Nonparametric classifier for Face recognition system

W. Drira, F. Ghorbel

GRIFT Research Group, CRISTAL Laboratory, National School of Computer Sciences,
University of Manouba, Tunisia

Wissal.drira@gmail.com, Faouzi.ghorbel@ensi.rnu.tn

Abstract

In this paper, we introduce an optimal face recognition system based on Bayes classifier under nonparametric and supervised assumption, using the modified kernel estimator within the adjustment of the smoothing parameter in the sense of Mean Integrated Squared Error. Various experiments on different face databases are provided to illustrate the importance of the proposed system in face recognition system practice.

Keywords

Face recognition, Bayes, classification, Optimization, dimension reduction, kernel method.

1 Introduction

Face recognition has been an active research area over the last thirty years. It has been studied by scientists from different areas of psychophysical sciences and those from different areas of computer sciences. Face recognition has applications mainly in the fields of biometrics, access control, law enforcement, and security and surveillance systems.

The problem of face recognition can be stated as mainly a classification problem: Training the face recognition system with images from the known individuals and classifying the newly coming test images into one of the classes is the main aspect of the face recognition systems. For the type of data, the number of variables is often in the hundreds or thousands and sometimes much larger than the number of observations. Classification of such a high-dimensional data is a difficult problem that appears in such application. Then, the difficulty of this problem that comes from the large number of variables present challenges to classification methods and makes many conventional techniques impractical. A natural solution is to add a dimension reduction step before proceeding to the classification phase. Many discriminate analysis methods have been proposed in the literature such as Linear Discriminate Analysis LDA [3], Approximation of the Chernoff Criterion ACC [2] and Information Discriminate Analysis IDA [7]. In our previous work [5,6], we introduced a nonparametric probabilistic discriminate analysis method based on a new estimate of the L^2-Probabilistic Dependence Measure (L^2-PDM) to the multi-class case. Since this stochastic approach is based on probabilistic distances, theoretically this method approaches indeed the nonparametric Bayes classifier. Unfortunately, in practice, the optimization of this algorithm cannot be solved analytically.

In this sense, we propose an optimization algorithm of classification fusion of the optimal L^2-PDM in the sense of the numerical maximization and the Bayes classifier basing on the adjusted kernel method estimation of the smoothing parameter, this optimization is illustrated in the second section of this article. With the aim to illustrate this fact, a performance evaluation by the mean of the misclassification error was performed and illustrated for different face databases in the last section of this article.

2 Optimization algorithm to achieve a nonparametric Bayes Classifier

The discriminate analysis based on stochastic method such the L^2-PDM prepares the application of Bayes classifier in a nonparametric case for high dimensions, since there is a kind of equivalence between these two quantities: This means that the proposed estimator of the L^2-PDM is theoretically given the lower miss classification error in this direction. On the practice field, this result can not be reached since the optimization method is not analytical. So to approach the Bayesian classifier in the general case, we proposed a series of non-parametric procedures, consists of the following tasks:

1) <u>Selection features and optimization</u>

First, to estimate the optimal linear transformation of the L^2-PDM introduced in [5], we

proceeded as follows:
For i from 1 to N_{it} :

1. **1.** Generate $(W_{IDA})_i$ for the system initialization as [7]:

$$(W_{IDA})_i = arg\ \max_{W\in R^{dxD}} \{\mu I(WX;Y)\}$$

Where μI is the mutual information presented by : $\mu I(X;Y=k) = \frac{1}{2}\log|S| - \sum_{k=1}^{K}\pi_k \log|S_k|$

$$S_k = \sum_{k=1}^{K}\pi_k\,\mu_k;\ S = \sum_{k=1}^{K}\pi_k\, S_k$$

2. **2.** Search $\{(W_i^*) \in \mathbb{R}^{Dxd}\}$ the linear transformation that maximizes the estimate of the L^2-PDM noted I_p presented in our previous work [6], using numerical optimization procedure "Trust Region". This solution, which depends on the Initialization $(W_{IDA})_i$ is obtained numerically according to this expression:

$$W^* = \boldsymbol{Arg\ \max_{W\in\mathbb{R}^{D\,x\,d}}}\left[\sum_{k=1}^{K}\left(\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}[\widetilde{K}_{(W_{IDA})_i,m_{nk}}(WV_i{}^k, WV_j{}^k)]\right.\right.$$
$$+\sum_{p=1}^{K}\sum_{i=1}^{n_p}\frac{1}{n_p}\left(\sum_{l=1}^{K}\sum_{j=1}^{n_l}\frac{1}{n_l}\widetilde{K}_{(W_{IDA})_i,\min(m_{np},m_{nl})}(WV_i{}^p, WV_j{}^l)\right)$$
$$\left.\left.-\,2\sum_{i=1}^{n_k}\sum_{l=1}^{K}\sum_{j=1}^{n_l}\frac{1}{n_l}\widetilde{K}_{(W_{IDA})_i,\min(m_{nk},m_{nl})}(WV_i{}^k, WV_j{}^l)\right)\right]$$

where $\{V_i{}^k\}_{i=1,\dots,n_k\,;\,k=1,..,K}$ denotes a supervised learning sample distributed according to the conditional random vector to the class k of dimension D, n_k is the size of the observation relative to the class k and K denotes the number of classes.

$\widetilde{K}_{m_N}(v, V_i)$ is the multidimensional kernel for a trigonometric system defined on $[-\pi, \pi]^d$:

$$\widetilde{K}_{W_0,m_N}(v, V_i) = \prod_{l=1}^{d}\frac{\sin\left[\left(\frac{2m_N+1}{2}\right)(x^l - X_i^l)\right]}{2\pi\sin\left[\frac{x^l - X_i^l}{2}\right]}$$

$v = \{x_l\}_{l\in[1-d]}$, $V_i = \{X_{i_l}\}_{l\in[1-d]}$

W_0 is the initialization transformation, used in the optimization algorithm.

3. **3.** For the d-reducers (W^*), the corresponding misclassification error $P_r(W^*, (W_{IDA})_i)$ is then calculated from a supervised test sample as following:

$$P_r(W_i^*, (W_{IDA})_i) = \frac{card\left\{Y'_j \neq Arg\left[\max_k\ \hat{n}_k\ \sum_{i=1}^{N} K\left(\frac{W_i^*V_j - W_i^*V_i}{h_N}\right) I_{[Y_i=k]}(W_i^*V_j)\right], j = 1,..,N\right\}}{N}$$

Thus, for this family of dimension reducers $\{W_i^*,\ i=1,..,\ n+N_{it}\}$ generated with different initializations of the numerical optimization procedure applied to the L^2-PDM, we will keep the one that minimizes the misclassification rate

$$W^* = arg\ \min_{i=1..n+N_{it}} P_r(W_i^*, (W_{IDA})_i)$$

2) Optimal classification in the sense of Bayes

The estimate of the misclassification error is then calculated by the optimal classifier in the sense of Bayes criterion. For this estimate, we used the method of modified kernel function parameter truncation adjusted h_N, defined by:

$$\widehat{f}_k(x) = \frac{1}{Nh_N^d}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h_N}\right) I_{[Y_i=k]}(x)$$

h_N is the smoothing parameter of the kernel method, K is a probability density and I is the indicator function of the event $[Y_i = k]$. N is the total number of supervised learning sample $\{(X_i, Y_i)\}_{i\in[1..N]}$, d the size of the reduced space.

The achievement of convergence quadratic mean requires that the following h_N tends to 0 and Nh_N^d tends to infinity when N tends to infinity. For the convergence, the condition is more restrictive that is made by the fact that $(Nh_N^d)^2$ must tend to infinity as N tends to infinity. An iterative algorithm called Plug in recently implemented to better adjust to the optimal in the sense of Mean Integrate Square Error MISE admits the following expression [4]:

$$h_N^* = [M(K)]^{\frac{1}{5}}\ [J(f_X)]^{-\frac{1}{5}}\,[N]^{-\frac{1}{5}}$$

where $M(K) = \int_{-\infty}^{+\infty} K^2(x)d$; $J(f_X) = \int_{-\infty}^{+\infty}(f_X''(x))^2 dx$

For high dimensions, the convergence of such estimator requires a very large sample size, in practice this size is often unrealistic. It is for this reason that we have gone through a dimension reduction. Once reduced space to 1-dimension enough so that the sample size is reasonable, the Plug in the method can be implemented. It consists of the following iterative algorithm [4]:

1. Calculate analytically or numerically M(K)
2. Initialize $J(f_X)$ in the Gaussian case
3. Calculate
   $h_N^* = [M(K)]^{\frac{1}{5}}\ [J(f_X)]^{-\frac{1}{5}}\,[N]^{-\frac{1}{5}}$
4. Evaluate f by the kernel method
5. Approach numerically the integral of the second derivative of f_X : $J(f_X)$
6. Return to Phase 3
7. Stop when the smoothing parameter is stabilized.

The use of these algorithms (figure 1) is not necessarily optimal solutions. So we cannot guarantee the realization of Bayesian classifier. Applications in face recognition will be described in the next section in order to illustrate the performance of this system in real datasets.

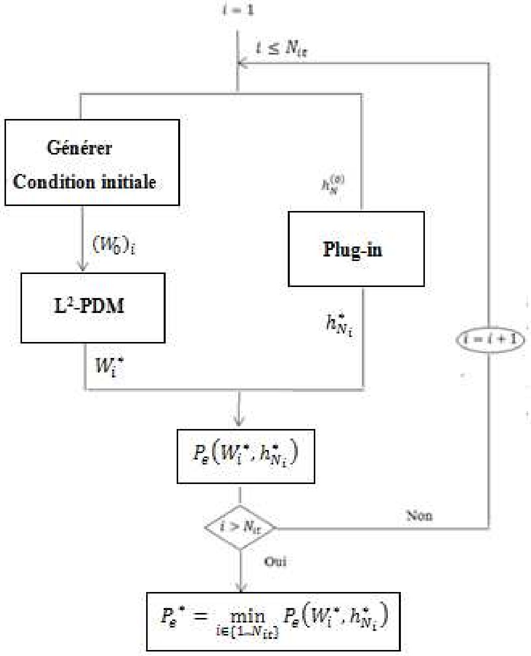

Figure 1- The global system for optimal Bayes classification

3 Experimental results in face recognition

Face recognition continues to be one of the most popular research areas of computer vision and machine learning. During the last decade, various methods of face recognition have been created and many of which are very efficient. However, the success of these methods depends largely on the quality of the results of the acquisition and detection of faces. That’s why, along with the development of face recognition algorithms, a comparatively large number of face databases have been collected. In our work we have used different databases that are publicly available and are demonstrated of use to others in the community like The ORL, Yale, Yale B, CMU and BIOID databases. In the following, we will review the databases used in our work.

- **The ORL** face database consists of 400 face images taken from 40 people, 10 images per person. For each person, it contains face images under different lighting conditions, facial expressions, and poses. All the images are against a dark homogeneous background with the subjects in an up-right, frontal position, with tolerance for some tilting and rotation of up to about 20 degrees. The images are in bitmap file format (bmp), grayscale with a resolution of 92x 112 pixels. There are variations in images of different persons like persons have beard, persons have glasses, persons have moustaches etc.
- **The Yale** Face Database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, and sleepy, surprised, and wink.
- **The Yale** B Database was collected to allow systematic testing of face recognition methods under large variations in illumination and pose. The subjects were imaged inside a geodesic dome with 64 computer-controlled xenon strobes. Images of 10 individuals were recorded under 64 lighting conditions in nine poses (one frontal, five poses at 12°, and three poses at 24° from the camera axis). Because all 64 images of a face in a particular pose were acquired within about 2 seconds, only minimal changes in head position and facial expression are visible.
- **The CMU PIE** database presents systematically samples a large number of pose and illumination conditions along with a variety of facial expressions. The PIE database contains 41,368 images obtained from 68 individuals. The subjects were imaged in the CMU 3D Room using a set of 13 synchronized high-quality cameras and 21 flashes. The resulting images are 640 × 480 in size

- **The BioID** Database contains 1521 frontal view images of 23 subjects. Images are 384 ×288 in size. Because the images were recorded at different locations, significant variations in illumination, background, and face size are present. The BioID Face Database images were manually marked up, 20 feature point were selected which are very useful for facial analysis and gesture recognition.

-

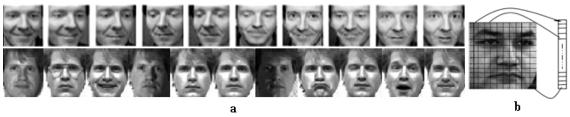

Figure 2 - a) Examples of YALE and ORL databases b) Holistic representation

Here we intend to present the results obtained after the dimensionality reduction process and the classification. The experimental setup used in this work is the following. For every data set used and for every possible d to reduce the dimension to, the experiment described below is conducted several times.

1. Using the holistic representation, the face descriptor in this face recognition technique is a vector that contains a lexicographic ordering of raw pixel values of each facial image in the database. Thus, for each 32 x 32 image, one obtains a 1024 dimensional feature vector per face (see fig2.b).

2. The evaluation consists of randomly divide the data set into ten non overlapping folds of equal size, then take seven folds for to compose the training set and three folds for the test set. This means that the training set compose 70% percent of the data and the test compose the other 30% .The transformation matrices W were estimated from the training data, which was then transformed to a subspace of appropriate dimension.

3. PCA is performed on the train set after which all principal components with an eigenvalue smaller than one millionth of the total variance, i.e., the trace of the total covariance matrix, are discarded. In this way, problems related to (near) singular covariance matrices are avoided and all four transformations can be properly determined.

4. Using the transformed train data, we determine the different LDR transformations and reduce the dimensionality of the train data to d.

5. In the d-dimensional reduced feature space, we determine the linear classifier, the nearest neighbor, the classical Bayes and the presented Bayes classifier (noted optimal Bayes) using the train data and, subsequently, classify the test data after transforming its instances in the same way as the train instances.

The classification error is estimated on the test data.

Many experiments were conducted on the different databases presented before in order to evaluate the different techniques and methods of facial feature extraction and classification mentioned in this work.

The per-data set performances of the several Linear Discriminate Reduction (LDR) techniques are compared. To this end, per classifier, data set and dimension d, the mean estimated classification error over the different runs is determined. This gives a final estimate of the classification error for the respective settings.

For every LDR transform, only the optimal dimensionality to reduce the data to and the corresponding mean classification error is reported. The different methods, give no direct means to determine an optimal dimensionality to reduce to. However, the observed optimal classification errors give an indication of the attainable performance and can be used to compare the several approaches. These numbers are presented in Tables 1, 2, 3, 4 and 5. The overall optimal classification error over all transforms is typeset in bold and a "*" is added in superscript. 'd*' represents the optimal dimension to reduce to, i.e. the dimension that gave us the minimum classification error. We have to mention that the transformed space of the LDA is constrained the number of classes in each database.

We start with two general observations: First, the ACC, the IDA and the L^2-PDM methods have outperformed the LDA in every database used. Second, the best performances were obtained by the optimal Bayes classifier combined with the L^2-PDM. We can notice also that the best classification rates are obtained by the CMU and the Yale B databases. It is obvious that these two data bases contain a large number of instances. Consequently many problems like the under sampled problems will be discarded.

The results taken from experiences on the Yale B database show how the classification error rate drop considerably as the number of feature extracted increases, specially using L^2-PDM, except with the Nearest Neighbour classifier that fails to give a good separation of the data.

It is interesting to mention that both the LDA and ACC methods were not able to perform the space reduction of the data when the initial dimension D is greater than the number of the training samples (the singularity problem). Taking into account the fact that both the LDA and ACC require the calculation of the covariance matrices, the singularity problems may occur. This problem of singularity of the covariance matrices does not figure in the case of IDA or the L^2-PDM. Concerning the IDA, this approach does not directly address the covariance matrices Σ_i , contrary; it is based on their representation in the transformed space. The major advantage of our system is that it is not limited by any assumption or condition, as the estimate of the L^2-PDM is done without any hypothesis on the type of the distribution of the data.

4 Conclusion

In this article, we have introduced an optimal classifier based on our previous reduction dimension approach, to approximate the Bayes classifier in a non-parametric supervised case, using the modified kernel estimator within the adjustment of the smoothing parameter in the sense of Mean Integrated Squared Error. In order to evaluate the performance of this system, we have detailed and explained the different steps of face recognition. The performance of the previously discussed classification methods were evaluated using several facial databases.

Table 1 Face Databases classification results

ORL	**LDA**	**d***	**IDA**	**d***	**L^2-PDM**	**d***
Linear classifier	0.8500	26	0.1459	210	0. 1447*	190
Nearest Neighbour classifier	0.9583	4	0.1351	215	0.1253*	210
Bayes classifier	0.8500	28	0.1295*	205	0.1295*	200
Optimal Bayes classifier	0.8500	28	0.1227 *	185	0.1227 *	180
YALE	**LDA**	**d***	**IDA**	**d***	**L^2- PDM**	**d***
Linear classifier	0.8444	2	0.1790*	105	0.1790*	100
Nearest Neighbour classifier	0.8667	2	0.2123	90	0.1908*	80
Bayes classifier	0.8444	2	0.1778*	100	0.1778*	110
Optimal Bayes classifier	0.8444	2	0.1737	115	0.1712*	95
YALE B	**LDA**	**d***	**IDA**	**d***	**L^2- PDM**	**d***
Linear classifier	0.3089	37	0.1070*	115	0.1080	100
Nearest Neighbour classifier	0.6003	16	0.1099	100	0.1007*	95
Bayes classifier	0.3236	37	0.1003*	110	0.1003*	120
Optimal Bayes classifier	0.3089	37	0.1057	125	0.1002*	110
CMU PIE	**LDA**	**d***	**IDA**	**d***	**L^2- PDM**	**d***
Linear classifier	0.1665	37	0.0852*	110	0.0870	100
Nearest Neighbour classifier	0.3255	37	0.1109	95	0.1051*	95
Bayes classifier	0.1137	37	0.0932*	120	0.0932*	120
Optimal Bayes classifier	0.1579	37	0.0803	120	0.0800*	110
BIOID	**LDA**	**d***	**IDA**	**d***	**L^2- PDM**	**d***
Linear classifier	0.4743	20	0.2398	190	0.2125*	200
Nearest Neighbour classifier	0.5342	19	0.2654*	180	0.2654*	195
Bayes classifier	0.4312	21	0.2864	200	0.2500*	205
Optimal Bayes classifier	0.4943	18	0.2706	180	0.2111*	210

References

[1] E.A. Patrick and F.P. Fisher, “Nonparametric feature selection”, *IEEE Trans. On Inf. Theory*, vol. IT-15, pp.577-584, 1969.

[2] Loog, M. et al. (2001). Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria, *IEEE trans. on PAMI.*, Vol. 23 N°7.

[3] R. A, Fisher. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol. 7, 179-188. 1936.

[4] S. Saoudi, et al. An Iterative Soft Bit Error Rate Estimation of Any Digital Communication Systems Using a Nonparametric Probability Density Function. EURASIP Journal on Wireless Communications and Networking, vol. 2009, 2009.

[5] W. Drira, et F. Ghorbel Un estimateur de la L^2 mesure de dépendance probabiliste pour la réduction de dimension vectorielle pour le multi classes, Traitement du Signal TS 2012.

[6] W. Drira, W. Neji et F. Ghorbel. Dimension reduction by an orthogonal series estimate of the probabilistic dependence measure. ICPRAM -International Conference on Pattern Recognition Applications and Methods, pp. 314-317, 2012

[7] Z. Nenadic, Information Discriminate Analysis: Feature Extraction with an Information-Theoric Objective, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol 29 N° 8, August 2007.