A Wavelet-based Framework for Face Recognition

Christophe Garcia, Giorgos Zikos, Giorgos Tziritas

ICS – Foundation for Research and Technology-Hellas – FORTH
P.O.Box 1385, GR 711 10 Heraklion, Crete, Greece
Tel.: +30 (81) 39 17 01, Fax: +30 (81) 39 16 01
E-mail: {cgarcia,gzikos,tziritas}@csi.forth.gr

Abstract

Content-based indexing methods are of great interest for image and video retrievial in audio-visual archives, such as in the DiVAN project that we are currently developping. Detecting and recognizing human faces automatically in video data provide users with powerful tools for performing queries. In this article, a new scheme for face recognition using a wavelet packet decomposition is presented. Each face is described by a subset of band filtered images containing wavelet coefficients. These coefficients characterize the face texture and a set of simple statistical measures allows us to form compact and meaningful feature vectors. Then, an efficient and reliable probalistic metric derived from the Bhattacharrya distance is used in order to classify the face feature vectors into person classes.

1 Introduction

Face recognition is becoming a very promising tool for automatic multimedia content analysis and for a content-based indexing video retrievial system. Such a system is currently developped within the Esprit project DiVAN ([5]) which aims at building and evaluating a distributed audio-visual archives network providing a community of users with facilities to store video raw material, and access it in a coherent way, on top of high-speed wide area communication networks. The video raw data is first automatically segmented into shots and from the content-related image segments, salient features such as region shape, intensity, color, texture and motion descriptors are extracted and used for indexing and retrieving information.

In order to allow queries at a higher semantic level, some particular pictorial objects have to be detected and exploited for indexing. We focus on human faces detection and recognition, given that such data are of great interest for users queries.

In recent years, considerable progress has been made on the problem of face detection and face recognition, especially under stable conditions such as small variations in lighting, facial expression and pose. A good survey may be found in [16]. These methods can be roughly divided into two different groups: geometrical features matching and template matching. In the first case, some geometrical measures about distinctive facial features such as eyes, mouth, nose and chin are extracted ([2]). In the second case, the face image, represented as a two-dimensional array of intensity values, is compared to a single or several templates representing a whole face. The earliest methods for template matching are correlation-based, thus computationally very expensive and require great amount of storage and since a few years, the Principal Components Analysis (PCA) method also known as Karhunen-Loeve method, is successfully used in order to perform dimensionality reduction ([9, 15, 12, 14, 1]). We may cite other methods using neural network classification ([13, 3]) or using a deformable model of templates ([10, 17]).

In this paper, we propose a new method for face recognition based on a wavelet packet decomposition of the face images. Each face image is described by a subset of band filtered images containing wavelet coefficients. From these wavelet coefficients which characterize the face texture, we form compact and meaningful feature vectors, using simple statistical measures. Then, we show how an efficient and reliable probalistic metric derived from the Bhattacharrya distance can be used in order to classify the face feature vectors into person classes. Experimental results are presented using images from the FERET and the FACES databases. The efficiency of our approach is analyzed by comparing the results with those obtained using the well-known Eigenfaces method.

2 The proposed approach

In the last decade, wavelets have become very popular, and new interest is rising on this topic. The main reason is that a complete framework has been recently built ([11, 4]) in particular for what concerns the construction of wavelet bases and efficient algorithms for its computation.

We based our approach on the wavelet decomposition of faces images for the reasons that we explain hereafter.

The main characteristic of wavelets (if compared to other transformations) is the possibility to provide a multiresolution analysis of the image in the form of coefficient matrices. Strong arguments for the use of multiresolution decomposition can be found in psychovisual research, which offers evidence that the human visual system processes the images in a multiscale way. Moreover, wavelets provide a spatial and a frequential decomposition of a the image at the same time.

Wavelets are also very flexible: several bases exist, and one can choose the basis which is more suitable for a given application. We think that this is still an open problem, and up to now only experimental considerations rule the choice of a wavelet form. However, the choice of an appropriate basis can be very helpful.

Computational complexity of wavelets is linear with the number (N) of computed coefficients ($O(N)$) while other transformations, also in their fast implementation, lead to $N \times log_2(N)$ complexity. Thus, wavelets are adapted also for dedicated hardware design (Discrete wavelet Transform). If the recognition task has real time computation needs, the possibility of embedding part of the process in Hardware is very interesting, like in compression tasks ([6]).

2.1 Wavelet packet decomposition

The (continuous) wavelet transform of a 1-D signal $f(x)$ is defined as:

$$(W_a f)(b) = \int f(x)\,\psi_{a,b}(x)\,dx \qquad (1)$$

$$\text{with } \psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)$$

The mother wavelet ψ has to satisfy the admissibility criterion to ensure that it is a localized zero-mean function. Equation (1) can be discretized by restraining a and b to a discrete lattice ($a = 2^n, b \in \mathcal{Z}$). Typically, some more constraints are imposed on ψ to ensure that the transform is non-redundant, complete and consitutes a multiresolution respresentation of the original signal. This leads to an efficient real-space implementation of the transform using quadrature mirror filters. The extension to the 2-D case is usually performed by applying a separable filter bank to the image. Typically, a low filter and a bandpass filter (H and G respectively) are used. The convolution with the low pass filter results in a so-called approximation image and the convolution with the bandpass filter in a specific direction results in so-called details image.

In classical wavelet decomposition, the image is split into an approximation and details images. The approximation is then split itself into a second-level approximation and details. For a n-level decomposition, the signal is decomposed in the following way:

$$A_n = \Big[H_x * \big[H_y * A_{n-1}\big]_{\downarrow 2,1}\Big]_{\downarrow 1,2} \qquad (2)$$

$$D_{n1} = \Big[H_x * \big[G_y * A_{n-1}\big]_{\downarrow 2,1}\Big]_{\downarrow 1,2} \qquad (3)$$

$$D_{n2} = \Big[G_x * \big[H_y * A_{n-1}\big]_{\downarrow 2,1}\Big]_{\downarrow 1,2} \qquad (4)$$

$$D_{n3} = \Big[G_x * \big[G_y * A_{n-1}\big]_{\downarrow 2,1}\Big]_{\downarrow 1,2} \qquad (5)$$

where $*$ denotes the convolution operator, $\downarrow 2,1$ ($\downarrow 1,2$) sub-sampling along the rows (columns) and $A_0 = I(x,y)$ is the original image. A_n is obtained by low pass filtering and is the approximation image at scale n. The details images D_{ni} are obtained by bandpass filtering in a specific direction and thus contain directional detail information at scale n. The original image I is thus represented by a set of subimages at several scales; $\{A_n, D_{ni}\}$.

The *wavelet packet decomposition*, that we use in our approach, is a generalization of the classical wavelet decomposition that offers a richer signal analysis (discontinuity in higher derivatives, self-similarity,...). In that case, the details as well as the approximations can be split. This results in a wavelet decomposition tree. Usually, an entropy-based criterion is used to select the deepest level of the tree, while keeping the meaningful information.

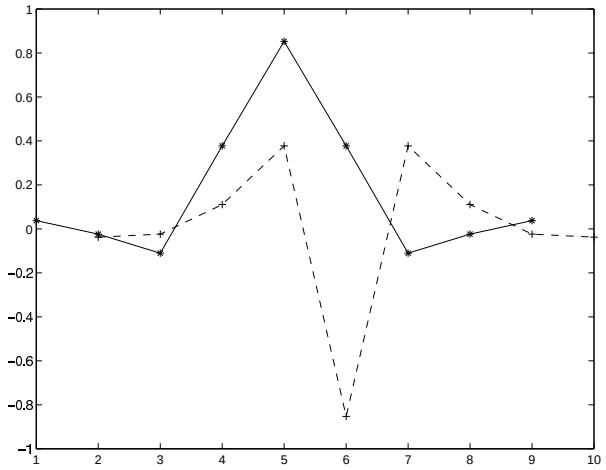

Figure 1. H (solid line) and G (dashed line) filters

In our experimentations, we have selected 2 levels of decomposition according to the size of the face images (as shown in figure 2) and we use the 16 resulting coefficient matrices which are displayed in figure 3. Figure 1 shows the H and G filters that have been applied. These filters have been selected based on trials during our experimentations. For each coefficient matrix, a set of statistical features is computed as described in the next section.

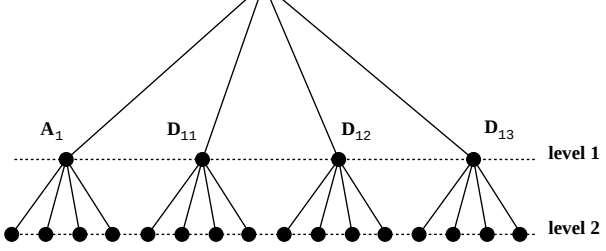

Figure 2. A wavelet packet tree

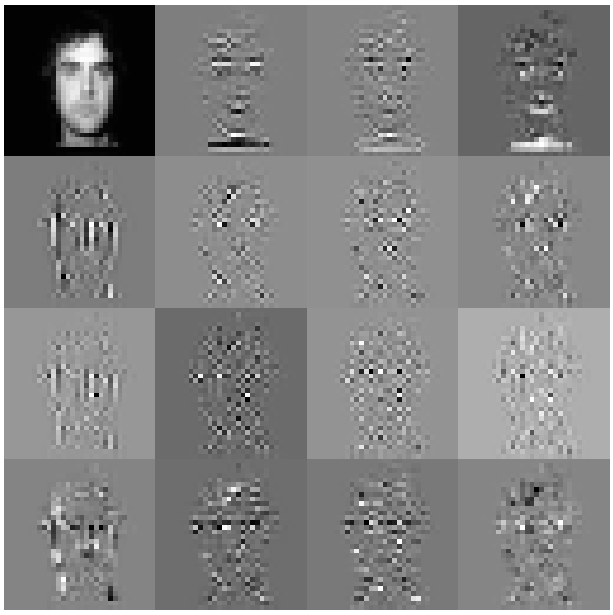

Figure 3. Level 2 of the wavelet packet tree

2.2 Feature vectors extraction

Before proceeding with wavelet packet analysis and feature extraction, we aim at segmenting the face image in order to separate the face from the background. Since the background is simple and homogeneous in the images that we process, (i.e., dark in the FACES database images and light in the FERET database images), we apply an iterative Lloyd quantization method ([8]) using 4 levels of quantization. Then, a rectangular area (bounding box) containing the face is obtained. After this step of preprocessing, the wavelet packet decomposition is performed on the whole image but the wavelet coefficients will be considered only in the face bounding box. An example of the quantization process results is presented in figure 4 . As mentionned above, a two levels wavelet packet decomposition is performed. There is no need to perform a deeper decomposition because, after the second level, the size of images is becoming too small and no more valuable information is obtained. At the second level of decomposition, we obtain one image of approximation (low-resolution image) and 15 images of details. Therefore, the face image is described by 16 wavelet coefficient matrices, which represent quite a huge amount of information (equal to the size of the input image).

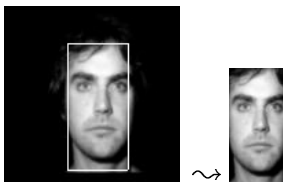

Figure 4. Lloyd quantization and extraction of the face bounding box

It is well-known that, as the complexity of a classifier grows rapidly with the number of dimensions of the pattern space, it is important to take decisions only on the most essential, so-called discriminatory information, which is conveyed by the extracted features. Thus, we are faced with the need of dimensinality reduction.

Each of the 16 coefficient matrices contains informations about the texture of the face. An efficient way of reducing dimensionality and characterizing textural information is to compute a set of moments. Thus, we extract 4 measures from the low-resolution image which are the mean value μ_{out} and the variance σ^2_{out} of the face outline by considering the border area (whose width is a percentage of the bounding box width, typically 30%) of the face bounding box, the mean value μ_{in} and the variance σ^2_{in} of the area inside the face bounding box (with less including hair or background). The outside area of the bounding box will give information about the face shape and the inside area will provide information about the face texture and the skin-tone. From the other 15 detail images, we extract the means μ_i and variances σ_i (i=2,..,16). In fact, the mean values μ_i are null, due to the design of the bank filters that we apply. Thus, the feature vectors contain a maximum of 19 components and are described as follows:

$$\mathcal{V} = \bigcup_{i=0}^{16} \left\{\mu_i, \sigma_i^2\right\} \tag{6}$$

where $\forall i \geq 2, \mu_i = 0$
$\mu_0 = \mu_{out}, \sigma_0^2 = \sigma_{out}^2$ and $\mu_1 = \mu_{in}, \sigma_1^2 = \sigma_{in}^2$.

In fact, after the extraction of all the vectors of the training set, we keep the most meaningful components by checking the mean value of each ot them for all the feature vectors. Only the components with a mean value above a predefined threshold are considered for feature vector formation. Typically, feature vectors of size 9 are built for a threshold value of 0.9.

2.3 Feature vectors classification

When solving a pattern recognition problem, the ultimate objective is to design a recognition system which will classify unknown patterns with the lowest possible probability of misrecognition. In the feature space defined by a set of features $X = [x_1, ..., x_n]$ which may belong to one of the possible m pattern classes

$\omega_i, i = 1, .., m$, an error probability can be defined but can not be easily evaluated ([7]). Thus, a number of alternative feature evaluation criteria have been suggested in the litterature [7]. One of these criteria is based on probalistic distance measures.

It is easy to show that, in the two-class case, the error probability e can be written:

$$e = \frac{1}{2}[1 - \int |p(X|\omega_1) P(\omega_1) - p(X|\omega_2) P(\omega_2)| dX] \quad (7)$$

According to equation (7), the error will be maximum when the integrand is zero, that is, when density functions are completely overlapping, and it will be zero when they don't overlap. The integral in (7) can be considered as the probalistic distance between the two density functions.

In our approach, the *Bhattacharyya distance* $\mathcal{B}$ is chosen as a probalistic distance:

$$\mathcal{B}(X) = -\ln \int \left[p(X|\omega_1) p(X|\omega_2)\right]^{\frac{1}{2}} dX \quad (8)$$

In the multi-classes case and to solve our problem, we make the assumption that the class-conditional probability distributions are Gaussian, that is, when the density functions are defined as:

$$p(X|\omega_i) = \left[(2\pi)^n |\Sigma_i|\right]^{-\frac{1}{2}} \times \exp\left\{-\tfrac{1}{2}(X-\mu_i)^T \Sigma_i^{-1} (X-\mu_i)\right\} \quad (9)$$

where μ_i and Σ_i are the mean vector and covariance matrix of the i^{th} class distribution respectively. The multivariate integrals in the measure can be evaluated which leads to:

$$B = \frac{1}{4}(\mu_2 - \mu_1)^T [\Sigma_1 + \Sigma_2]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \left[\frac{\left|\frac{1}{2}(\Sigma_1 + \Sigma_2)\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \right] \quad (10)$$

We consider that each component pair $\{\mu_i, \sigma_i^2\}$ is independent from the other component pairs of the feature vector $\mathcal{V}$. Thus, the distance between to feature vectors $\mathcal{V}_k$ and $\mathcal{V}_l$ is computed on a component-pair basis, that is, the distance is considered as a sum of distances relative to each of these component pairs. Using the Bhattacharrya distance, the distance $\mathcal{D}_i$ between the component pairs i of the two feature vectors $\mathcal{V}_k$ and $\mathcal{V}_l$ is:

$$\mathcal{D}_i(\mathcal{V}_k, \mathcal{V}_l) = \frac{1}{4} \frac{(\mu_{ik} - \mu_{il})^2}{(\sigma_{ik}^2 + \sigma_{il}^2)} + \frac{1}{2} \ln \left[\frac{\frac{1}{2}(\sigma_{ik}^2 + \sigma_{il}^2)}{\sqrt{\sigma_{ik}^2 \, \sigma_{il}^2}} \right] \quad (11)$$

with $\forall i = 2, .., n,\ \mu_{ik} = \mu_{il} = 0$ where $n+1$ is the size of the feature vectors.

As a consequence, the resulting distance $\mathcal{D}$ between two feature vectors $\mathcal{V}_k$ and $\mathcal{V}_l$ can be chosen as:

$$\mathcal{D}(\mathcal{V}_k, \mathcal{V}_l) = \sum_{i=0}^{n} \mathcal{D}_i(\mathcal{V}_k, V_l) \quad (12)$$

3 Experimental Results

In order to test the efficiency of the algorithm presented above, we performed a series of experiments using two different sets of test images. The first set is extracted from the FERET database. This is a collection of 234 images of 117 individuals (2 images per person). The second set is extracted from the FACES database of the MIT Media Lab used in the Photobook project ([12]), and contains 150 images of 50 individuals (3 images per person). In both of these databases, the images that belong to the same person (same class) usually present variations in expression, illumination. In addition, they are not well-framed (variations in position) in the FERET database.

Sample images from the two sets are displayed in figures 5 and 6.

Figure 5. Sample images from FACES database

Figure 6. Sample images from FERET database

3.1 Experiment 1

In this experiment, we first extract the feature vectors of all the images in the data set and then form the mean vectors of each class c (namely $\mathcal{V}_c^{mean}$), that is, we use an intra-class information. Then, we verify that each image k is classified into the correct class, looking for the minimum $\mathcal{D}(\mathcal{V}_k, \mathcal{V}_c^{mean})$ distance, for each class c. Every experiment was performed using fractions of the available images in the whole dataset. By this way, we are able to study how the size of the image dataset

affects the recognition performances. The results of the experiments are displayed in table 2 and table 1.

Number of Images	Number of Misclassified	Recognition rate
60	0	100.0%
90	0	100.0%
120	0	100.0%
150	6	96.0%

Table 1. Results for the FACES database, experiment 1

Number of Images	Number of Misclassified	Recognition rate
150	2	98.6%
160	2	98.7%
170	2	98.8%
180	2	98.9%
190	3	98.4%
200	6	97.0%
210	6	97.1%
220	6	97.2%
234	9	96.1%

Table 2. Results for the FERET database, experiment 1

From these results, it can be seen that the recognition rates vary from 100.0% to 96.0%, with scores of 96.0% and 96.1% for the whole set of images in FACES and FERET respectively. These results are good if we consider the quite significant number of faces to be classified. In the FACES database, perfect classification is obtained if we use up to 120 images. Above all, these results are very similar for both databases which may mean that the proposed method is stable and tolerant to changes in appearance as well as changes in position.

3.2 Experiment 2

This experiment was performed using the images of the FACES database. Since 3 images of each individual are available, we use the first two as training data (in order to compute the mean vector) and the third image as a test image. The results are displayed in table 3. It can be seen that the recognition rate for the whole dataset decreases from 96.0% to 92.0%, which means that only two available images of each class seem not to be enough to estimate a good mean class vector, according to the face variations. Therefore, using the mean class vector seems to improve the classification results.

Number of Images	Number of Misclassified	Recognition rate
60	1	98.3%
90	2	97.7%
120	3	97.5%
150	12	92.0%

Table 3. Results for the FACES database, experiment 2

3.3 Experiment 3

In order to check the discriminatory properties of our scheme, we perform the features vector classification as in experiment 2, but without using any class information, that is, without computing the class mean vectors. Results are presented in tables 4 and 5. The recognition rates for the both whole sets of images are 92.0% and 91.4% respectively, which are still high, given that no intra-class information is used.

Number of Images	Number of Misclassified	Recognition rate
60	2	96.6%
90	3	96.6%
120	4	96.6%
150	12	92.0%

Table 4. Results for the FACES database, experiment 3

Number of Images	Number of Misclassified	Recognition rate
150	13	91.3%
160	13	91.8%
170	13	92.3%
180	14	92.2%
190	17	91.0%
200	19	90.5%
210	19	90.9%
220	19	91.3%
234	20	91.4%

Table 5. Results for the FERET database, experiment 3

3.4 Comparison with the Eigenfaces method

In the Eigenfaces approach, each image is treated as a high dimensional feature vector by concatening the rows of the image together, using each pixel as a single feature. Thus, each image is considered as a sample point in a high-dimensional space. The dimension of the

feature vector is usually very large, on the order of several thousands for even small image sizes (in our case, the image size is $128 \times 128 = 1024$). The Eigenfaces method which uses PCA is based on linearly projecting the image space to a lower dimensional space, and maximizing the total scatter across all classes, i.e, accross all images of all classes ([15, 12]). The orthonormal basis vector of this resulting low dimensional space are reffered as eigenfaces and are stored. Each face to recognize is then *projected* onto each of these eigenfaces, giving each of the component of the resulting feature vector. Then, an euclidian distance is used in order to classify the features vector. In figures 7 and 8, the first 6 computed eigenfaces of the FACES and FERET databases respectively are displayed.

Figure 7. the first 6 eigenfaces of the FACES database

Figure 8. the first 6 eigenfaces of the FERET database

We applied the Eigenfaces method on both databases. We obtain very good results on the Faces database images which is actually not surprising. Indeed, in that case, the images have been normalized (well-framed) especially for the PCA method. We obtain a result of 99.33% good classification (1 error for 150 images) using 40 eigenfaces compared to 96.0% using our approach. But, one drawback of this method is that these eigenfaces (the number of eigenfaces has to be approximately one third of the total number of images) have to be stored, which supposes an amount of extraspace in the database. A second disadvantage is that images have to be normalized. In the FERET database case, the images are not normalized as in the FACES case, and the remaining error is 87 (i.e 62.82% good) even if more than 50 eigenfaces are used. Without any normalization needs and above all without any eigenface computation and storage, the results obtained by our approach are much better that those obtained by applying PCA in the FERET database case.

Another key point of our scheme, compared to the Eigenfaces method, is the compact size of the feature vectors that represent the faces and above all, the very high matching speed that we provide. Indeed, the time required to perform the wavelet packet analysis of a test image and to extract the feature vectors is of approximatively 0.05 s. on a SUN-Ultra 1 workstation, while the time for comparing a test image to the whole database (150 images) is 0.021 s. The PCA method requires quite a long time of training in order to compute the eigenfaces and the recognition process is as well expensive because it is correlation-based: the test image has to be correlated with each of the eigenfaces.

4 Conclusion

Our experiments show that a small transform of the face, including translation, small rotation and illumination changes, leave the face recognition performance relatively unaffected. For both databases, good recognition rates of approximately 96.0% are obtained. Thus, the wavelet transform proved to provide an excellent image decomposition and texture description. In addition to this, very fast implementations of wavelet decomposition are available in hardware form. We show that even very simple statistical features such as mean and variances provide an excellent basis for face classification, if an appropriate distance is used. The use of the Bhattacharyya distance proved to be very efficient for this purpose. As an extension of this work, we believe that it would be interesting to extract the statistical features from the wavelet decomposition of more specific facial features such as eyes, mouth and nose. That will not increase much the size of the feature vector but we will have previously to detect the features location in order to extract the values. However, detecting features is by itself a difficult and time consuming process so this strategy will increase the time that actually will be needed for recognition. Therefore, we will focus on a fast and efficient algorithm for features detection.

Acknowledgments

This work was funded in part under the DiVAN Esprit Project EP 24956.

References

[1] P. Belhumeur, J. Hespanha, D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 19(7):711–720, July 1997.

[2] R. Brunelli, T. Poggio. Face Recognition: Features versus Templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 11(6):1042–1052, 1993.

[3] Y. Dai, Y. Nakano. Recognition of facial images with low resolution using a Hopfield memory model. *Pattern Recognition* 31(2):159–167, 1998.

[4] I. Daubechies. The Wavelet Transform, Time-Frequency Localization and Signal Analysis. *IEEE Transactions on Information Theory* , 36(5):961-1005, 1990.

[5] EP 24956. Esprit Project. Distributed audioVisual Archives Network (DiVAN). http:\\divan.intranet.gr\info, 1997.

[6] M. Ferretti, D. Rizzo. Wavelet Transform Architectures: A system Level Review. *Proc. of the* 9^{th} *International Conference ICIAP'97*, Vol.2, pp. 77-84, Florence, Italy, September 1997.

[7] Y. Fu. Handbook of Pattern Recognition and Image Processing. *Academic Press*, 1986.

[8] A. Gersho, R.M. Gray. Vector Quantization and Signal Compression. Kluwer Academic Publisher, 1992

[9] M. Kirby, L. Sirovich. Application of the Karhunen-Loeve Procedure and the Characterization of Human Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 12(1):103–108, 1990.

[10] A. Lanitis, C.J. Taylor, T. F. Cootes. Automatic Interpretation and Coding of Face Images Using Flexible Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 19(7):743–756, July 1997.

[11] Mallat S., Multifrequencies Channel Decompositions of Images and Wavelets Models. *IEEE Transactions on Acoustics, Speech and Signal Processing* , 37(12),1989.

[12] A. Pentland, R.W. Picard, S. Sclaroff. Photobook: Content-Based Manipulation of Image Databases. In *Proceedings of the SPIE Storage and Retrieval and Video Databases II*, No. 2185, San Jose, 1994.

[13] J.L Perry, J.M Carney. Human Face Recognition Using a Multilayer Perceptron. In *IJCNN*, Washington D.C., pp. 413–416, 1990.

[14] D. L. Swets, J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 18(8):831–836, August 1996.

[15] M. Turk, A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Science* , 3(1):71–86, 1991.

[16] C. L. Wilson, C. S. Barnes, R. Chellappa, S. A. Sirohey. Face Recognition Technology for Law Enforcement Applications. *NISTIR 5465, U.S. Department of Commerce*, 1994.

[17] L. Wiskott, JM. Fellous, N. Kruger, C. Von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 19(7):775–779, July 1997.