m}_j$ is the average vector representing the j-th person. An unknown vector is then associated to the *nearest neighbor* in the databse, i.e. to the person which maximizes the probability of the measurement vector[4]. The effectiveness of the features in describing the available data has been investigated using the Karhunen-Loeve expansion (or principal component analysis, [14]). The feature vectors are expressed as a linear combination of the eigenvectors of the covariance matrix, sorted by decreasing eigenvalue: the first components in the new expression of the vectors are the most effective in capturing the variance of the data. The fraction of the total variance captured by the first n eigenvectors is reported in Fig. 7. The performance which can be achieved using the first n principal components is reported in Fig. 8.

A useful data on the robustness of the classification is given by an estimate of the intra class variability as opposed to the inter class variability. This can be done using the so called Min/Max ratio (hereafter R_{mM})(see [25], [26]) which is defined as the minimum distance to a wrong correspondence over the maximum distance to the correct correspondence. In our experiments each class was represented by a single element, the arithmetic average of the available examples, so that the maximum distance reduces to the distance from the representing vector. Should the Min/Max ratio be greater than 1 for every class, perfect discrimination could be achieved. The values reported for the different experiments are average values of the ratio computed for the different classes: the higher the value the better discriminable the classes are. It is important to note that the the vectors of geometrical features extracted by our system have low stability, i.e. the intra-class dispersion of the different features is of the same order of the inter-class dispersion (from three to two times smaller). This suggests that performance could be improved with more accurate feature detectors where the process of feature extraction (see also [19], [18] where the use of manually extracted features is studied). It is not clear, however, how to design more accurate feature detectors.

An important issue is how performance scales with the size of tha data base. To obtain these data, a number of recognition experiments have been conducted on randomly chosen subsets of classes at the different required cardinalities (200 random subsets at each cardinality). The plots in Fig. 10 report both recognition performance and the R_{mM} ratio. As expected, both data exhibit a monotonically decreasing trend for increasing cardinality of the class.

A possible way to enhance the robustness of classification is the introduction of a rejection threshold. The classifier can then suspend classification if the input is not sufficiently similar to any of the available models. Rejection could trigger the action of a different classifier or the use of a different recognition strategy (such as voice identification). Rejection can be introduced, in a metric classifier, by means of a rejection threshold: if the distance of a given input vector from all of the stored models exceeds the rejection threshold the vector is *rejected*. A possible figure of merit of a classifier with rejection is given by the recognition performance with no errors (vectors are either correctly recognized or rejected). The average performance of our classifier as a function of the rejection threshold is given in Fig. 11. [5]

3. Template Matching Strategy

The other class of approaches to automated face recognition has at its core a very simple recognition technique, based on the use of whole image grey-level templates. The most direct of the matching procedures is *correlation* and is the basis of the work of Baron (see [1]). The system we implemented is an extension of the little known work of Baron (extensively described in [1]). First, the image is normalized using the same technique described in the previous section. Each person is represented by a database entry whose fields are a digital image of his/her frontal view and a set of four masks representing eyes, nose, mouth, and face (the region from eyebrows downwards), as shown in Fig. 12. The location of the four masks relative to the (normalized) eye position is the same for the whole database.

When attempting recognition, the unclassified image is compared in turn to all of the database images, returning a vector of matching scores (one per feature) computed through normalized cross-correlation. The unknown person is then classified as the one giving the highest cumulative score.

The main difference between our approach and that of Baron lies in the window selection procedure. Whereas Baron's selection was done interactively by a human operator, which could select the windows (s)he considered to be most distinctive, our selection procedure is entirely automatic and the same set of features is selected for all of the available images. It is then possible to automatically add a complete database entry for an unknown person, thereby easing the update of available information. Another, minor difference can be found in the eye location procedure used for normalization: the one we propose uses a single template at different scales (as opposed to the sixteen of Baron's).

As already pointed out, correlation is sensitive to illumination gradients and the question arises whether there is a way to preprocess the compared images to get rid of this confounding effect. Trying to decide experimentally this point, we ran four different experiments of recognition using correlation on images preprocessed in different ways. The different normalization used in the comparison were:

- I : no preprocessing: a plain intensity image was used;
- $I/<I>$: intensity normalization using the ratio of the local value over the average brightness in a suitable neighborhood;
- *D(I)* : gradient magnitude: the intensity of the gradient, computed with an L_1 norm on a Gaussian regularized image was used ($|\partial_x I| + |\partial_y I|$);
- *DD(I)* : the Laplacian of the intensity image ($\partial_{xx} I + \partial_{yy} I$);

The recognition rates we have been able to obtain with the different preprocessing techniques are reported in Fig. 14. The best results, both in term of recognition have been obtained using gradient information.

[4] We could have chosen other classifiers instead of a Bayes Classifier. The HyperBF classifier, used in previous experiments of 3D object recognition [25], [6], allows the automatic choice of the appropriate metric, which is still, however, a weighted Euclidean metric.

[5] Experiments by Lee on a OCR problem ([21]) suggest that a HyperBF classifier would be significantly better than a NN classifier in the presence of rejection thresholds.

Another set of constraints derives from the fact that almost every face has two eyes, one nose, one mouth with a very similar layout. While this may make the task of face classification more difficult, it can ease the task of feature extraction. The following paragraphs briefly explore the implication of bilateral symmetry and expose some ideas on how anthropometric measures can be used to focus the search of a particular facial feature and to validate results obtained through simple image processing techniques ([4], [5]).

A very useful technique for the extraction of facial features is that of integral projections. Let $\mathcal{I}(x,y)$ be our image. The vertical integral projection of $\mathcal{I}(x,y)$ in the $[x_1,x_2] \times [y_1,y_2]$ rectangle is defined as:

$$V(x) = \sum_{y=y_1}^{y_2} \mathcal{I}(x,y) \tag{4}$$

The horizontal integral projection is similarly defined as:

$$H(y) = \sum_{x=x_1}^{x_2} \mathcal{I}(x,y) \tag{5}$$

This technique was successfully used by Takeo Kanade in his pioneering work ([18]) on recognition of human faces. Projections can be extremely effective in determining the position of features provided the window on which they act is suitably located to avoid misleading interferences. In the original work of Kanade the projection analysis was performed on a binary picture obtained by applying a laplacian operator (a discretization of $\partial_{xx} I + \partial_{yy} I$) on the grey-level picture and by thresholding the result at a proper level. The use of a laplacian operator, however, does not provide information on edge (that is gradient) directions. We have chosen therefore to perform edge projection analysis by partitioning the edge map in terms of edge directions. There are two main directions in our constrained face pictures: horizontal and vertical[3].

Horizontal gradients are useful to detect the left and right boundaries of face and nose, while vertical gradients are useful to detect the head top, eyes, nose base and mouth.

Once eyes have been located using template matching, the search for the other features can take advantage of the knowledge of their average layout (an initial estimate, to be updated when new faces are added to the database, can be derived by manually locating features on a single face).

2.2.1 Mouth and nose

Mouth and nose are located using similar strategies. The vertical position is guessed using anthropometric standards. A first, refined estimate of their real position is obtained looking for peaks of the horizontal projection of the vertical gradient for the nose, and for valleys of the horizontal projection of the intensity for the mouth (the line between the lips is the darkest structure in the area, due to its configuration). The peaks (and valleys) are then rated using their prominence and distance from the expected location (height and depth are weighted by a gaussian factor). The ones with the highest rating are taken to be the vertical position of nose and mouth. Having established the vertical position, search is limited to smaller windows.

The nose is delimited horizontally searching for peaks (in the vertical projection of horizontal edge map) whose height is above the average value in the searched window. The nose boundaries are estimated from the leftmost and rightmost peaks. Mouth height is computed using the same technique but applied to the vertical gradient component. The use of directional information is quite effective at this stage, cleaning much of the noise which would otherwise impair the feature extraction process. Mouth width is finally computed thresholding the vertical projection of the horizontal edge map at the average value (see Fig. 2).

2.2.2 Eyebrows

Eyebrow position and thickness can be found through a similar analysis. The search is once again limited to a focused window, just above the eyes, and the eyebrows are found using the vertical gradient map. Our eyebrows detector looks for pairs of peaks of gradient intensity with opposite direction. Pairs from one eye are compared to those of the other one: the most similar pair (in term of the distance from the eye center and thickness) is selected as the correct one. Given this information the upper and lower boundary of the left eyebrow are followed and the set of features shown in Fig. 5 is computed. No hairline information is considered because it may change considerably in time.

2.2.3 Face outline

We used a different approach for the detection of the face outline. Again we have attempted to exploit the natural constraints of faces. As the face outline is essentially elliptical, dynamic programming has been used to follow the outline on a gradient intensity map of an elliptical projection of the face image. The reason for using an elliptical coordinate system is that a typical face outline is approximately represented by a line. The computation of the cost function to be minimized (deviation from the assumed shape, an ellipse represented as a line) is simplified, resulting in a serial dynamic problem which can be efficiently solved (see [5]).

In summary, the resulting 35 geometrical features that are extracted automatically in our system and that are used for recognition, are as follows:

- eyebrow thickness and vertical position at the eye center position;
- a coarse description (11 data) of the left eyebrows arches;
- nose vertical position and width;
- mouth vertical position, width (upper and lower lips) and height;
- eleven radii describing the chin shape;
- bigonial breadth (face width at nose position);
- zygomatic breadth (face width halfway between nose tip and eyes).

A pictorial presentation of the features is given in Fig. 6.

2.3 Recognition Performance

Detection of the features listed above associates to each face a thirtyfive-dimensional numerical vector. Recognition is then performed with a Bayes classifier.

Our main experiment aims to characterize the performance of the feature-based technique as a function of the number of classes to be discriminated. Other experiments try to assess performance when the possibility of rejection is introduced.

We assume that the feature vectors for a single person are distributed according to a gaussian distribution. Different people are characterized only by the average value while the distribution is common. This allows us to estimate the shape of the

[3] A pixel is considered to be in the vertical edge map if the magnitude of the vertical component of the gradient at that pixel is greater than the horizontal one. The gradient is computed using a gaussian regularization of the image. Only points where the gradient intensity is above an automatically selected threshold are considered ([34], [4]).

fixed only approximately, so that scale variations of as much as 30 percent were possible.

In all of the recognition experiments the learning set had an empty intersection with the testing set.

2. Geometric, feature-based matching

As we have mentioned already, the very fact that face recognition is possible even at coarse resolution, when the single facial features are hardly resolved in detail, implies that the overall geometrical configuration of the face features is sufficient for discrimination. The overall configuration can be described by a vector of numerical data representing the position and size of the main facial features: eyes and eyebrows, nose and mouth. This information can be supplemented by the shape of the face outline. As put forward by Kaya and Kobayashi (see [19]) the set of features should satisfy the following requisites:

- estimation must be as easy as possible;
- dependency on light conditions must be as small as possible;
- dependency on small changes of face expression must be small;
- information content must be as high as possible.

The first three requirements are satisfied by the set of features we have adopted, while their information content is characterized by the experiments described later.

One of the first attempts at automatic recognition of faces by using a vector of geometrical features was due to Kanade (see [18]) in 1973. Using a robust feature detector (built from simple modules used within a backtracking strategy) a set of 16 features was computed. Analysis of the inter and intra class variances revealed some of the parameters to be ineffective, yielding a vector of reduced dimensionality (13). Kanade's system achieved a peak performance of 0.75 correct identification on a database of 20 different people using two images per person, one for reference and one for testing.

The computer procedures we implemented are loosely based on Kanade's work and will be detailed in the next section. The database used is however more meaningful (in the sense of being larger) both in the number of classes (47 different people) to be recognized, and in the number of instances of the same person to be recognized (4).

2.1 Normalization

One of the most critical issues in using a vector of geometrical features is that of proper scale normalization. The extracted features must be somehow normalized in order to be independent of position, scale and rotation of the face in the image plane. Translation dependency can be eliminated once the origin of coordinates is set to a point which can be detected with good accuracy in each image. The approach we have followed achieves scale and rotation invariance by setting the interocular distance and the direction of the eye-to-eye axis. We will describe the steps of the normalization procedure in some detail since they are themselves of some interest (an alternative more recent strategy which is even faster and has comparable performance can be found in [32]).

The first step in our technique resembles that of Baron ([1]) and is based on template matching by means of a normalized cross-correlation coefficient, defined by :

$$C_N(\mathbf{y}) = \frac{< I_T T > - < I_T >< T >}{\sigma(I_T)\sigma(T)} \tag{1}$$

where I_T is the patch of image I which must be matched to T, $<>$ the average operator, $I_T T$ represents the pixel-by-pixel product, and σ the standard deviation over the area being matched. This normalization rescales the template and image energy distribution so that their average and variances match.

The eyes of one of the authors (without eyebrows) were used as a template to locate eyes on the image to be normalized. To cope with scale variations, a set of 5 eye templates was used, obtained by scaling the original one (the set of scales used is 0.7, 0.85, 1, 1.15, 1.3 to account for the expected scale variation). Eye position was then determined looking for the maximum absolute value of the normalized correlation values (one for each of the templates). To make correlation more robust against illumination gradients, each image was prenormalized by dividing each pixel by the average intensity over a suitably large neighborhood.

It is well known that correlation is computationally expensive. Additionally, eyes of different people can be markedly different.

These difficulties[2] can be significantly reduced by using hierarchical correlation (as proposed by Burt in [10]). Gaussian pyramids of the prenormalized image and templates are built. Correlation is performed starting from the lowest resolution level, progressively reducing the area of computation from level to level by keeping only a progressively smaller area. Let α be in $(0, 1)$ and $i = 1, \ldots, n$ be the pyramid level, with L_1 the lowest resolution level. Let $U_i(L_j)$ be the operator that produces an image with the resolution of level i from an image at level j (by pixel replication if $i > j$ and by matched low-pass filtering and subsampling if $i < j$). At each level (starting from level 2) correlation is computed at pixel $\mathbf{x}$ only if the following requirement is satisfied:

$$U_{i+1}(C_{Ni}(\mathbf{x})) \geq \Theta = \max_{\theta}(\theta \mid 1 - \sigma_i(\theta) \geq \alpha) \tag{3}$$

where $\sigma_i(\theta)$ is the cumulative (frequency) distribution of the (computed) correlation values at level i and α is the fraction of *active* pixels (i.e. the sites where correlation was computed) which will be projected at the upper level.

Once the eyes have been detected, scale is pre-adjusted using the ratio of the scale of the best responding template to the reference template. The position of the left and right eye is then refined using the same technique (with a left and a right eye template). The resulting normalization proved to be good. The procedure is also able to absorb a limited rotation in the image plane (up to 15 degrees). Once the eyes have been independently located, rotation can be fixed by imposing the direction of the eye-to-eye axis, which we assumed to be horizontal in the natural reference frame.

2.2 Feature Extraction

Face recognition, while difficult, presents a set of interesting constraints which can be exploited in the recovery of facial features. The first important constraint is bilateral symmetry.

[2] Influence of eye shape can be further reduced by introducing a modified correlation coefficient. Let $\Omega_I(\mathbf{x})$ be a (small) neighborhood of point $\mathbf{x}$ in image I and $F_{\Omega_I(\mathbf{x})}(w)$ the intensity value in $\Omega_I(\mathbf{x})$ whose absolute difference from w is minimum: if two values qualify, their average (w) is returned. The new cross correlation is considered:

$$C'(\mathbf{y}) = \sum_{\mathbf{x}} F_{\Omega_I(\mathbf{x}+\mathbf{y})}(T(\mathbf{x}))\, T(\mathbf{x}) \tag{2}$$

whose normalized form is similar to Eq. 1. The newly introduced coefficient introduces the possibility of *local deformation* in the computation of similarity. The interplay of the two techniques (hierarchical correlation and modified correlation coefficient) proved very effective, yielding no errors on the available database.

Face Recognition: Features versus Templates

R. Brunelli[1], **T. Poggio**[2,1]
[1] Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, ITALY
[2]Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, USA

Abstract—**Over the last twenty years several different techniques have been proposed for computer recognition of human faces. The purpose of this paper is to compare two simple but general strategies on a common database (frontal images of faces of 47 people, 26 males and 21 females, four images per person). We have developed and implemented two new algorithms, the first one based on the computation of a set of geometrical features, such as nose width and length, mouth position and chin shape, and the second one based on almost-grey-level template matching. The results obtained on the testing sets, about 90% correct recognition using geometrical features and perfect recognition using template matching, favour our implementation of the template matching approach.**

One of the most remarkable abilities of human vision is that of face recognition. It develops over several years of childhood, is important for several aspects of our social life and, together with related abilities, such as estimating the expression of people we interact with, has played an important role in the course of evolution.

The problem of face recognition was considered in the early stages of computer vision and is now undergoing a revival, after nearly twenty years. Different specific techniques were proposed or reproposed recently. Among those, one may cite neural nets ([12]), elastic template matching ([8], [36]), Karhunen-Loeve expansion ([33]), algebraic moments ([17]) and isodensity lines ([22]). Typically, the relation of these techniques with standard approaches and their relative performance has not been characterized well or at all. Even absolute performance has been rarely measured with statistical significance on meaningful databases.

This paper focuses on two traditional classes of techniques applied to the recognition of digital images of frontal views of faces under roughly constant illumination. The first technique is based on the computation of a set of geometrical *features* from the picture of a face. This was the first approach towards an automated recognition of faces (for the pioneering work of Kanade, see [18]). The second class of techniques is based on *template matching*. We attempt here a comparative analysis of these two different approaches to face recognition.

Psychological studies of human face recognition suggest that virtually every type of available information is used ([35]). Broadly speaking we can distinguish two ways (see [29]) to get a one-to-one correspondence between the stimulus (face to be recognized) and the stored representation (face in the database):

Geometric, feature-based matching. A face can be recognized even when the details of the individual features (such as eyes, nose and mouth) are no longer resolved. The remaining information is, in a sense, purely geometrical and represents what is left at a very coarse resolution. The idea is to extract relative position and other parameters of distinctive features such as eyes, mouth, nose and chin. Goldstein ([15]) and Kaya ([19]) (see also [11], [3]) showed that a computer program provided with face features extracted manually could perform recognition with apparently satisfactory performance. A recent investigation can be found in [13], [2].

Template matching. In the simplest version of template matching, the image, represented as a bidimensional array of intensity values, is compared using a suitable metric (typically the euclidean distance) with a single template, representing the whole face. There are of course several, more sophisticated ways of performing template matching. For instance, the array of grey levels may be suitably preprocessed before matching (see also [2]). Several full templates per each face may be used to account for the recognition from different viewpoints. Still another important variation is to use, even for a single viewpoint, multiple templates. A face is stored then as a set of distinct(ive) smaller templates [1].

A rather different, and more complex, approach is to use a single template together with a qualitative, prior model of how a generic face transforms under a change of viewpoint. The deformation model is then heuristically built into the metric used by the matching measure: this is the idea underlying the technique of elastic templates (see [9], [8], [36]).

In order to investigate the two approaches described above we have developed two new algorithms and tested them on the same data bases. The experiments described here begin to answer questions such as:

- How discriminating are the single features?
- How does recognition performance depend on resolution?
- Is low-pass information less or more effective than high-pass information?
- How do the two different strategies compare with each other?

While we do not claim that our findings are relevant to how human recognition proceeds, they could well provide some hint as to how it could.

1. Experimental setup

The database we used for the comparison of the different strategies is composed of 188 images, four for each of 47 people. Of the four pictures available, the first two were taken in the same session (a time interval of a few minutes) while the other pictures were taken at intervals of some weeks (2 to 4). The pictures were acquired with a CCD camera at a resolution of 512×512 pixels as frontal views. The subjects were asked to look into the camera but no particular efforts were made to ensure perfectly frontal images [1]. The illumination was partially controlled: the same powerful light was used but the environment where the pictures were acquired was exposed to sun light through windows. The pictures were taken randomly during the day time. The distance of the subject from the camera was

[1]Visual inspection of the database revealed no significant deviation from a frontal view but no quantitative analysis was done