M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 245
Appeared in: *IEEE Conference on Computer Vision & Pattern Recognition, 1994*

View-Based and Modular Eigenspaces for Face Recognition

Alex Pentland, Baback Moghaddam, Thad Starner
Vision and Modeling Group, The Media Laboratory
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139

Abstract

In this work we describe experiments with eigenfaces *for recognition and interactive search in a large-scale face database. Accurate visual recognition is demonstrated using a database of* $O(10^3)$ *faces. The problem of recognition under general viewing orientation is also examined. A* view-based *multiple-observer eigenspace technique is proposed for use in face recognition under variable pose. In addition, a modular eigenspace description technique is used which incorporates salient features such as the eyes, nose and mouth, in an* eigenfeature *layer. This modular representation yields higher recognition rates as well as a more robust framework for face recognition. An automatic feature extraction technique using feature* eigentemplates *is also demonstrated.*

1 Introduction

In recent years considerable progress has been made on the problems of face detection and recognition, especially in the processing of "mug shots," i.e., head-on face pictures with controlled illumination and scale. The best results have been obtained for 2-D, view-based techniques based on either template matching (*e.g.,* [1], [2]), or matching using "eigenfaces," *i.e.* template matching using the Karhunen-Loeve transformation of a set of face pictures (*e.g.,* [10, 11, 5]).

However to date tests of these methods have been confined to datasets of only a few hundred images. For real-world applications, we must be able to reliably discriminate among thousands of individuals. Moreover, the problem of recognizing a human face from a *general* view remains largely unsolved, because transformations such as position, orientation, scale, and illumination cause the face's appearance to vary substantially. It is therefore important to ask if we can extend these successful 2-D, view-based recognition approaches to large databases with more general viewing conditions.

In this paper we first explore how the eigenface technique of Turk and Pentland [11] scales when applied to much larger recognition problems. We then generalize the approach to view-based and modular eigenspaces for detection and recognition. The view-based formulation allows for recognition under varying head orientations and the modular description allows for the incorporation of important facial features such as eyes, nose and mouth. These extensions account for variations in object pose and lead to a more robust recognition system.

Although the application reported in this paper is that of face recognition, the same techniques can be applied to recognition and detection of most rigid, roughly convex objects. The general applicability of eigenvector decomposition methods for appearance-based 3D object recognition has recently been convincingly demonstrated by Murase and Nayar [7].

2 A large face database

To date, most face recognition experiments have had at most a few hundred faces. Thus how face recognition performance scales with the number of faces is almost completely unknown. In order to have an estimate of the recognition performance on much larger databases, we have conducted tests on a database of 7,562 images of approximately 3,000 people.

The eigenfaces for this database were approximated using a principal components analysis on a representative sample of 128 faces. Recognition and matching was subsequently performed using the first 20 eigenvectors. In addition, each image was then annotated (by hand) as to sex, race, approximate age, facial expression, and other salient features. Almost every person has at least two images in the database; several people have many images with varying expressions, headwear, facial hair, etc.

2.1 Photobook: an image database tool

This database can be interactively searched using an X-windows browsing tool we have created called Photobook [8]. The user begins by selecting the types of faces they wish to examine; *e.g.,* senior Caucasian males with mustaches, or adult Hispanic females with hats. This subset selection is accomplished using an object-oriented database to search through the face image annotations. Photobook then presents the user with the first 21 of these images (as shown in Figure 1); the rest of the images can be viewed by "paging" through the set of image in groups of 21 images.

At any time the user can select a face from among those presented, and Photobook will then use the eigenvector description of that face to sort the entire set of faces in terms of their similarity to the selected face. Photobook then re-presents the user with the face images, now sorted by similarity to the selected face.

Figure 1 shows the typical results of such a similarity search using the eigenvector descriptors. The face at the upper left of each set of images was selected by the user; the remainder of the faces are the 20 most-similar faces from among the entire 7,562 images. Similarity decreases left

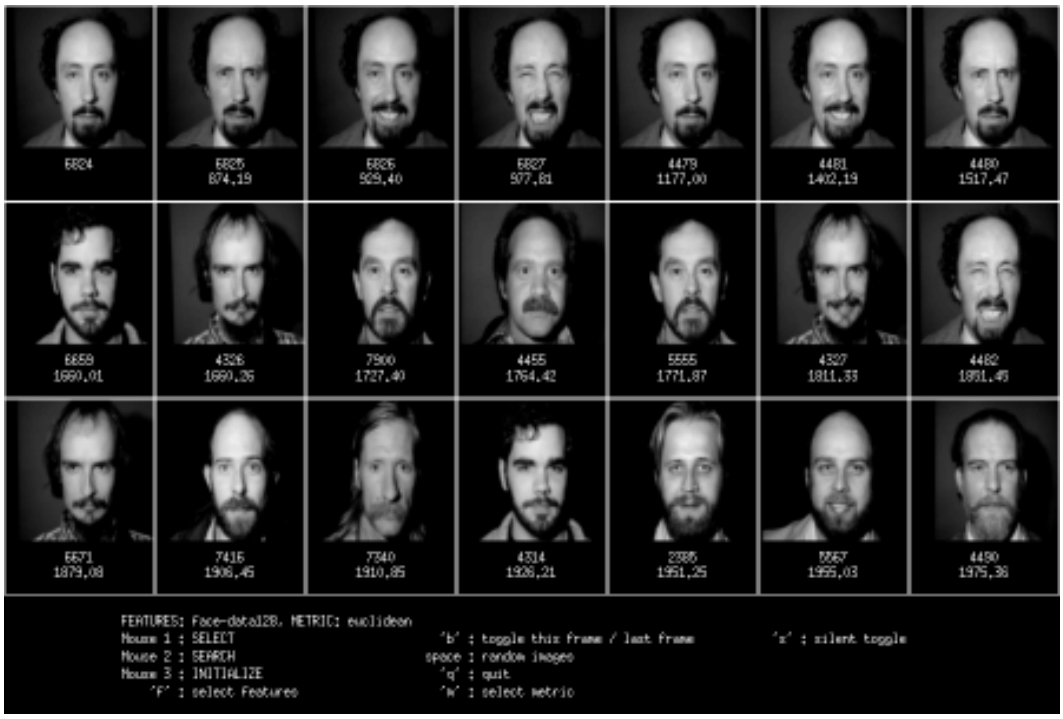


Figure 1: The face at the upper left was selected by the user; the remainder of the faces are the 20 most-similar faces found from among the entire 7,562 individuals in the database. Similarity decreases left to right, top to bottom. Note the ability to match an individual despite wide variations in expression.

to right, top to bottom. The entire searching and sorting operation takes less than one second on a standard Sun Sparcstation, because each face is described using only a very small number of eigenvector coefficients. Of particular importance is the ability to find the same person despite wide variations in expression and variations such as presence of eye glasses, etc.

To assess the average recognition rate, 200 faces were selected at random, and a nearest-neighbor rule was used to find the most-similar face from the entire database. If the most-similar face was of the same person then a correct recognition was scored. In this experiment the eigenvector-based recognition system produced a recognition accuracy of 95%. This performance is somewhat surprising because the database contains wide variations in expression, and has relatively weak control of head position and illumination. This accuracy was maintained across race and sex categories, although we have observed a possible (not statistically significant) trend toward lower performance on Oriental faces.

3 General viewing geometries

There are two ways of approaching the problem of face recognition under general viewing conditions. Given N individuals under M different views, one can do recognition and pose estimation in a universal eigenspace computed from the combination of NM images. In this way a single "parametric eigenspace" will encode both identity as well as viewing conditions. Such an approach, for example, has recently been used by Murase and Nayar [7] for general 3D object recognition.

An alternative formulation is to build a "view-based" set of M separate eigenspaces, each capturing the variation of the N individuals in a common view. The view-based eigenspace is essentially an extension of the eigenface technique to multiple sets of eigenvectors, one for each combination of scale and orientation. One can view this architecture as a set of parallel "observers" each trying to explain the image data with their set of eigenvectors (see also Darrell and Pentland [3].)

In this view-based, multiple-observer approach, the first step is to determine the location and orientation of the target object by selecting the eigenspace which best describes the input image. This is accomplished by calculating the residual description error (the "distance-from-face-space" metric [11]) using each viewspace's eigenvectors. Once the proper viewspace is determined, the image is encoded using the eigenvectors of that viewspace, and then recognized.

3.1 View-based vs. parametric methods

The main advantage of the parametric eigenspace method is its simplicity. The encoding of an input image using n eigenvectors requires only n projections. In the view-based method, M different sets of n projections are required, one for each view. However, this does not imply that a factor of M times more computation is necessarily required. By progressively calculating the eigenvector coefficients while pruning alternative viewspaces, the cost of using M eigenspaces can be greatly reduced.

The key difference between the view-based and para-

Figure 2: Some of the images used to test accuracy at face recognition despite wide variations in head orientation. Average recognition accuracy was 92%, the orientation error had a standard deviation of 15°.

metric representations can be understood by considering the geometry of facespace. In the high-dimensional vector space of an input image, multiple-orientation training images are represented by a set of M distinct regions, each defined by the scatter of N individuals. Multiple views of a face form non-convex (yet connected) regions in image space [1]. Therefore the resulting ensemble is a highly complex and non-separable manifold.

The parametric eigenspace attempts to describe this ensemble with a projection onto a single low-dimensional linear subspace (corresponding to the first n eigenvectors of the NM training images). In contrast, the view-based approach corresponds to M independent subspaces, each describing a particular region of the facespace (corresponding to a particular view of a face). The relevant analogy here is that of modeling a complex distribution by a single cluster model or by the union of several component clusters. Naturally, the latter (view-based) representation can yield a more accurate representation of the underlying geometry.

3.2 Recognition performance

We have evaluated both the view-based and parametric techniques with data similar to that shown in Figure 2. This data consists of 189 images consisting of nine views of 21 people. The nine views of each person were evenly spaced from $-90°$ to $+90°$ along the horizontal plane. Data were provided by Westinghouse Electronic Systems. Our experimental results show a slightly superior performance obtained with the view-based method. Two different testing methodologies were used to judge the relative performance of the parametric and view-based eigenspace methods.

In the first series of experiments the *interpolation* performance was tested by training on a subset of the available views $\{\pm 90°, \pm 45°, 0°\}$ and testing on the intermediate views $\{\pm 68°, \pm 23°\}$. The average recognition rates obtained were 90% for the view-based and 88% for the parametric eigenspace methods.

A second series of experiments tested the *extrapolation* performance by training on a range of views (*e.g.,* $-90°$ to $+45°$) and testing on novel views outside the training range (*e.g.,* $+68°$ and $+90°$). For testing views separated by $\pm 23°$ from the training range, the average recognition rates were 83% for the view-based and 78% for the parametric eigenspace method. For $\pm 45°$ testing views, the average recognition rates were 50% (view-based) and 43% (parametric).

4 Eigenfeatures

The eigenface technique is easily extended to the description and coding of facial features, yielding eigeneyes, eigennoses and eigenmouths. Eye-movement studies indicate that these particular facial features represent important landmarks for fixation, especially in an attentive discrimination task [14]. Therefore we should expect an improvement in recognition performance by incorporating an additional layer of description in terms of facial features. This can be viewed as either a modular or layered representation of a face, where a coarse (low-resolution) description of the whole head is augmented by additional (higher-resolution) details in terms of salient facial features.

This modularity in face description also has distinct advantages for face coding in teleconferencing. For example, a layered representation consisting of the face and eigenmouths has recently been implemented for low bit-rate transmission of visual telephony by Welsh and Shah [13]. In section 5, we will demonstrate the potential utility of eigenfeatures for face recognition.

4.1 Detection of facial features

An important pre-processing step in an eigenvector recognition system is that of registration. A face in an input image must first be located and registered in a standard-size frame before being processed. In addition to head detection and tracking, automatic detection of facial features is also an important component for face recognition. Over the years, various strategies for facial feature detection have been proposed, ranging from the early work of Kanade with edge-map projections [4], to more recent techniques using generalized symmetry operators [9] and multilayer perceptrons [12].

By far, the standard detection paradigm in computer vision is that of simple correlation or template matching. The eigenspace formulation, however, leads to a powerful alternative to simple template matching. The reconstruction error (or residual) of the principal component representation (referred to as the "distance-from-face-space" in the context of our earlier work [11]) is a an effective indicator of a match. The residual error is easily computed using the projection coefficients and signal energy. This detection strategy is equivalent to matching with *eigentemplates* and allows for a greater range of distortions in the input signal (including lighting, rotation and scale). In a statistical signal detection framework, the use of eigen-

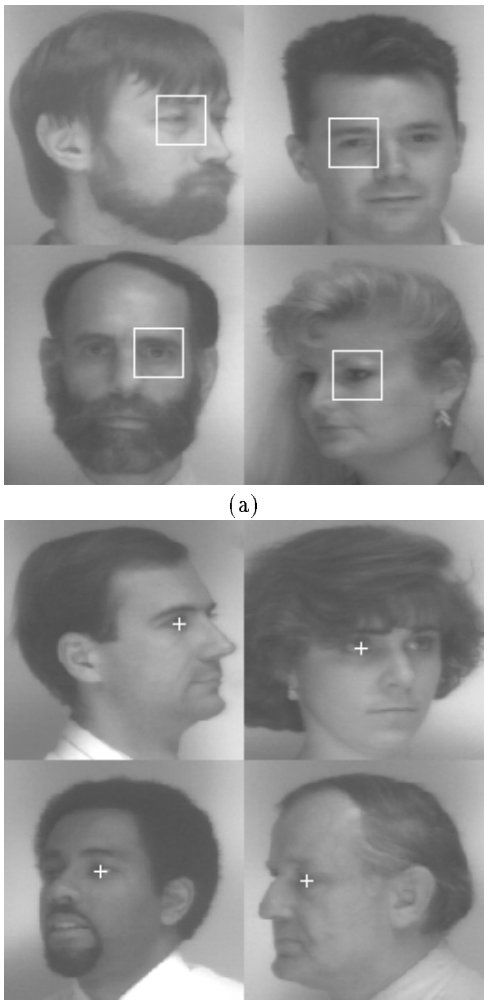


Figure 3: (a) Examples of multiple-view eye training templates and (b) typical detections on novel views.

templates has been shown to yield superior performance in comparison with standard matched filtering [6].

In the eigenfeature representation the equivalent "distance-from-*feature*-space" (DFFS) can be effectively used for the detection of facial features. Given an input image, a feature distance-map is built by computing the DFFS at each pixel. When using n eigenvectors, this requires n convolutions (which can be efficiently computed using an FFT) plus an additional local energy computation. The global minimum of this distance map is then selected as the best feature match.

4.2 View-invariant detection

The DFFS feature detection method can be extended to the detection of features under different viewing geometries. Here, once again, one faces the choice of using either a view-based eigenspace or a parametric eigenspace. Using our multiple-orientation database we tested the relative performance of these two methods for detection in the following manner.

First, a subset of the available views were selected $\{\pm 90°, \pm 45°, 0°\}$ and the appropriate eye templates were extracted for training. These training templates are shown in Figure 3(a). Then the DFFS metric was used to detect the corresponding eye in the intermediate views $\{\pm 68°, \pm 23°\}$. Typical (correct) detections are shown in Figure 3(b). The detections over the four novel views were averaged to yield an overall percentage of correct detection (a correct detection was defined as one within 5 pixels of the true feature location). The percent correct detections were 90% for the view-based detector and 70% for the parametric detector.

The relative performance of the view-based and parametric methods was similar for other facial features (noses and mouths). However, the detection rates for these features were lower due to the greater variations in appearance as a function of viewing geometry (due, for example, to the large depth range of the nose).

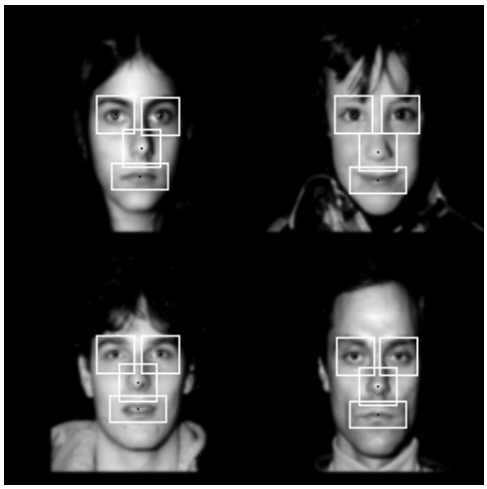
A complicating factor in facial feature detection across head orientation is the issue of feature occlusion and feature/background interaction. The former results in only some features being visible in some views (*e.g.,* right eye only in extreme right views) and the latter in the interaction of some features with the background (*e.g.,* the nose in profile views). However, an estimate of head orientation obtained with the view-based eigenspace method can be used to determine *interior* features that will be visible in the input image and consequently which features are to be relied upon for recognition.

4.3 Detection on a large database

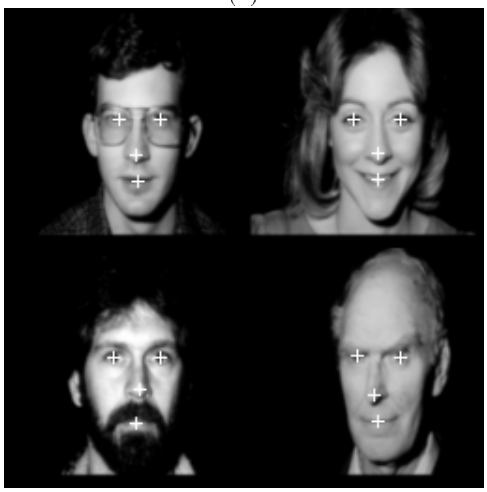
The DFFS feature detector was also used for the automatic detection and coding of the facial features in our large database of 7,562 faces. The same representative sample of 128 individuals used in computing the eigenfaces was used to compute a set of corresponding eigenfeatures. Figure 4(a) shows examples of the training templates used for the facial features (left eye, right eye, nose and mouth). The entire database was processed by using independent detectors for each feature (with the DFFS computed based on projection on the first 10 eigenvectors). The matches were obtained by independently selecting the global minimum in each of the four distance maps. Typical detections are shown in Figure 4(b).

To illustrate the effectiveness of the DFFS detector on this large database, the 7,562 feature detections were pooled into a feature accumulator array as follows: for each detection, the corresponding pixel location in the array was incremented by an amount inversely proportional to the DFFS score at the selected global minimum. Figure 5 shows the combined accumulator array for the four facial features as superimposed on the mean face. The peaks in the accumulator array are quite sharp since false detections are randomly distributed in the image and tend to have large DFFS values. Since the eyes were accurately aligned in the picture taking process, the corresponding eye peaks are quite sharp.

The detection peaks for the nose and the mouth are more diffuse (yet still accurate in location) due to the greater variation in appearance and position. The spatial spread is due to the variations in head shape and the relative positions of the nose and mouth with respect to the eyes. In



(a)



(b)

Figure 4: (a) Examples of facial feature training templates used and (b) the resulting typical detections.

addition, these features tend to have a lower detection rate and higher DFFS values. Although no ground truth data for feature locations is available, eye locations are quite consistent in this database. Using the mean eye location, peak detection rates for the eyes can be conservatively estimated as 94%.

The DFFS metric associated with each detection can be used in conjunction with a threshold — *i.e.,* only the global minima with a DFFS value *less* than the threshold are declared to be a possible match. Consequently we can characterize the detection vs. false-alarm tradeoff by varying this threshold and generating a *receiver operating characteristics* (ROC) curve. Figure 6 shows the ROC curves for the left eye using the first and first 10 eigenvectors in the DFFS detector. A correct detection was defined as a below-threshold global minimum within 5 pixels of the mean left eye position. Similarly, a false alarm was defined as a below-threshold detection located *outside* the 5-pixel radius. Global minima *above* the threshold were

Figure 5: Detection accumulator array

undeclared. The peak performance of the DFFS detector using the first 10 eigenvectors corresponds to a 94% detection rate at a false alarm rate of 6%. Conversely, at a zero false-alarm rate, 52% of the eyes were correctly detected. To calibrate the performance of the DFFS detector, we have also shown the ROC curve corresponding to a standard sum-of-square-differences (SSD) template matching technique. The templates used in this case were the mean features in each case. We observe that for the same probability of detection, the DFFS detector shows an order of magnitude improvement in false-alarm rate over the SSD.

Note that the SSD can be considered a *degenerate* case of a DFFS detector, corresponding to a zero-th order encoding — *i.e.,* using only the mean vector for description. The addition of the principal components results in incremental improvements in detection performance, resulting in a gradation of ROC curves similar to those shown in Figure 6. Naturally, the incorporation of each additional eigenvector means an extra correlation. However, the increase in computational cost is linear with the number of eigenvectors and is typically offset by the subsequent gain in performance. In fact, as the ROC curves indicate, by using only the first eigenvector (at the cost of one additional convolution over SSD) we have substantially increased detection performance.

Finally, we note that the detection of facial features can be made more robust by incorporating constraints on the geometry of a face in terms of relative feature locations. These constraints can be used to guide the search for matches and thus restrict the regions over which the DFFS is computed. Preliminary experiments with such constraints indicate that the detection rate of mouths and noses can be greatly improved by "anchoring" the search with respect to more easily detected features, such as eyes.

5 Modular eigenspaces

With the ability to reliably detect facial features across a wide range of faces, we can automatically generate a modular representation of a face. The utility of this layered representation (eigenface plus eigenfeatures) was tested on a small subset of our face database. We selected a representative sample of 45 individuals with two views per person, corresponding to different facial expressions (neutral vs. smiling). These set of images was partitioned into a training set (neutral) and a testing set (smiling).

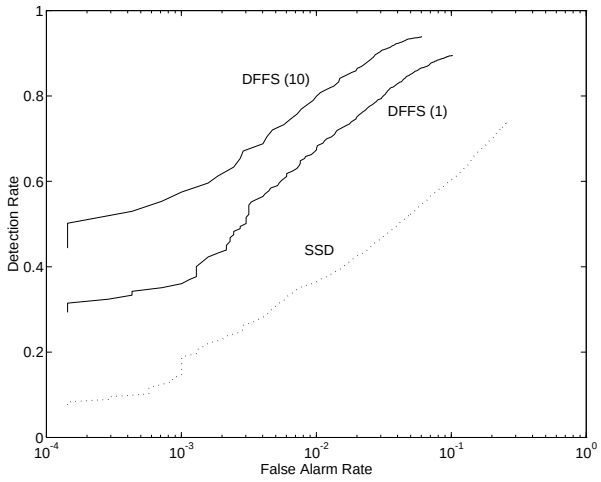


Figure 6: ROC curve for left eye using DFFS detectors with 1 and 10 eigenvectors. An SSD detector is shown for comparison.

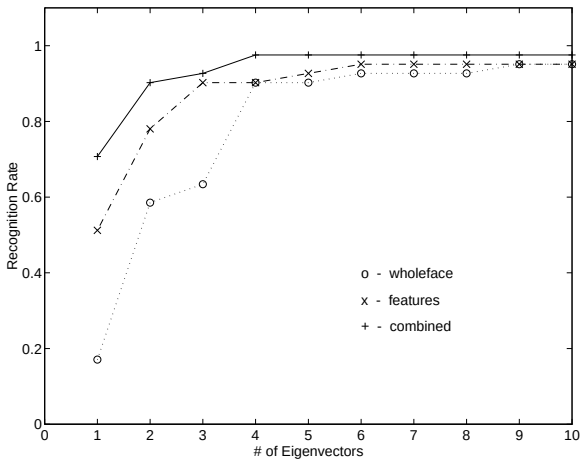


Figure 7: Recognition rates for eigenfaces, eigenfeatures and the combined modular representation.

Since the difference between these particular facial expressions is primarily articulated in the mouth, this feature was discarded for recognition purposes. Figure 7 shows the recognition rates as a function of the number of eigenvectors for eigenface-only, eigenfeature-only and the combined representation. What is surprising is that (for this small dataset at least) the eigenfeatures alone were sufficient in achieving an (asymptotic) recognition rate of 95% (equal to that of the eigenfaces). More surprising, perhaps, is the observation that in the lower dimensions of eigenspace, eigenfeatures outperformed the eigenface recognition. Finally, by using the combined representation, we gain a slight improvement in the asymptotic recognition rate (98%). A similar effect has recently been reported by Brunelli and Poggio [2] where the cumulative normalized correlation scores of templates for the face, eyes, nose and mouth showed improved performance over the face-only templates.

A potential advantage of the eigenfeature layer is the ability to overcome the shortcomings of the standard eigenface method. A pure eigenface recognition system can be fooled by gross variations in the input image (hats, beards, etc.). Figure 8(a) shows additional testing views of 3 individuals in the above dataset of 45. These test images are indicative of the type of variations which can lead to false matches: a hand near the face, a painted face, and a beard. Figure 8(b) shows the nearest matches found based on a standard eigenface classification. Neither of the 3 matches correspond to the correct individual. On the other hand, Figure 8(c) shows the nearest matches based on the eyes and nose, and results in correct identification in each case. This simple example illustrates the potential advantage of a modular representation in disambiguating false eigenface matches.

We are currently exploring strategies for the optimal fusion of the available information in the modular representation. One simple approach is to form a cumulative score in terms of equal contributions by each of the components (head, eyes, nose and mouth). Alternatively, psychophysical data can be used in formulating a more elaborate weighting scheme for classification (*e.g.*, eyes tend to be the most salient features). A more ambitious scheme would be to modulate the contribution of each component in a task or state-dependent manner.

An attractive recognition strategy is to combine a sequential classifier with a coarse-to-fine matching procedure, whereby a pyramid sequence of (low-resolution) eigenface projections is used to limit the database search to a local region of facespace, and finally a (high-resolution) facial feature description is used to perform the final classification. By embedding this mechanism in the framework of our view-based eigenspace method, the overall system can perform robust face recognition under varying head orientations.

6 Conclusions

Our experimental results have demonstrated the success of eigenspace techniques for object search and recognition in a large image database. We believe this is the first time accurate visual recognition has been reported using a database of 3,000 individuals.

We have generalized our technique to handle a variable viewing geometry, using a *view-based* approach by describing faces with a set of 2-D "aspects". The key to the success of such a view-based approach is the ability to localize the object (or features on an object) and identify the correct aspect. In this regard, we have shown that the *distance-from-feature-space* in a view-based eigenspace formulation is an effective tool for robust detection and pose estimation.

Finally, we have extended the approach to a modular representation by incorporating information from different levels of description. Once again, the ability of the DFFS filter to accurately and reliably detect features was critical for successfully incorporating a parts-based description. By using this modular approach we have been able to demonstrate robustness to localized variations in object appearance.

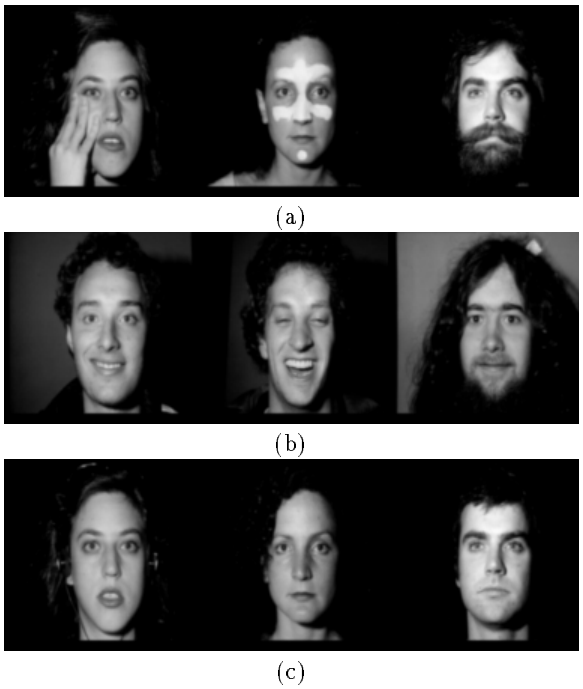


Figure 8: (a) Test views, (b) Eigenface matches, (c) Eigenfeature matches.

Acknowledgments

This research was funded by British Telecom. The authors also wish to thank Olorunfunmi Oliyide and Matthew Turk for their work in assembling the face database.

References

[1] Bichsel, M., and Pentland, A., "Topological Matching for Human Face Recognition," M.I.T. Media Laboratory Vision and Modeling Group Technical Report No. 186, Jan. 1992. *to appear* CVGIP: Image Understanding

[2] Brunelli, R., and Poggio, T., "Face Recognition: Features vs. Templates,", *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, vol. 15, no. 10, Oct. 1993.

[3] Darrell, T., and Pentland, A., "Space-Time Gestures," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, New York, NY, June 1993.

[4] Kanade, T., "Picture processing by computer complex and recognition of human faces," Tech. Report, Kyoto University, Dept. of Information Science, 1973.

[5] Kirby, M., and Sirovich, L., "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 1, Jan. 1990.

[6] Kumar, B., Casasent, D., and Murakami, H., "Principal Component Imagery for Statistical Pattern Recognition Correlators," *Optical Engineering*, vol. 21, no. 1, Jan/Feb 1982.

[7] Murase, H., and Nayar, S. K., "Learning and Recognition of 3D Objects from Appearance" in *IEEE 2nd Qualitative Vision Workshop*, New York, NY, June 1993.

[8] Pentland, A., Picard, R., and Sclaroff, S., "Photobook: Tools for Content-Based Manipulation of Image Databases," SPIE Storage and Retrieval of Image and Video Databases II, No. 2185, San Jose, Feb 6-10, 1994.

[9] Reisfeld, D., Wolfson, H., and Yeshurun, Y., "Detection of Interest Points Using Symmetry,", *ICCV '90*, Osaka, Japan, Dec. 1990.

[10] Turk, M., and Pentland, A., "Face processing: models for recognition," *Intelligent Robots and Computer Vision VIII*, SPIE, Philadelphia, PA, 1989.

[11] Turk, M., and Pentland, A., "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991.

[12] Vincent, J. M., Waite, J. B., and Myers, D. J, "Automatic Location of Visual Features by a System of Multilayered Perceptrons," *IEE Proceedings*, vol. 139, no. 6, Dec. 1992.

[13] Welsh, J. W., and Shah, D., "Facial-Feature Image Coding Using Principal Components,", *Electronic Letters*, vol. 28, no. 22, October, 1992.

[14] Yarbus, A. L., *Eye Movement and Vision*, B. Haigh (trans.), New York, Plenum Press, 1967.