FEATURE SELECTION FOR SUBJECT IDENTIFICATION IN SURVEILLANCE PHOTOS

Jie Wang, K.N.Plataniotis, A.N.Venetsanopoulos

The Edward S.Rogers Sr.Department of Electrical and Computer Engineering
University of Toronto, Toronto,M5S 3G4, ONTARIO, CANADA
{jwang, kostas, anv}@dsp.utoronto.ca

ABSTRACT

In this paper, a novel face recognition method is proposed for surveillance photo identification applications. In such a case, only a limited number of images per subject is available for training purposes. Furthermore, surveillance photos are usually different from the stored templates mostly due to aging, illumination and pose variations. It is common practice to apply unsupervised techniques such as principle component analysis (PCA) when the sample size for each subject is small. However, since PCA is performed without sample label considerations, the captured variation between images contains not only interpersonal variation but also intrapersonal variation which has an adverse effect on recognition performance. To overcome the problem, feature selection is performed in the PCA space to obtain a representation in which intrapersonal variation is minimized and interpersonal variation is maximized. Extensive experimentation following the FERET evaluation protocol indicates that the proposed scheme improves significantly the recognition performance.

1. INTRODUCTION

Subject identification in surveillance photos is one of the most important applications for face recognition. The identity of the subject in a surveillance photo is determined by comparing it against stored templates of known subjects. In realistic applications, only a limited amount of image data per subject is available to system. At the same time appearance of the subject of interest captured in surveillance photos significantly differs from the templates stored in the database due to aging and other environmental factors such as lighting and pose.

In literature, numerous FR algorithms have been proposed, with [1] surveying most of them. Among the various FR procedure, appearance based approaches, which treat the face image as a two dimensional holistic pattern, seem to be the most successful. In appearance based approaches, principle component analysis (PCA) and linear discriminant analysis (LDA) are the most commonly used for feature extraction. It is generally believed that supervised learning techniques such as LDA usually outperform unsupervised techniques such as PCA in recognition tasks. However, this is not always the case. Experimental analysis' indicate that when training sample size (number of available training samples / subject) is small, PCA will outperform LDA[2][3]. In surveillance applications such as the one considered here, it is not unreasonable to assume that the sample size per subject is equal to 1, forcing the use of PCA as feature extractor. In this work we propose to enhance the PCA based solution by applying a feature selection process in the PCA subspace. The objective is to maximize interpersonal variation as well as to minimize intrapersonal variation. Due to its universal acceptance, and in order to facilitate comparisons with existing solutions, the FERET database is used to support the claim under this work[4]. Following the FERET evaluation protocol, both gallery images and probe images are projected to the selected PCA subspace and identity authentication is performed by comparing the distance between probe and gallery image. Extensive experimentations on the FERET database indicate that the proposed feature selection scheme improves the recognition performance when large intrapersonal variation exists in the probe images, the application scenario most often encountered in practice.

The rest of the paper is organized as follows: Section 2 describes the system framework; In section 3, feature selection procedure in the PCA space is introduced; Experimentations on the FERET database with the analysis of the results are described in section 4 followed by a conclusion drawn in section 5.

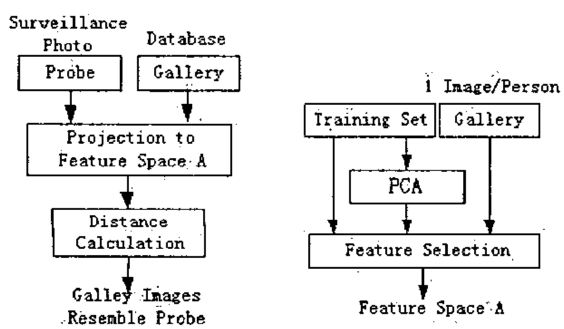
2. SYSTEM OVERVIEW

The system framework is shown in Fig.1. A surveillance photo of an unknown subject is fed to the FR system, which is asked to return the stored examples from the database which match most closely the input, along with the corre-

This work is partially supported by a Bell University Lab research grant.The authors would like to thank the FERET Technical Agent, the U.S. National Institute of Standards and Technology for providing the FERET database.

sponding labels. In order to facilitate comparisons with existing solutions, we formulate the problem using the FERET terminology. Surveillance photos with unknown subjects are called probe images while the database with known subject images is called gallery. To match the true operating characteristics of the surveillance paradigm, it is assumed that there is no overlap between gallery and probe. Furthermore, each subject in gallery is represented by a single frontal image.

The FR system is initially trained on a generic training set which is collected by the system provider. The generic training set does not overlap with gallery or the probe. PCA is used as a feature extractor followed by a feature selection procedure by using both training images and gallery images. After the initial training, the system projects each gallery image to the extracted feature space to generate their equivalent representations in the feature space. A probe image, considered here to be the actual input to the system, is also projected to the calculated feature space, and its distances from each feature-space projected gallery images are calculated. The gallery images reporting the smallest distances, in the feature space, are selected as candidates for subject identification.



(a) Operation Session (b) Training Session

Fig. 1. System Framework

3. FEATURE SELECTION IN THE PCA SPACE

The PCA feature extractor is an unsupervised linear technique which provides an optimum, in the mean square error sense, representation of input into a lower dimensional space[5]. However, since it is an unsupervised technique, there are two components in the PCA space, interclass subspace and intraclass subspace which are coupled together[6]. In the recognition application, intraclass variation is expected be small compared to the interclass variation for classification purposes. Therefore the subspace composed of the eigenvectors corresponding to the dominant eigenvalues may not be appropriate for classification due to large intraclass variation which may be captured[5]. Thus the "optimal" m-dimensional PCA space is not necessary comprised of m-largest eigenvectors and a feature selection procedure is necessary to improve the recognition performance.

Feature selection is performed in the PCA space to select m bases forming a m-dimensional feature space in which the intraclass variation will be minimized while the interclass variation will be maximized, i.e., maximize the ratio of interclass variation over intraclass variation. With such criteria, the obtained feature space has the property, which is advantageous to recognition performance, that image difference of same subject in the feature space is of small value while that of different subjects is large.

Let $\mathcal{G}$ be the gallery set with S images(identities), $\mathbf{g}_i, i = 1, 2, ..., S$. Let $\mathcal{T}$ be the training set of size $C \times L$ containing C identities, L images each. $\mathbf{t}_{i,j}$ is the jth image of identity i, $j = 1, 2, .., L; i = 1, 2, ..., C$. Therefore at most $C \times L - 1$ meaningful eigenvectors with non zero eigenvalues can be obtained by PCA when number of the training samples is less that the dimensionality of the image. Let $A = [\mathbf{a}_1, ..., \mathbf{a}_M]$ be the complete PCA feature set and the components are eigenvectors sorted in a descending order of their eigenvalues. M is chosen such that $\frac{\sum_{k=1}^{M} \lambda_k}{\sum_{k=1}^{C \times L-1} \lambda_k}$ is greater than a threshold, where $_i$ is ith eigenvalue. The reason for not using all available eigenvectors is that the eigenvectors with very small eigenvalues are usually very noisy due to limited training samples. Then the problem is becoming to select a subset A'_m with cardinality m from the complete set A. In a standard PCA methods, the criteria for selection is to maximize $\sum_{k=1}^{m} {}_k$, results in $A'_m = A_{1:m} = [\mathbf{a}_1, ..., \mathbf{a}_m]$. However, although the subspace with large eigenvalues will have large interclass variation, it may also contain large intraclass variation. Thus the purpose of maximizing the ratio between them may not be achieved. In this paper, we define the selection criteria as the one to maximize this ratio. The interclass variation is estimated by the distance between images in the gallery set. Based on the assumption that human faces share similar intraclass variation[6], intraclass variation of gallery images can be estimated from training samples. Therefore, the criteria used for selection are (a) gallery set: maximize the distance between images, which is represented by the minima of the mean value of the distance between image pairs (b) training set: minimize the mean value of the distance between images to their corresponding class centers, i.e.,

$$A'_m = arg \max_{A_F} J(A_F) \qquad J(A_F) = \frac{D_G(A_F)}{D_T(A_F)} \qquad (1)$$

where A_F is any subset of A with cardinality m, and,

$$D_G(A_F) = \min_i \left(\frac{1}{S-1} \sum_{k=1, k \neq i}^{S} ||A_F^T(\mathbf{g}_i - \mathbf{g}_k)|| \right) \qquad (2)$$

$$D_T(A_F) = \frac{1}{C \times L}\sum_{i=1}^{C}\sum_{j=1}^{L}||A_F^T(\mathbf{t}_{i,j}-\mu_i)|| \quad \mu_i = \frac{1}{L}\sum_{j=1}^{L}\mathbf{t}_{i,j} \tag{3}$$

$||\cdot||$ is the Euclidean distance between two vectors.

A Forward Selection algorithm based on the criteria discussed above is issued to select the optimal m combinations in A. We start with the most significant feature $\mathbf{a}_1$ with the largest eigenvalue. Although it may contain a large intraclass variation, it is believed to have the most of the interclass variation since interclass sample difference dominate the sample covariance[6]. Therefore, in order to avoid losing the important discriminant information, $\mathbf{a}_1$ is always be included. The whole selection procedure is as follows[7]:

(1) $A'_m(1) = [\mathbf{a}_1]$
(2) $For\ \ k = 1\ \ to\ \ m$
$A'_m(k+1)$
$= A'_m(k) \oplus arg\max_{\mathbf{a}_i \in A \quad A'_m(k)} J(A'_m(k) \oplus \mathbf{a}_i)$

Therefore A'_m is the m-dimensional feature space and the identification is performed by calculating distance between probe and gallery images in this feature space.

4. EXPERIMENTS AND RESULTS

4.1. Experiment Setup

The FERET database includes 14501 face images of 1209 subjects covering a wide range of variations in viewpoints, illuminations,facial expressions and so on. Since our system is partially automatic and does not include a face detection step, all the face images should be manually aligned and normalized which requires the coordinate information of eyes. Currently, only 3817 face images of 1200 persons in the FERET database are provided with such information. All the images are preprocessed according to the recommendation of the FERET protocol, which includes (1) images are rotated and scaled so that the centers of the eyes are placed on specific pixels and the image size is 150 × 130; (2) a standard mask is applied to remove nonface portions; (3) histogram equalized and image normalized to have zero mean and unit standard deviation. Then each image is finally represented as a vector of length 17154.

Among these 1200 subjects, there exist 226 subjects with 3 images per subject. These 678 images are formed as the training set. In addition, there are 1703 images of 256 persons with at least 4 images/subject. All these images are used as gallery and probe sets. Of these images, 1476 are frontal and 227 are non frontal. We randomly select 256 frontal images one per person to form the gallery set and the remaining images are treated as probe. The above data set configuration is also suggested in[3]. Then we further separate the probe as three subsets $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$ to test the algorithm under different conditions. $\mathcal{P}_1$ contains 914 images of 256 persons whose camera time difference with their corresponding gallery matches is less than half year($\leq 180days$). $\mathcal{P}_2$ contains 226 images of 75 persons whose camera time difference is greater than one and half year ($\geq 540days$). $\mathcal{P}_3$ contains 227 non frontal images of 48 persons. These three subsets illustrate three different scenarios. In $\mathcal{P}_1$, probe image is very close to its gallery match while in $\mathcal{P}_2$,$\mathcal{P}_3$, large intrapersonal variation is included.

The PCA algorithm is trained on the training set which results in a 677-dimensional meaningful space. We keep 270 eigenvectors as a complete feature set (A) for selection. These 270 eigenvalues contains more than 95% energy, i.e., $\frac{\sum_{k=1}^{270}\lambda_k}{\sum_{k=1}^{677}\lambda_k} > 95\%$. The feature subset ($A'_m$) size m is chosen from 1 to 100. Each gallery image and probe image are projected to the feature space. The evaluation is performed by calculating the Euclidean distance pair between each probe image and gallery image in the feature space, i.e., $d(i,j) = ||(A'_m)^T(\mathbf{g}_i - \mathbf{p}_j))||$, where $\mathbf{g}_i, i = 1,2,..,|\mathcal{G}|$ is the gallery image and $\mathbf{p}_j, j = 1,2,...,|\mathcal{P}|$ is the probe image. A probe is in the top K if the distance to its corresponding gallery match is among the K smallest distances for the gallery. Thus the recognition rate at rank K is the number of probe images in the top K divided by probe size.

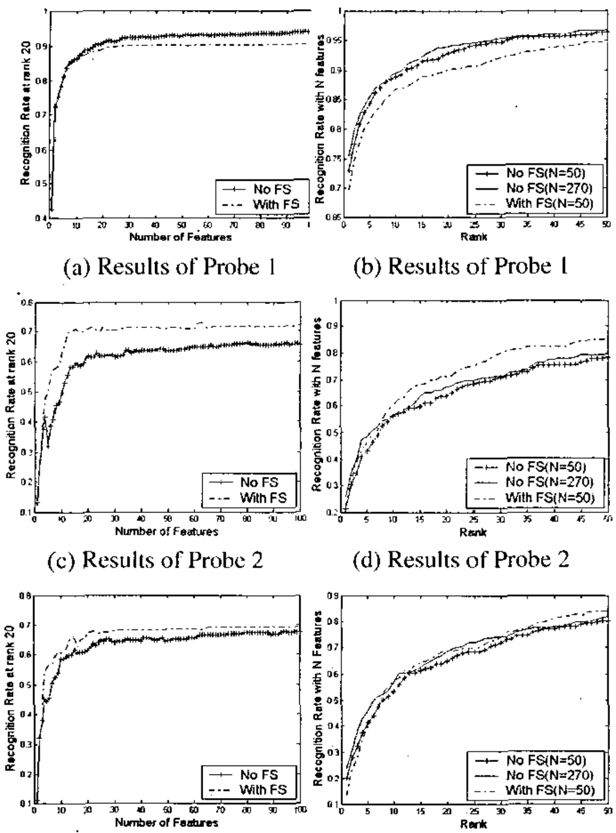
Table 1. Recognition Rate(%) at rank 20 with N features

	$\mathcal{P}_1$		$\mathcal{P}_2$		$\mathcal{P}_3$	
	N=20	N=50	N=20	N=50	N=20	N=50
PCA	90.92	92.89	61.95	63.72	61.23	64.76
FS	88.95	90.26	70.80	71.24	67.40	68.72
IMP	-1.97	-2.63	8.85	7.52	6.17	3.96

4.2. Results and Analysis

The performance of the proposed algorithm compared with PCA method is illustrated in Table.1 and Fig.2. Table.1 compares the recognition rate at rank 20 with 20 and 50 features. The results of PCA only(PCA), PCA+feature selection (FS) and performance improvement(IMP) are all listed.

The result shows that the proposed method will improve the performance in $\mathcal{P}_2$ and $\mathcal{P}_3$, however, will deteriorate in $\mathcal{P}_1$. In $\mathcal{P}_1$, the time difference between probe and gallery is within half year and the images of same identity in probe and gallery are very similar. However in $\mathcal{P}_2$ and $\mathcal{P}_3$, the difference between probe image and its corresponding gallery image is significant. The recognition is performed by project the difference image of probe and gallery to PCA space. We hope the difference of same identity (Δ_I) will have small value in PCA space while that of different identity (Δ_E) will have large value. As we discussed in section 3, PCA



(a) Results of Probe 1 (b) Results of Probe 1

(c) Results of Probe 2 (d) Results of Probe 2

(e) Results of Probe 3 (f) Results of Probe 3

Fig. 2. Performance of 3 Probe sets (with/without feature selection); left is recognition at rank 20 v.s. number of features, right is recognition with N features v.s. rank

space has two components, interclass subspace and intraclass subspace which are coupled together and both will affect the performance. However, if the probe image and its gallery match is very similar, i.e, Δ_I is very small, although intraclass variation is large, its effect on Δ_I is still very small. Therefore, in this case, interclass variation will dominate the performance. Our feature selection criteria is to maximize the ratio of interclass over intraclass variation. Although the ratio is maximized, the interclass variation will not necessary be large. Thus, PCA only will outperform that with feature selection, since PCA features with large eigenvalues usually capture large interclass variation. So in $\mathcal{P}_1$, feature selection is worse than PCA only. However, since we keep the first eigenvector which is believed to capture most of the interclass variation, the gap is not very large. On the contrary, if the probe image is significantly different with its match in gallery, i.e., Δ_I is large, intraclass variation will throw great effect on the performance. In this case, maximizing the ratio is more appororiate to improve the performance. Therefore, in $\mathcal{P}_2$ and $\mathcal{P}_3$, it shows an obvious improvement.

5. CONCLUSION

In this paper, the problem of subject identification in surveillance photos was addressed by introducing a feature selection mechanism operation on the PCA feature space. The proposed feature selector allows for the determination of a lower dimensionality feature space in which interclass variation is maximized while intraclass variations is minimized. Experimentations following the FERET protocol guidelines indicate that the proposed solution boosts the recognition performance, outperforming standard procedures such as the PCA approaches on recognition tasks when the probe images differ significantly from their corresponding gallery matches.

6. REFERENCES

[1] A.Rosenfeld W.Zhao, R.Chellappa and P.J. Phillips, "Face recognition: A literature survey (2000) ", *UMD Technical Report CAR-TR-948*,2000.

[2] A.M.Martinez and A.C.Kak, "PCA versus LDA", *IEEE Trans on PAMI*, 23(2) pp.228-233,2001.

[3] J.R.Beveridge, K.She, B.A. Draper, and G.H.Givens, "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition ", in *Proceedings of ICVPR 2001* pp.535-542.

[4] P.J. Phillips, H.Wechsler, J.Huang, and P.Rauss, "The FERET database and evaluation procedure for face recognition algorithms", *Image and Vision Computing*,16(5) pp.295-306,1998.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection", *IEEE Trans on PAMI*, 19(7) pp.711-720,1997.

[6] X.Wang and X.Tang, "Unified subspace analysis for face recognition ", in *Proceedings of ICCV 2003*, pp.679-686.

[7] S.Kumar, "Modular learning through output space decomposition", *Ph.D dissertation,Univ. of Texas at Austin*,2000.