Enhancing Face Recognition from Video Sequences using Robust Statistics

Sid-Ahmed Berrani

France Telecom R&D – TECH/IRIS
4, rue du Clos Courtel – BP 91226
35512 Cesson Sévigné Cedex. France
SidAhmed.Berrani@francetelecom.com

Christophe Garcia

France Telecom R&D – TECH/IRIS
4, rue du Clos Courtel – BP 91226
35512 Cesson Sévigné Cedex. France
Christophe.Garcia@francetelecom.com

Abstract

The aim of this work is to investigate how to enhance the performance of face recognition from video sequences by selecting only well-framed face images from those extracted from video sequences. Noisy face images (e.g. not well-centered, non-frontal poses...) significantly reduce the performance of face recognition methods, and therefore, need to be filtered out during the training and the recognition. The proposed method is based on robust statistics, and in particular, a recently proposed robust high-dimensional data analysis method, RobPCA. Experiments show that this filtering procedure improves the recognition rate by 10 to 20%.

1 Introduction

Face recognition is an important research subject in the pattern recognition field that has been extensively investigated in the last couple of years. It can be applied in different domains (biometrics, video surveillance, multimedia indexing, etc.).

The basic objective of a face recognition system is to identify an unknown face from a picture or a video sequence using a stored database of faces. While face recognition from still images has been very deeply investigated and many techniques have been proposed [13], face recognition from video sequences is just emerging. The fact that a person has to be learned and also recognized from a video sequence containing her/his face strongly modifies the deal, and introduces new elements that make the direct use of techniques designed for still images very difficult.

Video sequences are in general of a poor quality compared to still images. The resolution of video frames is low and they might contain very important illumination variations. In addition, face sub-images need to be automatically extracted from the frames of the video. Therefore, a face detection method is generally used. Due to the shortcomings of face detection methods, extracted face images are generally imprecisely cropped, may contain false detections and important pose variations. All these subsequent difficulties make the problem much more difficult than the case of recognition from still images. A video sequence provides, however, rich and redundant information that can be exploited to counterbalance these difficulties and enhance the performance of the recognition.

The recently proposed solutions dealing with face recognition from video sequences can be divided into two classes. The first one considers the problem as a recognition problem from still images by *independently* using all or a subset of the extracted face images. In general, the majority voting is used to come up with a final result. In the case where only a subset of faces is considered, ad hoc heuristics are used to select faces. It could be as simple as a random sampling, a selection of faces that are extracted from key frames or based on more sophisticated heuristics. In [5], an unsupervised learning method is proposed to extract the most representative faces (called “exemplars”): A non-linear dimensionality reduction is applied and followed by a K-Means clustering. Exemplars are chosen as the set of cluster centers. In [1], a principal component analysis (PCA) projection space is computed per person. The reconstruction error is then used as a decision criterion, that is, a query face is assigned to the person whose PCA projection space induces the least important reconstruction error of the feature vector associated with the query face. When the query is a video sequence, all of the extracted faces are used and the final decision is made using the face whose reconstruction error is the median value of the reconstruction errors.

The main drawback of these techniques lies in the fact that all the extracted faces are used. Therefore, noisy faces (not well centered, taken under poor illumination conditions, false detection, non-frontal faces...) are included in the training set and also used for recognition, which significantly decreases the recognition rate. In the case where a selection is performed (e. g. [5]), it does not take the quality of the faces into account. This also reduces the recognition rate as a decision can be made using noisy faces.

The second class makes use of all of the extracted faces together with or without considering their temporal order in the video. The idea is to integrate all of the information expressed by all of the face images into a single model (or a single descriptor). In particular, when the temporal order of the face images is taken into account, such a model is supposed to include information on the face dynamics. In [3], the ordered set of extracted faces are projected into the eigenspace and represented as a trajectory in that space. Recognition is then performed as a trajectory matching. In [12], a PCA space is computed for the extracted faces from each sequence. The similarity evaluation between two sequences is performed using the angle between the corresponding subspaces (using the mutual subspace method). In [9], the problem is considered from an information theory point of view. The proposed method classifies a set of face images using the Kullback-Leibler divergence between the estimated density of extracted faces and each of the densities associated with stored people (densities are assumed to be Gaussian). In [7], Liu *et al.* use adaptive Hidden Markov Models (HMM) to learn temporal statistics and the dynamics of the faces. One HMM is learned for each person from the associated video sequence. In the presence of a query sequence, it is analyzed by each HMM. The recognition is then performed using a maximum *a posteriori* rule.

In addition to specific limitations of each technique, once again, the main drawback of this second class of techniques is the use of all of the faces, including those that are noisy. These generate outlying feature vectors that, when used for training and/or recognition, significantly reduce the performance of recognition. As will be explained later, the reason of this phenomenon lies in the fact that all the introduced techniques are mainly based on high-dimensional data analysis methods like PCA or models like HMM. They are therefore very sensitive to outliers as they are based on the computation and the analysis of the first and the second order moments (the mean and the covariance matrix).

In this paper, we are interested in this problem and we propose a method for selecting only images that are of good quality from the extracted ones. The idea is to take advantage of the abundance of face images in this context to filter out those responsible for performance degradation. The adopted strategy is to consider the problem as an outlier detection problem and make use of robust statistics to perform filtering. In particular, we propose to use RobPCA, a recently proposed method by Hubert *et al.* [6].

We will illustrate the introduced concepts and show the effectiveness of our approach using two face databases and two methods for face recognition (PCA [10] and LDA2D [11]).

The rest of the paper is organized as follows. Section 2 analyzes the impact of outliers on high-dimensional data analysis methods in general. Section 3 introduces the proposed approach and presents our method for automatically selecting well-framed face images. It also explains how the proposed selection procedure can be included in a whole system for face recognition from video sequences. Experimental results are presented in section 4. Section 5 concludes the paper and discusses our future works.

2 On the impact of outliers

Face recognition methods being generally based on statistical methods, are therefore very sensitive to outliers. To analyze this problem, we will focus in a first instance on the well-known method of eigenfaces [10] in the context of face recognition from still images. The principle of this method is to make use of the statistical properties of feature vectors associated with face images to compute a projection space. Face images are projected in this space and their similarity is evaluated simply as an Euclidean distance. The identification of an unknown face is achieved by finding the face in the database whose projection vector is the closest to the projection vector of the unknown face (nearest-neighbor classification). In this method, the feature vector is just the vector obtained by concatenating the rows or the columns of the face image. The projection space is computed using PCA. It is defined by the most “important” eigenvectors of the covariance matrix of the feature vectors. These eigenvectors encode most of the variations between face images, that is, the discriminant information that allows matching or differentiating faces.

Ideally, the encoded variations should only be those related to faces (e. g. shapes of the face, the eyes, the nose and the mouth) and not those inherent to the acquisition conditions or to the expression changes. Unfortunately, this is very difficult in practice and it is extremely difficult to get rid of external variations. Therefore, we consider that a face image is *noisy* if it presents external variations that the face recognition method cannot be invariant to. We classify these variations into three categories:

1. Important illumination changes,
2. Imprecise face cropping,
3. Non-frontal poses.

A sample of face images presenting these kinds of variations is shown in figure 1. Faces from the second and third category arise in general from an automatic extraction of faces from still images or video sequences (imprecision inherent to face detection methods).

To illustrate the impact of these variations on face recognition, we have evaluated the recognition rate using the method of eigenfaces (PCA) on two databases: PF01 to study the impact of important illumination changes and

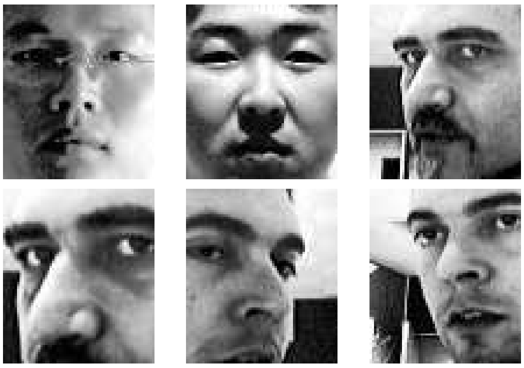

Figure 1: Example of face images considered as outliers.

FDB15 for imprecise face cropping and non-frontal poses[1]. The recognition rates have been evaluated twice: The first time using all of the images, and the second time after having removed the noisy images of each person from the training and the test sets. The obtained results clearly showed that the recognition rate increases by about 10 % when noisy faces are removed from the train and the test sets.

On the other hand, if we study the three first eigenvalues from PCA, we notice that their proportion is much more important when noisy faces are included during the analysis. We recall that the proportion of an eigenvalue w.r.t. the sum of all eigenvalues corresponds to the proportion of the variation expressed by the corresponding principal component.

Table 1 gives the cumulated proportion of the first three eigenvalues w.r.t. the sum of all eigenvalues for PF01 and FDB15. This table corroborates the analysis presented in the beginning of this section. It clearly shows that despite their relatively small number (2 among 11 face images per person for PF01, and 6 among 21 face images per person for FDB15), noisy faces in the training sets significantly modify the PCA, and hence decrease its effectiveness as they affect the discriminant information encoded by the first principal components.

	PF01	FDB15
Using all the faces	35.51	35.02
Without noisy faces	25.68	29.13

Table 1: Cumulated proportion of the first three eigenvalues w.r.t. the sum of all eigenvalues.

As for the context of face recognition from video sequences, this analysis applies in a straightforward manner to all of the PCA-based techniques (e.g. [1], [3], [12]). For the other methods that use HMM or Kullback-Leibler divergence, the impact of outliers could be easily analyzed in the same manner. These methods rely strongly on the estimation of Gaussian probability densities. They therefore require the computing of covariance matrices.

3 The proposed solution

In the previous section, we analyzed the problem of outliers in general and we showed that a small subset of noisy faces significantly reduces the performance of PCA-based face recognition methods. A good solution to get rid of this problem, whatever the used method of face recognition, is to filter out the noisy face images.

In this section, we first present the method we propose to filter out noisy faces. In the second part of this section, we present the application of the filtering procedure to face recognition from video sequences.

3.1 Outlier detection using RobPCA

Lets consider a set of n face images $I_1, \dots, I_n$ that might have been extracted from a video sequence. $x_1, \dots, x_n$ are the associated d-dimensional feature vectors. These vectors are arranged together in a single matrix $X_{n,d}$.

To isolate noisy face images, we consider the problem from a statistical point of view: A face is filtered out if its feature vector is an outlier. To achieve that, we propose the use of the RobPCA method introduced by Hubert *et al.* [6]. This method has been proposed to perform a robust principal component, i.e. finding principal components that are not influenced too much by outliers. It also provides a useful method to flag outliers.

RobPCA combines the ideas of two different approaches for robust estimation of principal components. The first approach aims at finding a subset of vectors whose covariance matrix has the smallest determinant, that is, the most compact in the space. The mean and the covariance matrix is computed on this subset. The second approach uses Projection Pursuit techniques. The idea is to maximize a robust measure of spread to sequentially find the principal axes.

Robust Mean and Covariance Matrix Estimation

To estimate the robust mean ($\hat{\mu}$) and the robust covariance matrix ($\hat{C}$) of a dataset $X_{n,d}$, the RobPCA proceeds in three steps:

1. Data vectors are processed using a classical PCA. The objective is not to reduce the dimension but only to remove superfluous dimensions.

2. The h "least outlying" vectors are searched, where $h < n$ and $h - n$ is the maximum expected number of outliers. To do that, a measure of "outlyingness" is used. This measure is computed by projecting all of

[1] PF01 is available following the URL: http://nova.postech.ac.kr/. FDB15 is a database we have created specially for the needs of this paper. These two databases are further described in section 4.

the vectors on a set of lines and by measuring the degree of "outlyingness" of each vector w.r.t. the spread of projections. A PCA is then performed on the found h vectors and the dimension is reduced.

3. The final $\hat{\mu}$ and $\hat{C}$ are estimated using an MCD estimator, i.e. based on the h vectors whose covariance matrix has the smallest determinant. To find these vectors, a FAST-MCD algorithm [8] is used. The principle of FAST-MCD is to draw a set of random subsets and to refine them iteratively:

   - Compute the mean (m) and covariance matrix (C) of the h vectors,
   - Compute the C-Mahalanobis distances of all the vectors to m,
   - Choose a new set composed of the h vectors with the smallest Mahalanobis distances. The determinant of the covariance matrix of these new h vectors is smaller than the determinant of C.

   This procedure is repeated until convergence, i.e. no further improvements are obtained.

Outliers Detection

Once $\hat{\mu}$ and $\hat{C}$ have been estimated, the vectors are projected into a lower dimensional space defined by the eigenvector of $\hat{C}$. Let $Y_{n,k}$ be the new data matrix:

$$Y_{n,k} = (X_{n,d} - 1_d\, \hat{\mu}^t) P_{d,k}, \tag{1}$$

where 1_d is d-dimensional vector of all components equal to 1 and $P_{d,k}$ is the projection matrix. $P_{d,k}$ is computed from a spectral decomposition of $\hat{C}$:

$$\hat{C} = P_{d,k}\, L_{k,k}\, P_{d,k}^t, \tag{2}$$

where $L_{k,k}$ is the diagonal matrix of eigenvalues $l_1, ..., l_k$.

The outliers are then determined by analyzing the distribution of the two following distances (computed for the vector i):

$$D1_i = \sqrt{\sum_{j=1}^{k} \frac{y_{ij}^2}{l_j}}, \tag{3}$$

$$\text{and} \qquad D2_i = ||x_i - \hat{\mu} - P_{d,k}\, y_i^t||. \tag{4}$$

The first distance is the distance to the robust center of the vectors. It evaluates the proximity of x_i to the vectors cloud in the projection space; whereas the second distance is the orthogonal distance to the projection space. Two thresholds are then derived from the distribution of these distances. If a vector has at least one of the two distances greater than the associated threshold then it is considered an outlier.

The distribution of D1 can be approximated by a χ_k^2 distribution because it is a Mahalanobis distance of normal vectors. Therefore, the associated threshold can merely be for example $\sqrt{\chi_{k,0.975}^2}$. The distribution of D2 however is not exactly known. Therefore, we use the approximation proposed in [6], i.e. D2$^{3/2}$ is approximately normally distributed. The associated threshold is hence $(m + \sigma z_{0.975})^{3/2}$, where m and σ are respectively the robust estimations of the mean and the standard deviation and $z_{0.975}$ is the 97.50% quantile of the normal distribution.

3.2 Application to face recognition

Filtering out noisy faces is needed twice in the whole recognition process: (1) off-line, to filter out face images of the training set, and (2) on-line, to decide whether to take a query face in the recognition process into account or not. We recall that this filtering procedure is applied independently of the used face recognition technique.

From the training video sequences, the face images are first extracted. In order to filter noisy faces, all that is needed is to transform face images into vectors and apply the outlier detection procedure described in the section 3.1. This can be done simply by concatenating the rows or the columns of each face image. However, as we are not concerned with all of the details of the images at this stage, we suggest reducing the resolution of the images. This allows the acceleration of the filtering process and avoids taking small details of the images into account.

This filtering procedure is applied per person if there are enough face images per person. It can also consider all of the faces in the training dataset simultaneously. The only important condition is to make sure that the set of images contains a majority of well-framed face images. This condition is inherent to the principle of our approach which considers noisy faces as outliers, and therefore makes the assumption that noisy faces are in the minority. This condition is however not very restricting because the acquisition of the training video sequences could be controlled.

For recognition, the situation is different. We do not have any control over the quality and the number of query faces. In particular, we cannot impose any condition concerning the proportion of well-framed face images. The proposed solution is as follows:

- Extract query face images from the query sequence,
- Insert all or a subset of the query faces into a set of well-framed face images from the training set,
- Apply the filtering procedure on that set and keep only query faces that have not been isolated as outliers.

The only condition that needs to be fulfilled is to have a majority of well-framed images in the set in which query faces are inserted.

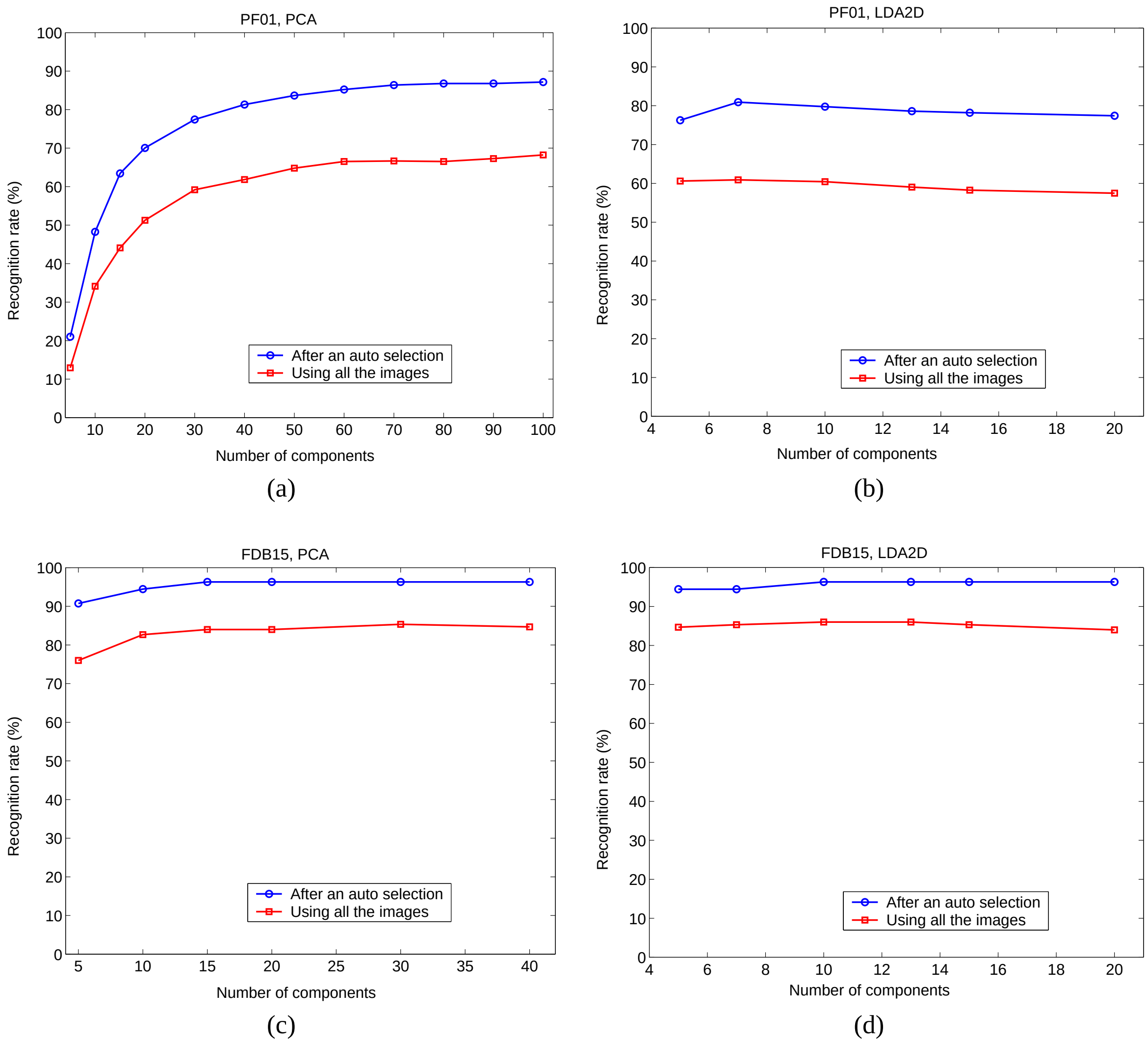

Figure 2: Recognition capabilities of PCA and LDA2D with and without an automatic selection of face images.

4 Experimental results

To assess the effectiveness of the proposed filtering procedure, we have performed experiments on two datasets and we have considered two face recognition methods.

Face Image Databses

1. The Asian Face Image Database PF01[2] contains 17 different views of each each person: 1 normal face, 4 illumination variations, 8 pose variations, 4 expression variations. We have used 11 images for training and 6 images for testing per person.

2. FDB15 is a database of 15 people. It has been created by automatically extracting faces from video sequences using the CFF detector [4]. In the training set, we have used 15 well-framed images and 6 noisy faces per person. In the test set, we have used 5 well-framed images and 5 noisy faces per person. The variations addressed by this database concern the position of the face in the image and the head pose.

The size of the face images of the two databases has been set to 65×75. It has been however reduced to 32×37 during the filtering step.

Face Recognition Techniques

In this evaluation, we have used two face recognition techniques, PCA (eigenfaces [10]) and LDA2D [11]. We have performed the recognition considering the problem as a recognition problem from still face images as there are no standard databases for face recognition from video se-

[2] Available following the URL: http://nova.postech.ac.kr/

quences. However, this has absolutely no affect on the conclusions of this evaluation. The objective of this study is only to show how our filtering procedure can significantly enhance the performance of face recognition techniques.

The eigenfaces method has been previously introduced in section 2. LDA2D is a recently proposed method that has the same principle as the well-known LDA method (referred to as Fisherfaces in [2]). LDA aims to define a projection space into which the classes are as compact as possible and their centers are the most far away from each other (a class refers here to the set of feature vectors extracted from the face images of a person). The LDA is carried out via scatter matrix analysis. LDA2D considers the faces images as matrices (and not as vectors). The advantages are an important gain in storage, a better numerical stability and an increased recognition rate.

Results

The filtering procedure of noisy faces has been carried out for the training sets on each person separately. The selection rates of face images is presented in table 2 for the training set and also for the test set.

	PF01	FDB15
Training set	65.25	63.81
Test set	40.03	36.00

Table 2: The selection rates of face images (%).

To assess the effectiveness of this procedure, we have evaluated the recognition rates using the databases without noisy faces that have been automatically filtered out by our method. We have then compared them to the recognition rates obtained without any filtering, i.e. using the whole images in the databases.

The obtained results are summarized in figure 2. We can notice that, overall, the recognition rates have been improved by the filtering procedure by 10 to 20%.

5 Conclusions and Future Work

In this paper, we have analyzed the problem of sensitivity to outliers of face recognition methods and we have proposed a filtering method to isolate noisy faces that are responsible for performance degradation.

The experimental study has clearly showed the effectiveness of the proposed filtering method on two face recognition methods (PCA and LDA2D). These two methods are at the root of the first class of methods presented in the introduction. In our future work, we will study the outlier problem with the second class of methods, i.e. the methods that use all of the face images to build a single model. Even if we would expect the same conclusions (these methods are mainly based on density estimation, covariance matrices, etc.), other phenomena inherent to the included temporal information might appear and might need to be deeply studied.

References

[1] E. Acosta, L. Torres, A. Albiol, and E. Delp. An automatic face detection and recognition system for video indexing applications. In *Proc. of the* IEEE *Int. Conf. on Acoustics, Speech, and Signal Processing (IV), Orlando, Florida, USA*, pages 3644–3647, May 2002.

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

[3] Z. Biuk and S. Loncaric. Face recognition from multi-pose image sequence. In *Proc. of the 2nd Int. Symp. on Image and Signal Processing and Analysis, Pula, Croatia*, June 2001.

[4] C. Garcia and M. Delakis. Convolutional face finder: A neural architecture for fast and robust face detection. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, 26(11):1408 –1423, November 2004.

[5] A. Hadid and M. Pietikainen. From still image to video-based face recognition: An experimental analysis. In *Proc. of the 6th* IEEE *Int. Conf. on Automatic Face and Gesture Recognition, Seoul, Korea*, pages 813 – 818, May 2004.

[6] M. Hubert, P. Rousseeuw, and K. V. Branden. Robpca: A new approach to robust principal component analysis. *Technometrics*, 1(47):64 – 79, February 2005.

[7] X. Liu and T. Cheng. Video-based face recognition using adaptive hidden markov models. In *Proc. of the Conf. on Computer Vision and Pattern Recognition (I), Madison, Wisconsin, USA*, pages 340 – 345, June 2003. RecoFaceVideo.

[8] P. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212 – 223, 1999.

[9] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. of the 7th European Conf. on Computer Vision, Copenhagen, Danemark*, pages 851–868, May 2002.

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

[11] M. Visani, C. Garcia, and J.-M. Jolion. Two-dimensional-oriented linear discriminant analysis for face recognition. In *Proc. of the Int. Conf. on Computer Vision and Graphics, Varsovie, Pologne*, November 2004.

[12] O. Yamaguchi, K. Fukui, and K. ichi Maeda. Face recognition using temporal image sequence. In *Proc. of the 3rd Int. Conf. on Face and Gesture Recognition, Nara, Japan*, pages 318 – 323, April 1998.

[13] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. ACM *Computing Surveys*, 35(4):399 – 458, December 2003.