A Pruning Approach Improving Face Identification Systems

Anis CHAARI[12], Sylvie LELANDAIS[1]
[1]IBISC Laboratory, CNRS FRE 3190
Evry University
Evry, FRANCE
anis.chaari@ibisc.fr, s.lelandais@iut.univ-evry.fr

Mohamed BEN AHMED[2]
[2]RIADI Laboratory, National School of Computer
Science, Manouba University
La Manouba, TUNISIA
mohamed.benahmed@riadi.rnu.tn

***Abstract*—We propose, in this paper, a new biometric identification approach which aims to improve recognition performances in identification systems. We aim to split the identity database into well separated partitions in order to simplify the identification task. In this paper we develop a face identification system and we use the reference algorithms of *Eigenfaces* and *Fisherfaces* in order to extract different features describing each identity. These features, which describe faces, are generally optimized to establish the required identity in a classical identification process. In this work, we develop a novel criterion to extract features used to partition the identity database. We develop database partitioning with clustering methods which split the gallery by bringing together identities which have similar features and separating dissimilar features in different bins. Pruning the most dissimilar bins from the query identity features allows us to improve the identification performances. We report results from the XM2VTS database.**

Keywords-Biometry; face identification; clustering; feature extraction; image database.

I. INTRODUCTION

We propose in this paper a new approach improving identification performances in biometric recognition systems. The classical process of identification consists in matching features of the unknown identity with the equivalent features of all users already stored in the database system (gallery). The unknown person to be recognised or the probe (image containing his face) is identified like the user (enrolled person in the database) having the features which resemble the more to the probe features (1:N match). Within the framework of identification, we don't have any a priori information of the probe identity. Thus, identification systems have a significant decrease of performances according especially to the potential number of system users (size of the database) [1]. They become more complex and very computing expensive especially with large biometric databases [2].

The system we propose aims to reduce the complexity and to improve the performances of biometric identification. To achieve this goal, we add, to the classic identification scheme, an offline partitioning phase of the gallery upstream the online searching phase. We propose a partitioning scheme of the database according to the similarity of the most discriminatory features we could extract from the facial images. Therefore, our system is based essentially on two critical tasks namely the extraction of discriminating features and the partition of the database.

Several studies in this context of databases partitioning have been developed to identify individuals by their fingerprints [3, 4, 5]. In fact, fingerprints possess a global morphological structure that helps to discriminate and separate them into well defined classes. Global ridge and furrow structure forms special patterns in the central region of the fingerprint. These patterns are learned and supervised methods of classification are used to separate them into the five well defined classes, namely, left loop, right loop, whorl, arch and tented arch. The identification of a query fingerprint is performed in a subset of the gallery, which corresponds to its membership class.

Having other perspectives to identify the sex or the ethnicity of a person, classification of facial images according mainly to their sex membership was extensively studied [6, 7, 8, 9, 10]. Although the classes are known a priori, data extracted from faces are generally implicit and any explicit pattern has been found to discriminate classes. Thus the attributes extracted from facial images could not be compatible with the learned distribution of classes and therefore not sufficiently discriminating.

Unlike these systems which require supervised classification methods, we address unsupervised learning methods to classify and partitioning facial images according to the extracted features. Few studies have attempted to cluster any biometric modality which hasn't morphological features defining some categories unlike fingerprints. The authors in [11, 12] have proposed an approach which performs clustering using hand geometry and signature features to reduce the search space upstream the exhaustive identification. In addition, such clustering approaches are not available for other biometry such as face or iris.

Unsupervised classification (or clustering) aims to provide homogenous and well separated classes. Indeed, the general rule of data clustering is to minimize the intra-class and to maximize the inter-class data variance. We develop, in this paper, clustering with K-means algorithm in order to partition the face database and to reduce the search space to a subset of the whole database.

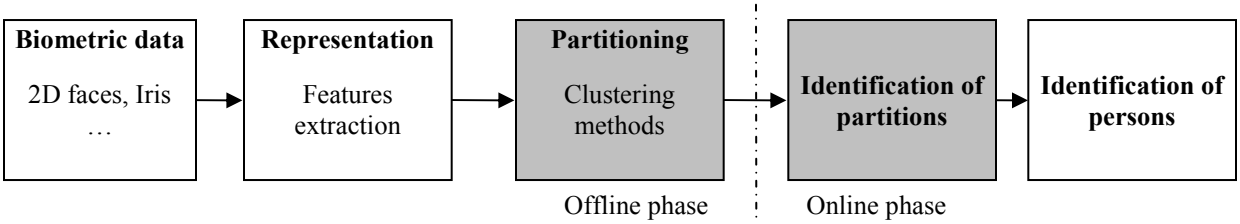

Figure 1. The different steps of our identification system.

II. CONCEPTUAL APPROACH

We describe, in this section, our conceptual approach of biometric identification systems. Figure 1 describes the hierarchy of the different processes, which constitute our identification system. The gray rectangles show the original stages we conceive in addition to the conventional identification steps. We introduce mainly a clustering based approach to divide the gallery into several subsets. This partitioning process aims to group the data (identities) that have similar characteristics and separate dissimilar ones into different bins. We propose, through this partitioning process of the gallery, to retain the closest clusters from the query image and to keep their representing identities for the final step of identification. Pruning the most dissimilar partitions with the query identity features lead to simplify the search in the gallery. This simplification is induced by reducing the number of identity from the gallery. Finally, we operate the identification process on the retained identities from the gallery simply by similarity measurement.

We develop hereafter the partitioning stage which aims to divide the database into several clusters. The extracted vectors of features for the partitioning task are assimilated to multi-dimensional data points occupying the face space F. Data clustering algorithms have the aim to identify the sparse and the crowded places of this space and divide consequently the distribution points into groups in such way to minimize the intra class variance and maximize the inter class variance. Clustering problem is then formalised as follows: given the desired number of clusters K and a dataset of N points, and a distance-based measurement function, it is a question of finding a partition of the dataset that minimizes the value of the measurement function [13]. Data points are then assigned to clusters C_i so that each cluster must contain at least one data point; and each data point may belong to one and only one cluster as show (1):

$$\bigcup_{i=1}^{K} C_i = \Phi \quad and \quad C_i \cap C_j = \emptyset, \; i \neq j \tag{1}$$

Following such binning, the biometric database will be partitioned such that features (or templates) in each bin are similar and correspond to a statistical class. We show in figure 2 some representation of clusters by using *Eigenfaces* features. We remark that faces in each bin are homogenous and present some common patterns. These results demonstrate the relevance of our clustering approach.

We differentiate features which will be used for the database partitioning and those which will be used for the final identification process. We can develop these two kinds of features from the same algorithm like Eigenfaces. However, in this case partitioning features are extracted differently from identification features. In fact, partitioning features constitute the input of clustering algorithm, whereas identification features will be matched with a similarity measure in the identification process. On the other hand, the use of different algorithms (multi-algorithmic) that represents as independently as possible the information contained in facial images is needed between the classification and the identification processes. Our aim is to generate several and different features for each person in the gallery. This allows to operate the partition of the gallery following some patterns different from those used for identification. The dependence level of these patterns will affect undoubtedly the final results of identification after the pruning process.

We develop in the next section the subspaces methods of *Eigenfaces* and *Fisherfaces*. *Eigenfaces* method tries to produce a new face space whose axes define the data variation. On the other hand, *Fisherfaces* find axes which better discriminate the different samples of each identity. We develop in this work these two different algorithms in order to combine them for partitioning and identification tasks in our approach framework.

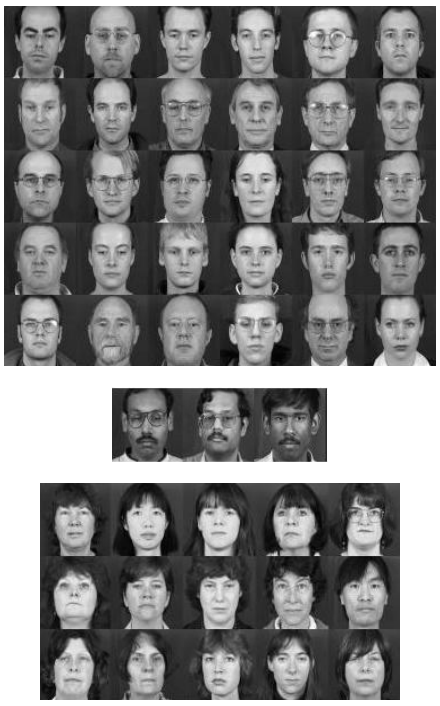

Figure 2. Some clusters from the XM2VTS partitioning.

III. FEATURE EXTRACTION

A very significant point in a pattern recognition problem is to extract pertinent and discriminative features which will constitute the data input to the classification module. We distinguish, through our experiments, the main pertinent features as well for identification task as for clustering task. We report results in term of Identification Rate and the Top Match Rank as show the figure 3.

The Identification Rate (*IR*) is the most commonly used evaluation result and criteria for identification systems. But in the case of error, it is useful to know the rank of the right response among a number of ordered scores. Then we trace the Cumulative Match Score (*CMS*) curve which represents the probability that the right identity is among a set of nearest scores.

The shape behavior of the cumulative match score curve could have a considerable importance for many pruning applications. For instance, we can cite the law enforcement applications where we analyze in posteriori the content of surveillance systems. In these systems, the goal is to reduce the potential identities of the gallery which could be matched to the query one in order to improve the decision stage of identification. For these systems, we seek to retain the minimum number of identities from the gallery, in which we find potentially the searched identity. To assess this functional point, we search the rank from which we certainly have the right identity that we denote by Top Matches Rank - *TMR*. This rank is determined as the x-axis coordinate where the CMS curve reaches 100% identification rate. The *TMR* may give an indication of the maximal distance that has a probe feature with its corresponding one from the database. The choice of feature vector that minimizes the *TMR* induces minimization of this distance. Since clustering module separates people into partitions, based primarily on the distance between their features, we opt mainly on this criterion (*TMR*) to select features for the clustering task.

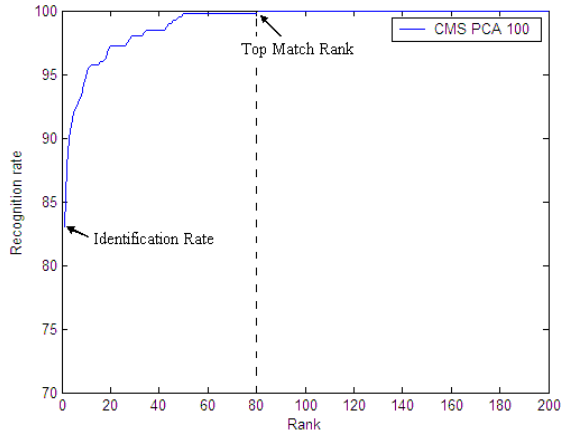

Figure 3. Eigenfaces Cumulative Match Score curve indicating Identification Rate and Top Match Rank.

The choice of features for the identification task is obviously guided by those giving the best identification rate. On the other hand, for the database partitioning, we choose smallest features (with minimal size) which mainly minimize the *TMR*. Smallest features are preferred in order to reduce the dimensionality of the classification problem. We can deduce already a simplification of the gallery through the ratio between the *TMR* and the database size. Thus this criterion (*TMR*) is very important for our partitioning problem in order to prune a subset of the database.

Eigenfaces and *Fisherfaces* are reference face recognition methods which build different face spaces. In order to produce the feature vector, these two methods proceed similarly by projecting the facial image on the new space. A two dimensional image of size p by q pixels is generally represented as a vector $\boldsymbol{x}$ of size n ($n = p \times q$) in a high dimensional space. A sample of face vector $\boldsymbol{x}$ can be expressed as linear combination in a basis Φ of the new face space $\mathcal{F}$ which has a lower dimension ($m << n$):

$$x = \sum_{i=1}^{n} a_i x_i \approx \sum_{i=1}^{m} \alpha_i \varphi_i \qquad (2)$$

We discuss in the next sections the choice of *Eigenfaces* and *Fisherfaces* features $\{\alpha_i\}$.

A. Eigenfaces Features

To determine the optimal value of g, we can study the eigenvalue spectrum λ_i which correspond to the eigenvectors φ_i. A natural algorithm determining g is to seek the threshold from which the eigenvalues are very small.

Kirby and Sirovitch [14] have introduced the first and the most used criterion for eigenvectors selection that is the inertia of the dimension [15]. The ratio of inertia from the first j eigenvectors is:

$$\frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \qquad (3)$$

λ_j is the eigenvalue associated to the j^{th} eigenvector.

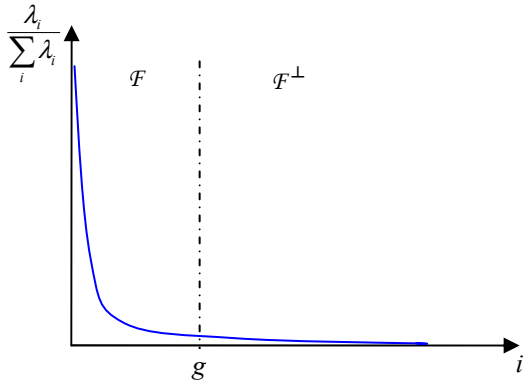

Figure 4. Typical aspect of eigenvalues sort by a decreasing order.

Swets et Weng [16] advocate the use of a ratio of 95% of inertia, while Kirby [15] uses a rate of 90%. Kirby [15] introduced another criterion, defined as the ratio $s_i = \lambda_i / \lambda_1$ between the eigenvalue of φ_i axis and the largest eigenvalue λ_1. Only eigenvectors which s_i are higher than a threshold τ are retained (τ = 1%). Both these methods provide roughly the same performances in terms of identification rate.

For our partitioning problem, we aim to minimize the *TMR* as explained above. We assess the different sizes of face spaces following the XM2VTS protocol of Lausanne [17] with a gallery of 200 persons. Figure 5 shows the *TMR* curve for the different subspaces inertia corresponding to the different feature sizes.

We see from Figure 5 that the *TMR* is minimal (around 80) with subspaces inertia from 80 to 85% of the total face space. The corresponding sizes of features are ranging from 44 to 66. In order to reduce the dimensionality of the classification problem, we choose rather the feature vector of size 44 to operate clustering. By obtaining almost a *TMR* of 80 identities out of 200 in the gallery, we carry out already a simplification of 60% of the gallery.

B. *Fisherfaces Features*

Fisherfaces method is based on solving the Linear Discriminant Analysis (LDA) using the following algorithm. The first step consists to perform PCA in the original face space by retaining M eigenvectors. The second one projects the within class scatter matrix S_w and the between class scatter matrix S_b on the reduced space. Finally it is to rebuild a PCA on the reduced scatter matrices. The deduced *Fisherfaces* space contains a number g of axes. Thus the *Fisherfaces* technique is dependent on two parameters M and g that Belhumeur et al. [18] choose to set at $M = N - k$ (maximum value of M such that S_w is full rank), and $g = k - 1$. N is the number of image in the learning set and k is the number of identities in this set. Swets and Weng presented a similar technique in [16], but based on the use of $M < N - k$ axis. They choose the value of the parameter M using the dimension inertia of 95%.

We have studied, many values of $M < N - k$ according to the technique of Swets et Weng [16]. The choice of M is achieved by using the dimension inertia. We tried different thresholds of inertia ranging from 10% to 90% in order to build the first step components. On the other hand, we chose the g value as the minimum of $N - k$ and M. Table 1 shows the Identification Rates (*IR*) and the Top Match Rank (*TMR*) values corresponding to the chosen levels of inertia.

Sub spaces containing 30% of the total inertia of the learning set of faces optimize the *TMR*, while sub spaces with 65% of the total inertia optimize the identification rate. We retain features of these dimensions for the clustering and identification process respectively.

IV. IDENTIFICATION OF PARTITIONS

Given a facial image that we want to establish its identity, we compute its partitioning feature that we compare to the center of each bin. The searched identity is potentially in the nearest bin. However, by choosing identities that belong only to this partition, we increase the probability of an error discard of the searched identity. This error called Research Error Rate (*RER*) is introduced in [19]. The *RER* tends to compromise the next identification step and degrade its accuracy. Therefore, we make the following study to choose the partition which minimize and cancel this error rate.

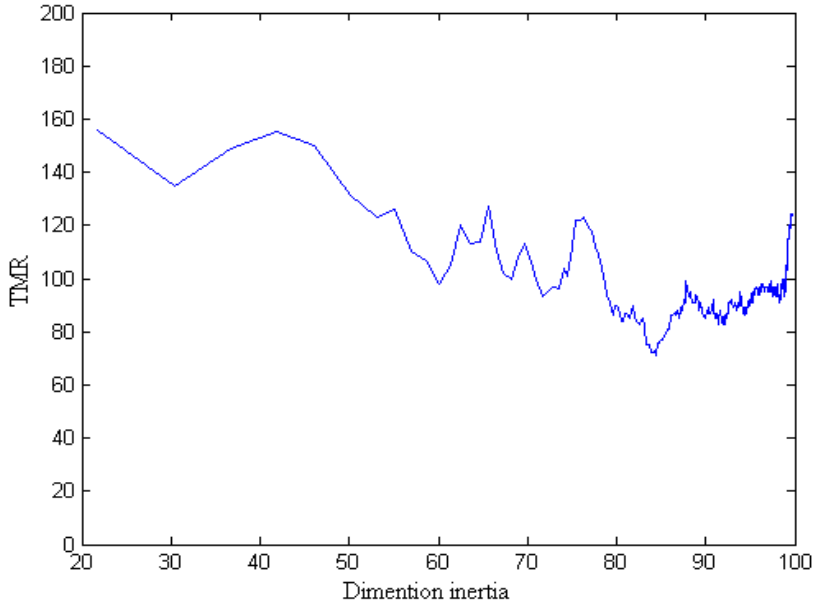

Figure 5. The TMR behaviour by increasing the subspaces inertia of Eigenfaces.

TABLE I. *IR* AND *TMR* PERFORMANCES OF FISHERFACES ALGORITHM IN FUNCTION OF M AND M PARAMETERS.

Inertia	90%	85%	80%	75%	70%	65%	60%	55%	50%	40%	35%	30%	25%	20%	15%	10%
M	400	338	284	238	200	165	136	111	90	56	42	31	22	15	9	5
g	199	199	199	199	199	165	136	111	90	56	42	31	22	15	9	5
IR	88,5	88,75	87,25	85,75	88,25	89,25	89	88,75	89	87,5	85,5	82,5	76,5	66,25	55,5	34,5
TMR	122	119	155	136	72	71	46	55	52	60	58	49	97	85	114	155

By subdividing the database into several partitions, the probe identity is potentially found in the closest bin. However, the probability to find the query identity in the closest bin is relatively small and decrease by increasing the number *K* of bins. To prevent missing the bin in which lies the query identity, we propose to perform research into the *P* closest bins. The *P* closest bins constitute the subset on which we perform the last task of identification. If we retain a large number *P* of bins, the simplification of the gallery is marginal and the identification accuracy could hardly be improved. However, if we accept one or relatively few classes as a subset of research from the gallery, the *RER* will surely affect the identification performance. Thus, the estimation of the number *P* of bins is quite difficult. In fact, the parameter *P* is related to the total number *K* of bins, to the intrinsic quality of extracted features and to the clustering method.

By analogy with the Cumulative Match Score curve used to assess identification performances, we delineate the Binning Cumulative Match Score curve that we denote by *BCMS*. In fact, to evaluate our partitioning approach, we compute the classification rate that is the probability to find the query identity in the closest bin. But in case of error, it is useful to know the rank of the bin where the probe identity is. Fig. 6 illustrates a *BCMS* curve with a clustering into a number *K* of 10 bins. For a probe face, the *P* closest bins constitute the retained partition of the gallery. Thus, we control an *RER* equal to 0%.

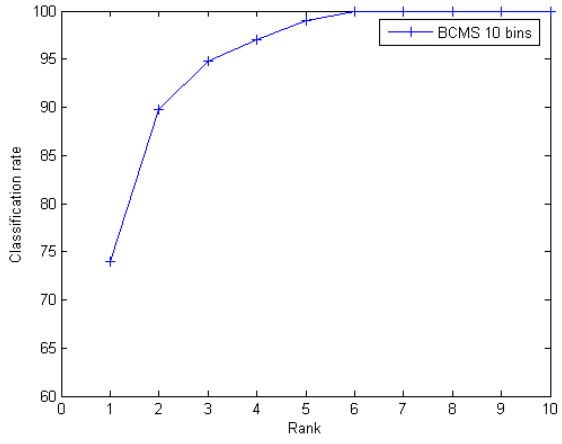

Figure 6. Binning Cumulative Match Score curve with a clustering on 10 classes.

The penetration rate *R* (∈ [0,1]) is the percentage of the retained partition compared to the total size of the gallery. The simplification rate can be derived as 1 - *R*. By assuming that the built bins contain the same number of identities, an approximate value of the penetration rate is given as follows:

$$R \approx \frac{N_B P}{N} \approx \frac{P}{K} \tag{4}$$

where N_B is the average number of identity features per bin, *N* is the database size and *K* is the number of bins.

However, we don't find usually the same number of identities per bin. For instance, for a 10 classes binning of our gallery which is composed of 200 individuals, the average number of data points per each class is 20. But these bins include really a number which varies from 3 to 35 identities. This inequality is due to the non-uniformity occupation of the face space where there are inevitably non uniform distributions with sparse and crowded places. Fig. 2 shows an example of a first bin which contains 30 identities and a second bin representing only with three persons. Thus, the effective penetration rate is the ratio of the number of identities from the retained partition and the size of the gallery. We report in the table below the effective penetration rate *R* from various clustering experiments on *Eigenfaces* features with different number *K* of bins.

We remark that raising the number of bins decreases the penetration rate until a number about 70 bins. After that, the penetration rate makes a climb. By considering the complexity of the clustering problem, we retain the 70 bins classification in order to simplify the search space. This procedure avoids any false detection and ensures the simplification of a substantial part of the database. The effectiveness of this procedure is confirmed by using a different validation sets and various learning strategies. All this experiences give the same results in terms of penetration and simplification rates.

TABLE II. THE KEPT NUMBER OF BINS AND THE PENETRATION RATE.

K	10	20	30	40	50	70	90	100
P	6	13	15	24	26	29	38	48
R	142.5	156.7	136.3	133.7	129	107	105	117

V. IDENTIFICATION RESULTS

Combination of the *Eigenfaces* and *Fisherfaces* representation algorithms does not improve the identification rate. However, our approach doesn't degrade the performance of the classical identification. To examine the behavior of all responses returned by our system of identification, we plot the cumulative match scores curve. Fig. 7 illustrates the CMS curve of our approach in red and that of classical identification in blue. We combined *Eigenfaces* and *Fisherfaces* methods alternately for clustering and identification. Given that the best results for identification are provided by *Fisherfaces*, the best combination shown in figure 7 is achieved by *Eigenfaces* features (g=44) for clustering and *Fisherfaces* one (g=90) for identification.

VI. CONCLUSIONS

We have proposed in this paper a framework of clustering and partitioning facial databases in order to improve biometric identification within these databases. Clustering was achieved on *Eigenfaces* features using the k-means clustering algorithm. The final step of identification is performed through *Fisherfaces* features. We improve just a little bit results of identification. But the use of more independent features from those used in the database clustering could improve more and more identification performances. Our future work involves to develop other face representing and clustering methods. We plan also to extend the evaluation of our recognition algorithms with a large scale database gathered from various benchmark datasets like FERET, FRGC, ORL and IV² databases.

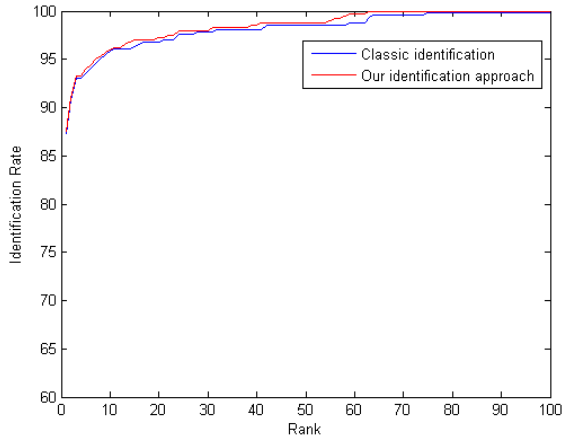

Figure 7. Comparison of the Cumulative Match Score curves of our approach and the classical approach of identification.

REFERENCES

[1] S. Z. Li, A. K. Jain, Handbook of face recognition, Springer, 2004.

[2] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone, FRVT 2002: Evaluation report, March 2003.

[3] A. K. Jain, S. Pankanti, "Fingerprint classification and matching". Handbook for Image and Video Processing, A. Bovik (ed.), Academic Press, 2000.

[4] X. Tan, B. Bhanu, Y. Lin, "Fingerprint identification: classification vs. indexing", IEEE Conference on Advanced Video and Signal Based Surveillance, pp.151-156, 2003.

[5] J. Gu, J. Zhou, Analysis of singular points in fingerprints based on topological structure and orientation field, International Conference on Biometrics, 2007.

[6] A. S. Tolba, Invariant gender identification, DIGITAL SIGNAL PROCESSING, 11(3), pp. 222 – 240, 2001.

[7] N.P. Costen, M. Brown, S. Akamatsu, Sparse models for gender classification. IEEE Proceeding in Automatic Face and Gesture Recognition, pp. 201- 206, May 2004.

[8] X. Lu, H. Chen, A. K. Jain, Multimodal Facial Gender and Ethnicity Identification, International Conference on Biometrics (ICB), pp.554-561, 2006.

[9] Z. Yang, M. Li, H. Ai, "An Experimental Study on Automatic Face Gender Classification," Proc. 18th IEEE Int'l Conf. Pattern Recognition, vol. 3, pp. 1099-1102, Aug. 2006.

[10] Z. Xu, L. Lu, P. Shi, A Hybrid Approach to Gender Classification from Face Images, the 19th International Conference on Pattern Recognition, 2008.

[11] S. Palla, S. Chikkerur, V. Govindaraju, Classification and indexing in large biometric databases, Biometrics Consortium Conference, Crystal City, VA, September 2004.

[12] A. Mhatre , S. Palla , S. Chikkerur, V. Govindaraju, "Efficient Search and Retrieval in Biometric Databases", SPIE Defense and Security Symposium, Vol - 5779, pages:265-273, March-2005.

[13] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: a review", ACM Computing Surveys, 31(3) pp. 264-323, 1999.

[14] 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, Jan. 1990.

[15] M. Kirby, Dimentionality reduction and pattern analysis : an empirical Approach, Wiley, New York, 2000.

[16] D. L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Analysis and Machine Intelligence, 18, 831– 836, 1996.

[17] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maître, XM2VTSBD: the extended M2VTS database, Int. Conf. on Audio & Video-based Biometric Authentication (AVBPA), pp. 72-77, 1999.

[18] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19, pp. 711-720, July 1997.

[19] D. Maltoni, D. Maio, A.K. Jain, S. Prabhakar, "Handbook of fingerprint recognition", Springer, 2005.