Class-dependent Feature Selection for Face Recognition

Zhou Nina and Lipo Wang

School of Electrical and Electronic Engineering
Nanyang Technological University
Block S1, 50 Nanyang Avenue, Singapore 639798

Abstract. Feature extraction and feature selection are very important steps for face recognition. In this paper, we propose to use a class-dependent feature selection method to select different feature subsets for different classes after using principal component analysis to extract important information from face images. We then use the support vector machine (SVM) for classification. The experimental result shows that class-dependent feature selection can produce better classification accuracy with fewer features, compared with using the class-independent feature selection method.

1 Background

Automatic face recognition has experienced a relatively long process from the 1960s until now. In the 1960s, the first semi-automated face recognition system was developed [1]. The recognition process of this system included the location of features like eyebrows, eyes, noses and so on, calculation of distances and ratios to a common reference point and template matching. Later in the 1970s, some subjective features like hair color and lip thickness were developed by Goldstein et al. [2] to automate the recognition system. However, these two early solutions have one drawback, namely manually computing the measurements and locations. In order to deal with this problem, in the 1980s, Kirby and Sirovich [3] applied principal component analysis (PCA) to extract important information by singling out important face features. This application was thought of the first successful example of automatic face recognition systems. Following that, more feature extraction techniques like independent component analysis (ICA) [4] and linear discriminant analysis (LDA) [5], [6] were proposed.

Face recognition has three basic sequential processes: preprocessing, feature extraction or selection, and recognition. For different images or databases, preprocessing can vary from noise removal, normalization, to space transformation, e.g., Fourier transformation [7]. Since a large amount of information is stored in images and it is impractical to use all information in computation, feature extraction or feature selection [8], [9] is very necessary. Various feature extraction techniques, for example, PCA [10], [11], [12] and its variants [13], [5], fisher linear discriminant analysis (FLDA) [5], [6], [7], general tensor discriminant analysis

(GTDA) [14], and ICA [4], have been used to extract face features [7]. Feature selection is another way of obtaining compact face information. This involves first determining some characteristics of faces, such as the distance between eyes, the width of nose, and the length of jaw line [15], [16], then combining all those information to form a feature vector. Although features obtained by this method can be easier to interpret than those by feature extraction, it is not cost-effective work to determine which features are desirable for face images and then compute those features. In this case, feature extraction is more reliable to obtain fully representative information of face images, which will be our first step in dimensionality reduction. Recognition of faces also has many kinds of techniques, such as template matching [1], [11] and various classifiers [17], [18], [19], [20], [21].

Considering the possibility that different features have different classification power for different classes, in this paper, we propose to adopt the class-dependent feature selection [22], [23], [24] method for further dimensionality reduction in the second step. Class-dependent feature selection chooses a feature subset for each class, i.e., different feature subsets for different classes. The usual class-independent feature selection method chooses a common feature vector for all classes. This is an novel application which is not to show that our method outperforms all other existing methods on classification performance, but intends to show that class-dependent feature selection is better than class-independent feature selection. Therefore, this application will provide the possibility to employ the idea of class-dependent feature selection with many other feature extraction methods.

This paper is organized as follows. In section 2, we briefly review PCA and then describe the class-dependent feature selection method. In section 3, we utilize the SVM to realize the classification on the ORL data set [25] and compare results of class-dependent feature selection method with those of class-independent feature selection method, and also some published results. In section 4, we make a discussion about the present work.

2 Methodology

As an efficient dimensionality reduction technique in data analysis and pattern recognition, PCA [10], [13], [11], [7], [26] has already been widely used in face recognition systems. PCA [27], [10], [12] computes principal components of images, thereby transforming training images (denoted as matrix X with N samples in rows and p features in columns) into a new space of the principal components. The basic steps are: (1) calculating the covariance matrix of data matrix X; (2) determining eigenvalues and eigenvectors of this covariance matrix; (3) selecting m ($m < p$) significant eigenvectors to form transformation matrix T with the first row corresponding to the most important eigenvectors; and (4) obtaining the projected images by calculating $Y^T = TX^T$, here Y is a matrix with N rows and m columns. Through PCA, the dimension of original images is reduced to m . The dimension of kept components decides the amount of information lost.

After feature extraction, we propose to select class-dependent features from obtained principal components. For the case of high-dimensional features, it is impractical for us to sequentially add all features and evaluate all feature subsets. Since in this paper we will adopt the ORL face database [25] which has a very high dimension in our experiment, each time we will add 3 features into the previous feature subset and stop at a predetermined threshold of the dimension, e.g., 30. Each feature subset is evaluated by the SVM and the feature subset with the highest classification accuracy is chosen as the superior one of the current class.

The class-dependent feature selection is described as follows:

1. Based on the strategy of "one-against-all" [28], [29], we convert a multi-class problem into several two-class problems. For example, the problem of classifying face images is converted into 2-class classification problems, where each problem only includes two classes, i.e., one being the original class and the other one consisting of all the other classes.
2. For each 2-class problem, we adopt the class separability measure (CSM) [30], [29] to evaluate features' ranking for each class. The CSM evaluates how well two classes are separated by a feature vector. The greater the distance between different classes, the easier the classification task. For example, if S_w denotes the within-class distance and S_b denotes the between-class distance, the ratio S_w/S_b can be used to measure the separability of the classes. The smaller the ratio, the better the separability. The importance of a feature may be evaluated by ratio S_w/S_b calculated *after the feature is removed from the data set*, i.e., S'_w/S'_b. The greater S'_w/S'_b is, the more important the removed attribute is. Hence we may evaluate the importance level of the attributes according to the ratio S_w/S_b with an attribute deleted each time in turn. Each class will have a feature importance ranking list. For example, for problem 1, its ranking list of features measures the importance of features in classifying class 1 from the other classes. Therefore, this feature importance ranking list is specific to class 1. Likewise for class 2, class 3,..., and class C.

$$S_w = \sum_{c=1}^{C} P_c \sum_{j=1}^{n_c} \left[\left(\mathbf{X}_{cj} - \overline{m}_c\right) \left(\mathbf{X}_{cj} - \overline{m}_c\right)^T \right]^{1/2} \tag{1}$$

$$S_b = \sum_{c=1}^{C} P_c \left[\left(\overline{m}_c - \overline{m}\right) \left(\overline{m}_c - \overline{m}\right)^T \right]^{1/2} \tag{2}$$

   Here P_c is the probability of the c-th class, and n_c is the number of samples in the c-th class. $\mathbf{X}_{cj}$ is the j-th sample in the c-th class, $\overline{m}_c$ is the mean vector of the c-th class, and $\overline{m}$ is the mean vector of all samples in the data set.
3. According to feature importance ranking lists obtained in step 2, we need to determine a feature subset for each class. We can choose a classifier, e.g., the

SVM, to evaluate feature subsets and determine the most contributive one. For each class, each feature subset is formed by sequentially adding one or several features into the previous subset. The feature subset with the highest classification accuracy is chosen as the most contributive one. Usually, we ranked all d features and formed all d feature subsets as follows. The top 1 feature consists of the first feature subset. The second feature subset consists of the top two ranked features. The d-th feature subset includes all features. Whereas, in practical situation, e.g., for the case of high-dimensional features, it will be computationally expensive for us to sequentially form all feature subsets and evaluate them. In this paper, we will adopt ORL face database [25] which has a very high dimension in our experiment. Each time we will try add 3 features into the previous feature subset and stop at a predetermined threshold of the dimension, e.g., 30. We also evaluated all 30 feature subsets, i.e., top first, top first and second feature, ..., top 30 features and found that accuracies are increasing. When more features are added, the accuracies of the formed feature subsets keep stable. Each feature subset is evaluated by the SVM and the feature subset with the highest classification accuracy is chosen as the superior one of the current class.

In order to conveniently describe the class-dependent feature subset, we attempt to use a feature mask to express the state of each feature. The feature mask only has two elements '0' and '1', in which '0' represents the absence of a particular feature and '1' represents the presence of the feature. For example, considering a data set with 5 features $\{x_1, x_2, x_3, x_4, x_5\}$, if the optimal feature subset obtained is with the second and forth features deleted, the feature mask for this feature subset should be {1, 0, 1, 0, 1}.

After selecting class-dependent feature subsets, we adopt the SVM with RBF kernel [30], [29] for the classification because of many advantages of the SVM, such as fast speed, high recognition rate. Since class-dependent features can not directly be input into the original SVM, we adopt the class-dependent SVM classifier as [29] described. Based on the class-dependent feature mask, we construct a classifier model for each class, i.e., forming class-dependent models. Each model is trained using feature subsets specific to the corresponding class. Each testing data is filtered by the feature mask of the corresponding class before input into one model. The maximum value of all models' outputs determines the class of the testing data.

3 Experiments

3.1 Data description and preprocessing

In this experiment, we selected the Cambridge ORL face database [25] as our experimental data. It includes 40 subjects (faces or classes), each of which has 10 slightly different face images. Therefore, the total number of face images is 400. Each face image is in gray scale and has 112 by 92 pixels. For processing conveniently, we reshaped each original image matrix into a column vector, which has the dimension 10304. All image vectors constitute a new image matrix with 400 samples in rows and 10304 features in columns. Considering the effect of illumination on different face images, we adopted the normalization as [7] described, i.e., subtracting the mean of each image and divide by the variance. After that, all images were equally distributed in terms of energy [7].

3.2 Implementation and Results

Since normalized face images have a high dimension, PCA is adopted to extract principal components for each face image. The dimension of the original face image is reduced to 399 by PCA. The class-dependent feature selection method is used to further reduce features for each class. Before feature selection, we separate all 400 images faces into 200 training images and 200 testing images. For the training set, we calculate features' importance ranking for each class. Although principal components (i.e., features obtained from PCA) are sorted according to another importance measure, i.e., the eigenvalue, we believe that features' ranking for different classes are different. In Table 1, we provide the number of features selected for each class. We can see that class-dependent feature selection method selected rather different features for different classes, i.e., as many as 8 features for several classes and as few as 1 features for other classes (See Table 1). When combining features subsets of all classes in Table 1, we include 26 different features in the union set. Through the SVM, we obtained different feature masks for different classes. For example, in Table 2, we provide feature masks of the first 5 classes. For classification, the 200 testing samples were processed in the same way as the training set and tested on a class-dependent SVM classifier to produce a 98% classification accuracy, which is better than the classification result by the class-independent feature selection method (see Table 3). Table 3 provides classification results of different number of features (principal components) by using normal class-independent feature selection method. The class-independent feature selection method selected 30 features to produce classification accuracy 97.5%. Finally we compared our result with that of some existing methods in Table 4. All those experiments were done on the ORL database but with different recognition methods. The result shows that our method is very comparable.

4 Summary and Discussion

In this paper, we applied the class-dependent feature selection methods [29] on face recognition problems, after processing the whole data set using PCA. Although many fully-fledged feature extraction techniques like PCA, LDA and elastic bunch graph matching [31] exist and have good successful experiences in face recognition, we still proposed this novel application on face recognition to detect if our proposed class-dependent method has advantages in classification performance over the class-independent feature selection method. In the process of experiments, PCA was used as a preprocessing step to extract face features. Then both class-dependent and class-independent feature selection were used to

Table 1. Number of features selected for each classes

Class No.	1	2	3	4	5	6	7	8	9	10
Number of features	2	3	4	7	3	1	3	2	3	2
Class No.	11	12	13	14	15	16	17	18	19	20
Number of features	3	3	1	2	4	5	3	2	1	5
Class no.	21	22	23	24	25	26	27	28	29	30
Number of features	2	2	3	1	6	3	2	8	2	1
Class no.	31	32	33	34	35	36	37	38	39	40
Number of features	2	5	3	2	6	8	5	4	3	8

Table 2. Feature masks of the first 5 classes

Class 1	1 1 0 0 0 0 0 0
Class 2	1 1 1 0 0 0 0 0
Class 3	1 1 1 1 0 0 0 0
Class 4	1 1 1 1 1 1 1 0
Class 5	1 1 1 0 0 0 0 0

Table 3. Classification accuracy for variant numbers of features using class-independent feature selection method (10 fold cross validation)

Number of components (features)	5	10	15	20	25	30
Accuracy	80.50%	87.50%	92.00%	94.75%	95.25%	97.50%

Table 4. Comparisons of classification accuracies with existing methods

Method	Proposed	Class-independent after PCA	Kim et al.'s [13]	Lu et al's [6]
Classifier	SVM with RBF kernel	SVM with RBF kernel	Linear SVM	Nearest Neighbor
Number of selected features	26	30	120	22
Accuracy	98.0%	97.5%	97.5%	96.0%

select features for recognition. The classification results showed that the introduction of our proposed method selected features specific for each class (face). Therefore, the feature dimension of each class is less than that produced by the class-independent feature selection. Also a better classification performance is obtained compared to normal class-independent feature selection methods. Besides, the corresponding face recognition system after class-dependent feature selection has one more advantage over the normal face recognition system: when adding one more class (face) into the existing system, class-dependent recognition system does not need to re-train whole system [22]. However, the recognition system based on class-independent feature selection needs to re-train the whole system if one more class is added.

Noticeably the class-dependent feature selection method is likely to be more computationally expensive than other conventional feature selection methods. However, the extra computational cost may be worthwhile in certain applications where improvements of accuracy or reduction of data dimensionality are very important and meaningful.

References

[1] Li, S.Z., Jain, A.K.: Handbooks of face recognition. Springer, New York (2005)

[2] Goldstein, A.J., Harmon, L.D., Lesk, A.B.: Identification of Human Faces. In: Proceedings of the IEEE. Volume 59. (1971) 748–760

[3] Sirovich, L., Kirby, M.: Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America **3**(4) (1987) 519–524

[4] Liu, C.J., Wechsler, H.: Independent component analysis of Gabor features for face recognition. IEEE Trans. on Neural Networks **14**(4) (2003) 919–928

[5] Liu, C.J.: Gabor-based kernel PCA with fractional power polynomial models for face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **26**(5) (2004) 572–581

[6] Lu, J.W., Plataniotis, K.N., Venetsanopoulos, A.N.: Face recognition using LDA-Based algorithms. IEEE Trans. on Neural Networks **14**(1) (2003) 195–200

[7] Vishnubhotla, S.: Face recognition. Project Work (2005)

[8] Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York (1973)

[9] Fukunaga, K.: Statistical Pattern Recognition. Academic Press, New York (1989)

[10] Joo, S.W.: Face recognition using PCA and FDA with intensity normalization, Project Work. Availabel at: http://www.cs.umd.edu/ swjoo/reports/739Q_report.pdf (2003)

[11] Perlibakas, V.: Distance measures for PCA-based face recognition. Pattern Recognition Letters **25**(6) (2004) 711–724

[12] Wang, L.P., Fu, X.J.: Data Mining with Computational Intelligence. Springer, Berlin (2005)

[13] Kim, K.I., Jung, K.C., Kim, H.J.: Face Recognition Using Kernel Principal Component Analysis. IEEE signal processing letters **9**(2) (2002) 40–42

[14] Tao, D.C., Li, X.L., Wu, X.D., Maybank, S.J.: General Tensor Discriminant Analysis and Gabor Features for Gait Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) **29**(10) (2007) 1700–1715

[15] Craw, I., Tock, D., Bennett, A.: Finding face features. Lecture Notes in Computer Science **588** (1992) 92–96

[16] Johnson, R., Bonsor, K.: How facial recognition systems work, Available at: http://computer.howstuffworks.com/facial-recognition.htm

[17] Liu, Q.S., Lu, H.Q., Ma, S.D.: A non-parameter Bayesian classifier for face recognition. Journal of Electronics **20**(5) (2003) 362–370

[18] Liu, S., Wang, Z.: A face recognition classifier based on the RBPNN model. Computer Engineering and Science (2) (2006)

[19] Zhang, Y.K., Liu, C.Q.: Face recognition based on support vector machines and nearest neighbor classifier. Journal of Systems Engineering and Electronics (3) (2003)

[20] Zhuang, L., Ai, H.Z., Xu, G.Y.: Training support vector machines for video based face recognition. In: The 2nd International Conference on Image and Graphics. Volume 4875. (2002) 737–743

[21] Zhuang, L., Ai, H.Z., Xu, G.Y.: Video based face recognition by support vector machines. In: International Conference on Computer Vision, Pattern Recognition and Image Processing, Durham, North Carolina, USA (2002)

[22] Baggenstoss, P.: Class-specific feature sets in classification. In: Proceedings of IEEE International Symposium on Intelligent Control (ISIC). (1998) 413–416

[23] Oh, I.S., Lee, J.S., Suen, C.Y.: Analysis of class separation and combination of class-dependent features for handwriting recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **21**(10) (1999) 1089–1094

[24] Zhou, N.N., Wang, L.P.: A novel support vector machine with class-dependent features for biomedical data. In Verma, B., Blumenstein, M., eds.: Pattern Recognition Technologies and Applications: Recent Advances, Hershey, New York, Information Science Reference (2008) 284–298

[25] Cambridge, A..T.L.: The ORL Databases of Faces, Available at: http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att_faces.zip (1992-1994)

[26] Zhao, H.T., Yuen, P.C., Kwok, J.T.: A novel incremental principal component analysis and its application for face recognition. IEEE Trans. on Systems, Man and Cyber. - part B: Cyber. **36**(4) (2006) 873–886

[27] Belhumeur, P.N., Hespanha, J.P., Kreigman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. on Pattern Analysis and machine intelligence **19**(7) (1997) 711–720

[28] Hsu, C.W., Lin, C.J.: A Comparison of methods for multi-class support vector machines. IEEE Trans. on Neural Network **13**(2) (2002)

[29] Wang, L.P., Zhou, N.N., Chu, F.: A general wrapper approach to selection of class-dependent features. IEEE. Trans. on Neural networks (Coming issue on Sep. 2008)

[30] Fu, X.J., Wang, L.P.: Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Trans. on System, Man, and Cyber.–Part B: Cyber. **33**(3) (2003) 399–400

[31] Wiskott, L., Fellous, J.M., Kuiger, N., von der Malsburg, C.: Face recognition by Elastic Bunch Graph Matching. IEEE Trans. on Pattern Analysis and Machine Intelligence **19**(7) (1997) 775–779