Using Random Sampling to Enhance Linear Discriminant Analysis for Video-based Face Recognition

Weiwu Jiang, Zhifeng Li, *Member, IEEE* and Helen Meng

***Abstract*—Video-based face recognition has attracted considerable research interests in recent years. The major advantage of video-based face recognition, in comparison with facial image recognition, is that more information is available in a video sequence than in a still image. However, such an advantage comes at a high cost because video data often involve many dimensions, which may lead to problems such as high processing complexity, numeral instability and training data sparsity which lead to overfitting. In order to address these problems, we propose a novel algorithm using random sampling to improve linear discriminant analysis (LDA) for video-based face recognition. Random sampling is applied on the training set, as well as the feature space. In this way we can train multiple stable LDA classifiers, whose outputs are combined to produce the final classification output. Significant performance improvement on face recognition is achieved based on the XM2VTS face video database.**

***Index Terms*—Video based face recognition, random sampling, classifier design and fusion**

I. Introduction

VEDIO-BASED face recognition has attracted considerable research interests in recent years. One major driver for this trend is the availability of high-quality video acquisition devices and techniques. Another major driver is the significant advantage offered by video-based face recognition over facial image recognition – while it is possible to penetrate a system fraudulently with a pre-recorded facial image; it is much more difficult to forge a video sequence before a live video camera. Furthermore, the video sequence contains much more information about the subject than a single image. Proper utilization of such additional information may heighten face recognition performance.

A major research challenge presented by video-based face recognition is that video data often involves many dimensions, leading to "the curse of dimensionality". This creates several serious problems, such as high processing complexity, numeral instability and training data sparsity which lead to overfitting. One approach to counteract these problems is to extract a compact set of features for data description, so that classification can be conducted in a space with significantly reduced dimensions. This should enhance classification efficiency and robustness. Hence, subspace techniques have been popular in previous work. Two main methods for subspace analysis include principal component analysis (PCA) and linear discriminant analysis (LDA). The PCA method is also known as the Eigenface method [1]. It uses the Karhunen-Loeve Transform (KLT) to produce a most expressive subspace for face representation and recognition. However, the goal of PCA is compression and does not target recognition [2]. In contrast, the LDA method, also known as the Fisherface method [3], aims to pursue the discriminant subspace that maximizes class separability. LDA-based subspace methods offer simplicity in computation and effectiveness in classification. However, when dealing with high-dimensional video face data, the LDA method encounters two main problems: First, the large number of feature dimensions renders the application of subspace analysis very costly. Second, high-dimensional feature vectors require a large amount of training data and consequently overfitting often occurs to due insufficient training data. In order to address these problems, we investigate the use of random sampling techniques for video-based face recognition. Two popular random sampling techniques are random subspace and bagging. In the random subspace method [4], multiple classifiers are generated by random sampling the feature space. The decisions from these classifiers are then combined by a final decision rule to generate the final decision to strive for improved classification performance. In the bagging method [5], multiple training data subsets are generated by random sampling the training set. A classifier is then constructed from each training data subset, and the results of all the classifiers are finally integrated. In video based face recognition, the problem of "the curse of dimensionality" becomes more

Manuscript received October 9, 2001. (Write the date on which you submitted your paper for review.) This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

Weiwu Jiang, Zhifeng Li and Helen Meng are with the System Engineer and Engineer Management Department, the Chinese University of Hong Kong, Hong Kong, (corresponding author to provide phone: 852-68415057; fax: 852-26035505; e-mail: {wwjiang, zfli, hmmeng}@ se.cuhk.edu.hk).

serious than image based face recognition. In order to better address this problem, our proposed approach utilizes both methods of random subspace and bagging for video-based face recognition – we randomly sample the feature space as well as the training set. Multiple classifiers are constructed and then combined to generate the final decision. We investigated the effectiveness of this method by experimentation with the largest standard XM2VTS video face database [6].

II. THE XM2VTS DATABASE

Please The XM2VTS database is a multi-modal face database project, which was collected at University of Surrey within the M2VTS (Multi-Modal Verification for Teleservices and Security Applications) project. This large multi-modal database has four sessions on 295 subjects over a four month period captured by high quality digital video. Each recording contains a speaking head shot and a rotating head shot. Sets of data taken from this database include high quality color images, 32 KHz 16-bit sound files, video sequences and a 3d Model. The persons in the video read two numeric sequences, “0 1 2 3 4 5 6 7 8 9” and “5 0 6 9 2 8 1 3 7 4”.

III. USING RANDOM SAMPLING TO ENHANCE LINEAR DISCRIMINANT ANALYSIS FOR VIDEO BASED FACE RECOGNITION

A. *High-dimensional Feature Vector*

Our approach begins by extracting a set of key video frames from each video sequence by means of the spatio-temporal synchronization method, as developed in our previous work [7]. The spatio-temporal synchronization method uses the waveform of the audio signal to allocate desired frames in each video for further analysis. The objective of this key frame extraction procedure is to locate a small set of distinct video frames to represent the characteristics of the video sequence. In our experiments, 21 key video frames are extracted from each video sequence using the spatio-temporal synchronization method. These key frames need to be combined for subsequent analyses and classification of the video. A straightforward approach is to append the key video frames into a single large vector, and then conduct regular subspace analysis for feature extraction. Although this approach of feature level fusion utilizes all the data in video, there is an overly large number of feature dimensions (21images×41×27pixels = 23,247dimensions). The high dimensionality leads to costly computations and overfitting problems. These issues are common in facial image recognition, but are vastly aggravated in video-based face recognition.

B. *Linear Discriminant Analysis (LDA)*

LDA is a popular subspace face recognition technique. It uses the within-class scatter matrix and the between-class scatter matrix to measure the class separability. They are defined as,

$$S_w = \sum_{i=1}^{c} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^T \qquad (1)$$

$$S_b = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T \qquad (2)$$

where μ_i denotes the mean of the class C_i, μ denotes the mean of all classes, *c* denotes the number of classes and N_i denotes the number of samples in class C_i.

LDA uses the optimal projections W_{opt}, which maximizes the ratio of the determinant of between-class matrix to that of the within-class matrix, defined as:

$$\mathrm{W}_{\mathrm{opt}} = [\mathrm{w}_1, \mathrm{w}_2, \ldots \mathrm{w}_\mathrm{f}] = \mathrm{argmax} \frac{\left\| \mathrm{W}^\mathrm{T} \mathrm{S}_\mathrm{b} \mathrm{W} \right\|}{\left\| \mathrm{W}^\mathrm{T} \mathrm{S}_\mathrm{w} \mathrm{W} \right\|} \qquad (2)$$

Mathematically it is equivalent to the leading eigenvectors of $S_w^{-1} S_b$. High dimensionality in the input feature vector space, where the training data will become relatively sparse, affects the computations of S_w and S_b. More specifically, it becomes difficult to accurately estimate S_w. A slight disturbance of noise will greatly change the inverse of S_w. Furthermore, S_b focuses primarily on the class centers and the boundary structure cannot be captured effectively from sparse data. In principal, any sample pairs from different classes can become candidates for estimating S_b. These factors will significantly affect the performance of LDA [6].

In order to alleviate these problems, we use two random sampling methods, random subspace and bagging to enhance the LDA. First, we use the random subspace method to randomly sample the feature space, in order to reduce the feature length with high dimensionality. Second, to address the problem caused by insufficient utilization of training set, we select specific sample pairs from different classes to better estimate the discriminant subspace, in order to better estimate S_b. It has been shown that sample pairs near the boundary contain more useful information and thus deserve greater emphasis. This favors the use of inter-class sample pairs (sample pairs from different classes) with smaller distances. More specifically, we use the bagging technique to randomly sample the ensemble of all inter-class sample pairs to generate multiple inter-class sample pair subsets for constructing multiple between-class scatter matrices S_b. We combine two random sampling technologies by developing an LDA classifier for each subspace and with each between class-matrix S_b derived from bagging. The entire framework with random sampling for classification is depicted in Fig. 1.

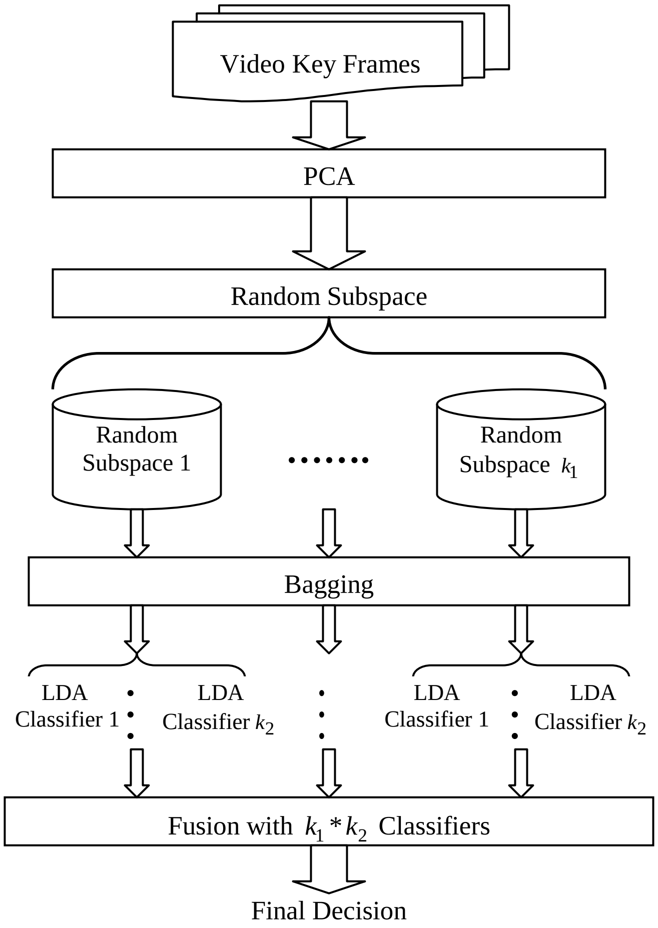

Fig. 1. Block diagram of the classification framework based on random sampling

Import The detailed algorithm is described in the following.

Training procedures:

1. Apply PCA to the training set of the video feature vectors. Keep all eigenvectors with non-zero eigenvalues as candidates to construct the random subspaces. These retained eigenvectors are denoted by $E = \{e_1, e_2, ..., e_{M-1}\}$, where M is number of training samples.
2. Construct k_1 random subspaces $\{S_i\}_{i=1}^{K_1}$ with each spanned by $D_1 + D_2$ dimensions. The first D_1 dimensions are fixed according to the first D_1 eigenvectors with the largest eigenvalues in E, which can preserve the main intra-personal variations. The remaining D_2 dimensions are randomly selected from the remaining eigenvectors in E.
3. In each random subspace, construct k_2 different between-class scatter matrix. Each between-class scatter matrices is composed of two components.

$$S_b^n = S_b + \sum_{j=1}^{p} (x_{j1} - x_{j2})(x_{j1} - x_{j2})^T, \qquad (4)$$

   where $n = 1, 2, \ldots, k_2$. The first term is the standard between-class scatter matrix as defined in Eq. (2) and is used to preserve the main inter-personal variations. The second term is calculated from the p inter-class pairs which are randomly selected from L inter-class pairs with smallest distances among all inter-class classes ($L >> p$), where (x_{j1}, x_{j2}) is the j-th selected inter-class pair from the subset of the p inter-class pairs.
4. A LDA classifier is then constructed based on each between-class scatter matrix. Hence in each subspace we generate k_2 LDA classifier with k_1 subspace. We generate $k = k_1 * k_2$ LDA classifiers in total.

Testing procedures:

The testing video data is fed to the k subspace classifiers in parallel, and the outputs are combined using a fusion scheme to make the final decision. Fusion involves either majority voting or the sum rule [4][11].

IV. EXPERIMENTS

We perform extensive experiments on the XM2VTS face video database [6]. For the training data, we use the 295*3 video sequences from the first three sessions. The test data is composed of a gallery set and a probe set. The gallery set is composed of the 295 video sequences of the first session, and the probe set is composed of the 295 video sequences of the last session. As described earlier, 21 key frames are selected from each video be the means of the spatio-temporal synchronization technique. The frame images are then normalized through the following steps: (1) rotate the face images to align with a vertical face orientation; (2) scale the face images so that the distances between the two eyes are the same for all images; (3) crop the face images to remove the background and the hair region; (4) apply histogram equalization to the face images for photometric normalization. The normalized images are fed into proposed random sampling classification framework. The framework parameters are selected as: $k_1 = 5$, $k_2 = 5$, L = 1000, p = 200, $D_1 = 50$, $D_2 = 350$.

We begin with comparing the proposed approach with conventional subspace methods, namely, Eigenface [1] and Fisherface [3]. Here all approaches directly use image gray scale values as facial features. Comparative results are shown in Table 1. As discussed above, when data is of high dimension, a single classifier constructed on the limited training samples is unstable. Traditional subspace methods suffer from training data sparsely and experiment results clearly verify this point. In the face of high dimensionality in feature space, our random subspace analysis framework greatly improves the recognition accuracy comparing to the conventional subspace methods.

We also compare the proposed approach with existing subspace models for video-based face recognition methods. They include:

(1) The nearest frame method [10], which matches two video sequences by selecting the pair of frames that are closest across the two videos.

(2) The mutual subspace method [10][11], which uses the video frames for each person separately to compute many individual eigenspaces for recognition.

(3) The multi-level subspace analysis method [7], which takes advantage of the spatio-temporal information in the video sequence to extract a set of key video frames and then conduct a two-level subspace analysis to extract the discrminant subspace features for recognition.

Results in Table 2 indicate that the proposed method gave superior performance over other the existing subspace methods for video based recognition. Comparing to the best result of the previous methods, the error rate is reduced by 50%.

V. Conclusion

In this paper, we propose an effective algorithm for video-based face recognition. In order to overcome the problems caused by the high dimensionality of video face data, we developed classification framework that incorporate two random sampling technologies---random subspace and bagging. They sample the feature space and training set to train multiple stable LDA-based classifiers. Their classification outputs are then combined into a final classification decision output. Experiments on the XM2VTS face video database show that the algorithm is effective in improving recognition performance. Nearly perfect recognition results are achieved by the new algorithm.

Acknowledgment

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.

References

[1]M. Turk and A. Pentland, "Face recognition using eigenfaces", IEEE International Conference Computer Vision and Pattern Recognition, pp. 586-591, 1991.

[2]K. Fukunnaga, “Introduction to statistical pattern recognition,” Academic Press, 1991.

[3]V. Belhumeur, J. Hespanda, and D. Kiregeman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,” IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 711-720, July 1997.

[4]T. K. Ho, “The random subspace method for constructing decision forests,” IEEE Trans. On PAMI, Vol. 20, Vol. 20, No. 8, pp. 832-844, 1998.

[5]L. Breiman, “Bagging predictors,” Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996.

[6]K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Matitre, “XM2VTSDB: The Extended M2VTS Database,” Second International Conference on AVBPA, March 1999.

[7]X. Tang and Z. Li, “Frame synchronization and multi-level subspace analysis for video based face recognition,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), June 2004.

[8]T. K. Ho, J. Hull, and S. Srihari, "Decision Combination in Multiple Classifier Systems," IEEE Trans. on PAMI, Vol. 16, No.1, pp. 66-75, Jan. 1994.

[9]L. Xu, A. Krzyzak, and C. Y. Suen, "Method of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," IEEE Trans. on System, Man, and Cybernetics, Vol. 22, No. 3, 418-435, 1992.

[10] S. Satoh, “Comparative evaluation of face sequence matching for content-based video access,” In Proceedings of IEEE International Conference on Automatic Face and Gesture, Page(s): 163–168, 2000.

[11] O. Yamaguchi, K. Fukui, and K. Maeda, “Face recognition using temporal image sequence,” In Proceedings of IEEE International Conference on Automatic Face and Gesture, Page(s): 318 –323, 1998.

First A. Author (M’76–SM’81–F’87) and the other authors may include biographies at the end of regular papers. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year degree was earned. The author’s major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author’s title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, the biography will be indented around it. The photograph is placed at the top left of the biography. Personal hobbies will be deleted from the biography.

TABLE I

Comparison between the proposed random sampling framework with conventional subspace methods

Method	Accuracy (%)
Eigenface	77.3
Fisherface	86.8
Random sampling based framework with fusion by majority voting	99.0
Random sampling based framework with fusion by sum rule	99.0

TABLE II
COMPARISON BETWEEN THE PROPOSED RANDOM SAMPLING CLASSIFICATION FRAMEWORK WITH EXISTING METHODS FOR VIDEO-BASED FACE RECOGNITION

METHOD	Accuracy (%)
Mutual Subspace	79.3
Nearest frame using Euclidean distance	81.7
Nearest frame using LDA	90.9
Nearest frame using unified subspace analysis	93.2
Multi-level subspace analysis	98.0
Random sampling based framework with fusion by majority voting	99.0
Random sampling based framework with fusion by sum rule	99.0