MEDWELL PUBLISHING
International Journal of Soft Computing 4 (2): 103-108, 2009
ISSN: 1816-9503

Learning Kernel Subspace Classifier for Robust Face Recognition

[1]Bailing Zhang, [1]Sheng-Uei Guan and [2]Hanseok Ko
[1]Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou, China
[2]School of Electronics and Computer Engineering, Korea University, Seoul, 136-713, Korea

Abstract: Subspace classifiers are very important in pattern recognition in which pattern classes are described in terms of linear subspaces spanned by their respective basis vectors. To overcome the limitations of linear methods, kernel based subspace models have been proposed in the past by applying the Kernel Principal Component Analysis (KPCA). However, the projection variance in the kernel space as applied in the previously proposed kernel subspace methods, is not a good criteria for the data representation and they simply fail in many recognition problems. We address this issue by proposing a learning kernel subspace classifier which attempts to reconstruct data in the input space through the kernel subspace projection. Comparing with the pre-image methods, we emphasize the problem of how to use a kernel subspace as a model to describe input space rather than finding an approximate pre-image for each input by minimization of the reconstruction error in the kernel space. Experimental results on occluded face recognition demonstrated the efficiency of the proposed method.

Key words: Subspace classifier, principal component analysis, kernel method, robust face recognition

INTRODUCTION

Subspace classifier is a traditional pattern recognition method that has been broadly applied in signal processing and computer vision. Subspace classifier classifies a pattern based on its distance from a number of subspaces representing given classes and the basis vectors spanning a subspace correspond to the significant eigenvectors of the co-variance matrix from the corresponding class. One of the first subspace classifiers is the CLASFIC (class feature information compression) (Oja, 1983), which employs the Principal Component Analysis (PCA) to compute the basis vectors. One of the advantages of subspace method is its ability to represent feature vectors in a low dimensional space. The earlier subspace classifier CLASFIC has been extended in many ways. For example, it was found better performance could be attained if the subspaces are modified in an error-driven way, which was termed as Averaged Learning Subspace Method (ALSM) (Laaksonen and Oja, 1996).

Among many important properties of subspace classifier, the best reconstruction of input is the best known. In fact, the straightforward motivation behind the subspace classification method is the optimal reconstruction of multidimensional data with linear principal components that carry the most significant representative features. For many image recognition problems, this offers an efficient way of handling missing pixels and occlusions that frequently appear in practices. On the other hand, it has been argued that reconstruction should be considered as a constraint to classification due to the fact that visual cortex accomplishes the essential task. A network performing reconstruction and classification was proposed to model a portion of hippocampus (Gluck and Myers, 2001).

The linear subspace methods, however, are limited in performance if non-linear features are involved. Furthermore, PCA encodes data based on the second order statistics and ignores the higher-order dependencies, which may contain important discriminant information for recognition. As a solution, kernel representations can be introduced by projecting the input attributes into a high dimensional feature space, through which the complex nonlinear problems in the original space will more likely be formulated as near-linear ones (Bakir *et al*., 2004; Rosipal *et al*., 2001; Rosipal and Trejo, 2001).

In the past, several rersearches have been reported on combining the kernel method with subspace classifiers (Tsuda, 1999; Maeda and Murase, 1999). These earlier

Corresponding Author: Bailing Zhang, Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou, China

researches shared the same idea of applying the kernel principal component analysis (KPCA) (Scholkopf *et al*., 1998) and establishing the classifier based on the projection variance in the kernel space. More specifically, the kernel trick is used to map each class of input data into their respective feature space F and then PCA is performed in F to produce the nonlinear subspace of the corresponding class. A test data is then projected to all of the nonlinear subspaces and the projection variance in the kernel space is used as the discriminant for classification. Our extensive experiment on applying this idea to face recognition problems, however, showed that the simple kernel subspace classifier does not work.

In addition to Tsuda (1999), Maeda and Murase (1999) and Scholkopf *et al*. (1998) a number of applications of KPCA in pattern classification have also been discussed in recent years, for example, image denoising via pre-image algorithms (Mika *et al*., 1998; Scholkopf *et al*., 1999). Pre-image algorithms attempt to find the approximate pattern in the input space which correspond to a feature vector in the kernel space F and the algorithms are usually based on the minimization of reconstruction error in F. Pre-image algorithms, however, do not give the measurement of discrepancy between an input vector and the reconstruction in the input space from its projection onto the kernel subspace. This is more important in kernel subspace classifier as the reconstruction error in input space directly indicates the representation capability of the corresponding kernel subspace. We attempted the problem by formulating the objective as best reconstructing input data from the kernel principal component projections. To be concise, we term our new approach as learning kernel subspace classifier and our experiments on robust face recognition problems showed the superiority of the proposed method over the pre-image algorithms and some complex occlusion robust face recognition schemes published in the last several years.

REVIEW OF SUBSPACE CLASSIFIER

Assume that each of the data classes forms a lower-dimensional linear subspace distinct from the subspaces spanned by other data classes (Oja, 1983; Laaksonen and Oja, 1996), then the subspace representing a class can be defined in terms of basis vectors spanning the subspace. And a testing data item is classified based on the lengths of its projections onto each of the subspaces or, alternatively, on the distances of the test vector from these subspaces.

Let $X = (x_1,..., x_N)$ be the training data matrix belongs to a class $\omega^{(c)}$, $c = 1,..., C$, where, x_i is a training data vector, C is the number of classes. A set of orthonormal vectors p_i can be obtained by, for example, the principal component analysis of the correlation matrix $X^T X$, i.e., $p_i^T p_j = \delta_{ij}$. The basis vectors $p_i \in R^n$, $i = 1,..., d$ ($d<n$) spans a subspace for the class, which can be expressed as L: $L = L(p_1,..., p_d)$. Denote P the matrix whose column vectors are p_i, $P = [p_1,..., p_d]$.

When an unlabeled sample x is classified by the subspace classifier, the distance between x and each of the subspace is calculated by the projection $y = P^T x$. Then, x is classified into the class with the smallest distance. The distance between x and L is described as:

$$d = \|x\|^2 - \|P^T x\|^2 \quad (1)$$

Since, the first term is independent from the class, the discriminant function indicating the membership of x belonging to $\omega^{(c)}$ can be written as:

$$f_c(x) = \|P_c^T x\|^2, \; c = 1,...,C \quad (2)$$

In the training stage, the sum of the squared distances between the training samples and the subspace is minimized, {\it i.e.}, the reconstruction of x,

$$\hat{x} = \sum_{i=1}^{m} y_i p_i = \sum_{i=1}^{m} (x^T p_i)\, p_i \quad (3)$$

will be minimized. This is also equivalent to the maximization of the projection variance in Eq. (2) and the standard solution is the principal component analysis on the correlation matrix.

KERNEL PCA AND KERNEL-BASED SUBSPACE CLASSIFIER

Subspace classifiers combined with kernel methods have been proposed in Tsuda (1999) and Maeda and Murase (1999) which are all based on the direct application of Kernel Principal Component Analysis (KPCA) in the feature space. KPCA maps all data samples to a higher-dimensional feature space via the so-called kernel trick and then finds the subspace in this transformed space through the PCA for each class separately.

Suppose a high dimensional feature space, F, is related to the input space by the (nonlinear) map $\Phi(x)$: $R^n \rightarrow F$. The map Φ and the space F are determined implicitly by the choice of a kernel function k, which computes the dot product between two input examples x and y mapped into F via,

$$k(x,y) = \Phi(x) \cdot \Phi(y) \quad (4)$$

where, (·) is the vector dot product in F. The most commonly used kernel is:

$$k(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2}) \quad (5)$$

where, σ is the width of the kernel. And the space F is called\a Reproducing Kernel Hilbert Space (RKHS) generated by k (Scholkopf *et al*., 1998). The input space is then mapped to F in the way that a sample v is transformed to the kernel function centered on v:

$$v \rightarrow k(x,v) \quad (6)$$

Kernel PCA performs the same procedure as PCA in the feature space F. For a set of N patterns x_i, i = 1, 2,..., N in R^n, the N×N kernel matrix K can be formed:

$$K_{ij} = k(x_i, x_j) \quad (7)$$

The kernel matrix K should be centralized with the result as the estimate of the covariance matrix of the new feature vector in F. Then the linear PCA is simply performed on it by finding a set of principal components in the span of vectors $\{\Phi(x)_r\}$, which represents the principal axes in the kernel space.

Let $\alpha^k = [\alpha_i^k,..., \alpha_N^k]^T$ be the normalized eigenvectors and $\lambda_1 \leq ... \leq \lambda_N$ be the eigenvalues of the matrix K such that $\lambda_k (\alpha^k, \alpha^k) = 1$ for all k = 1,..., N where, $\lambda_p > 0$. It can be shown that the eigenvectors in F can be expressed as linear combinations of the mapped training samples (Scholkopf *et al*., 1998):

$$v_k = \sum_{i=1}^{N} \alpha_i^k \Phi(x_i) \quad (8)$$

with known coefficients α_i^k. For a test data point x with image
Φ(x) in the kernel space, the projection of a mapped point Φ(x) on the eigenvector v_k is therefore given by:

$$\beta_k = (v_k, \Phi(x)) = \sum_{i=1}^{N} \alpha_i^N k(x_i, x) \quad (9)$$

In RKHS, the conventional subspace classifier can be simply performed by replacing the inner product in Eq. (2) by the one from RKHS (Tsuda, 1999; Scholkopf *et al*., 1998). The discriminant function in RKHS can then be described as follows:

$$f_c(x) = \left\| \sum_{k=1}^{r} \beta_k \right\|^2 = \sum_{k=1}^{m} \sum_{i=1}^{N} \alpha_i^k k(x_i, x) \quad (10)$$

where, r is the number of principal components in F.

LEARNING KERNEL SUBSPACE CLASSIFIER IN INPUT SPACE

The kernel subspace classifier based on Eq. (9) means performing PCA in F with optimal reconstruction of Φ(x) (the map of a test point x in F) based on its projections, i.e.,

$$\rho(x) = \|P_r \Phi(x) - \Phi(x)\|^2 \quad (11)$$

is minimized for a mapped test point with its projection onto the subspace spanned by the first r eigenvectors:

$$P_r \Phi(x) = \sum_{k=1}^{r} \alpha_k v_k \quad (12)$$

where, P_r is the projection operator in F.

However, distance in Eq. (11) does not give the reconstruction of x in the input space. For the kernel subspace classifier to be efficient in data classification, the reconstructed pre-image of Φ(x) should be as close to x as possible. Consequently, a kernel subspace classifier based on Eq. (10 and 11) can not be guaranteed to work well. For the KPCA to be efficient in data classification, the reconstructed pre-image of Φ(x) should be as close to x as possible, following the same principle of PCA.

The subject of data reconstruction has been discussed in the past with the name data de-noising or pre-image of KPCA (Bakir *et al*., 2004; Mika *et al*., 1998; Scholkopf *et al*., 1999). That means, we are looking for an explicit vector $z \in R^n$ satisfying $\Phi(z) = P_n \Phi(x)$. In other words, pre-image concerns the best reconstruction of mapped data in the kernel space and the solution can be approximated by minimizing the squared distance ρ(z) between the Φ-image of a vector *z* and the reconstructed pattern in F:

$$\rho(z) = \|\Phi(z) - P_n \Phi(x)\|^2 \quad (13)$$

For kernels satisfying k (x, x) = const, ∀x, an optimal z can be determined by an iterative update scheme as follows

$$z_{t+1} = \frac{\sum_{i=1}^{N} \gamma_i \exp(-\|z_t - x_i\|^2 / c) x_i}{\sum_{i=1}^{N} \gamma_i \exp(-\|z_t - x_i\|^2 / c)} \quad (14)$$

The popular kernel type which satisfies k(x, x) = const, ∀x is the RBF kernel. Though Eq. (14) seems applicable with the kernel subspace classifier paradigm.

While, pre-image of KPCA addresses the minimization of reconstruction error in kernel space F, we emphasize the data reconstruction in the input space after the KPCA projection, as this will explicitly express the representation capability of the kernel subspace for the data class. We formulated the problem as learning kernel subspace, with the objective of minimization of reconstruction error for the input data. The objective can be simply solved based on the kernel principal component regression (Rosipal *et al*., 2001), which defines the data reconstruction as the following regression problem from the kernel space:

$$\hat{x} = \Phi\xi + \varepsilon \tag{15}$$

where:

$\Phi(x)$ = An matrix composed of vector $\Phi(x)$
ξ = A vector of regression coefficients
ε = The error term

Performing PCA on $\Phi^T\Phi$ will result in M eigenvalues $\{\lambda_j\}_{j=1}^{M}$ and corresponding eigenvectors $\{v_j\}_{j=1}^{M}$ The projection of the $\Phi(x)$onto the k-th principal components is given by Eq. (7). By projecting all the $\Phi(x)$ onto the principal component, the above equation becomes

$$\hat{x} = \Psi w + \varepsilon \tag{16}$$

where, $B = \Phi V$ is an N×M matrix and Vis an M×M matrix with V^k as its k-th column. The least squares estimate of the coefficients w becomes:

$$\hat{w} = (B^TB)^{-1}B^Tx = \Lambda^{-1}B^Tx \tag{17}$$

where, Λ = diag $(\lambda_1,..., \lambda_M)$.

The proposed model (Eq. 17) has been discussed earlier by the author from the point of view of auto-associator model (Zhang, 2005), which is a direct result of applying the kernel principal component regression (Rosipal *et al*., 2001). The classification scheme has proved its efficiency in general face recognition problem (Zhang, 2005) and in cancer classification (Zhang, 2006). Formulating the methodology in the framework of kernel subspace classifier not only justifies the model theoretically but also clarifies some confusions arose from recent works on kernelization of subspace classification.

In summary, the kernel subspace classifier model provides a description of the nonlinear relationships between input and features from the kernel space. The model building involves 2 operations. The first is the kernel operation which transforms an input pattern to a high-dimensional feature space. The 2nd is the mapping of the feature space back to the input space. The proposed kernel subspace classifier shows satisfactory performance on some benchmarking robust face recognition problems, as explained in next section.

EXPERIMENTS

Experiment with the AR faces: AR face database (Martinez and Kak, 2001) is one of the most used and cited databases of face recognition research, consisting of frontal facial images of 135 subjects (76 male and 59 females), with 26 different images for each subjects. For each subject, the images were recorded in two different sessions separated by 2 weeks, each session consisting of 13 images. Each image has 768×576 pixels.

Following the practice in Martinez and Kak (2001), we used the images of 50 subjects (the first 25 males and 25 females). In the pre-processing step, the original images were converted into gray scale, aligned by the eyes, resized and cropped to size 104×85. In our experiments, the non-screaming and non-occluded images from both sessions were used for the training of each subject's kernel subspace classifier and the remaining occluded images by sunglasses and scarf and images of the screaming expression were used for testing. The first row of Fig. 1 gives examples of training images of the third subject from the AR database, while the second row of Fig. 1 contains the test images for the same subject.

Recently, occlusion robust face recognition has attracted much attention and several algorithms have been published. In Oh *et al*. (2006) a Selective Local Nonnegative Matrix Factorization (SL-LNMF) technique was proposed, which includes the occlusion detection step and the selective LNMF-based recognition step. Paper Park *et al*. (2005) proposed a Face-ARG matching scheme in which a line feature based Face-ARG model is used to describe face images. Based on robust estimation, (Fidler *et al*., 2006) propounded a classification method that combines reconstructive and discriminative models. To be brief, we term it as Reconstructive and Discriminative Subspace (RDS) model. These published recognition performances on the AR face are compared in the Table 1. It is worthy to note that the experiment settings from these publications are not exactly same as ours except the RDS in Fidler *et al*. (2005). Therefore the comparison can only give an intuitive meaning.

Figure 2 further explains the researches of the proposed method. The first column displays the probe images from sunglasses/scarf occluded faces and the screaming face. The images from the second column to the 6th column are the first 5 best reconstructed images from the corresponding probe by applying the kernel

Fig. 1: Sample images from the AR database. First row: training images. Second row: test images with occlusion by sunglasses/scarf and with screaming expression

Fig. 2: Reconstruction of probe images from the kernel subspace classifier. 1st column: probe images; 2nd column to 6th column are the first 5 best reconstructed images from the corresponding probe image

Fig. 3: Reconstruction of probe images from the pre-image algorithm Eq. (13). 1st column: probe images; 2nd column to 6th column are the first 5 best reconstructed images from the corresponding probe image

subspace classifier. It can be observed that for sunglasses occluded face and screaming face, the kernel subspace classifier gives reasonably good reconstructions, thus yielding high recognition accuracies as shown in Table 1. For the scarf occluded face, however, the reconstruction is poor, which is consistent with the low accuracy 51%.

As the pre-image problem of KPCA is relevant to the kernel subspace classifier, we also applied it to the AR faces with result shown in Table 1. The poor performance

Table 1: Comparison of the recognition accuracies

Method	Sunglass (%)	Scarf (%)	Scream (%)
RDS	84.0	93.0	87.0
IS-ICA	65.0	NA	NA
S-LNMF	90.0	92.0	44
Face-ARG	80.7	85.2	66.7
Pre-Image	50.0	13.0	43
Kernel Subspace	92.0	51.0	95

Table 2: Comparison of the recognition accuracies

Method	Sunglass (%)	Others (%)
Pre-image	43	47
Kernel subspace	80	86

Fig. 4: Top row: training samples from the UPC data set; Bottom: some of the testing mages

Fig. 5: First column: probe images; 2nd column to 6th column are the first 5 best reconstructed images from the corresponding probe

is in agreement with the visualization of the reconstructed images from the 3 kind of probe face images, as illustrated in Fig. 3.

Experiment with the UPC faces: In the second experiment, we used the UPC faces data provided by Universitat Politecnica de Catalunya (http://gps-tsc.upc.es/GTAV), which was specially created for the purpose of testing the robustness of face recognition algorithms against strong occlusion, pose and illumination variations. This database includes a total of 18 persons with 27 pictures per person which correspond to different pose views. In our experiments, we chose 8 near-front images per person for training while used occluded images for testing, with occlusions from sunglasses or hands, as illustrated in Fig. 4.

We tested the recognition performances on two different occlusions. The first is sunglasses occlusion

which is similar to the AR face scenario. The second is occlusion by hand as shown in the bottom of Fig. 4. The recognition accuracies from our proposed kernel subspace are 80 and 86%, respectively. Figure 5 illustrates the corresponding reconstructed images from the 2 probe faces. For comparison, the recognition accuracies from the pre-image algorithm are 43 and 47%, which shows again its unacceptability in kernel subspace classification (Table 2).

CONCLUSION

In this study, a new kernel subspace classifier algorithm is proposed which is based on the KPCA image reconstruction in the input space after the KPCA projection. With the objective of minimizing the reconstruction error in the input space, the least square regression is applied to map the KPCA projection from the implicit feature space to the input space. Our experiments on some occluded face recognition problems using the AR face and UPC face showed encouraging performance, which also compared favorably with some very complex occlusion robust face recognition methods proposed in recent years.

REFERENCES

Bakir, G., J. Weston and B. Scholkopf, 2004. Learning to Find Pre-Images. Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, USA, pp: 449-456.

Fidler, S., D. Skocaj and A. Leonardis, 2006. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. IEEE. Trans. Pattern Anal. Machine Intell., 28: 337-350.

Gluck, M. and C. Myers, 2001. Gateway to Memory. An Introduction to Neural Network Modeling of the Hippocampus and Learning. The MIT Press.

Kim, J., J. Choi, J. Yi and M. Turk, 2005. Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion, IEEE. Trans. Pattern Anal. Machine Intell., 27: 1977-1981.

Laaksonen, J. and E. Oja, 1996. Subspace Dimension Selection and Averaged Learning Subspace Method in Handwritten Digit Recognition, Proceedings of ICANN, pp: 227-232.

Maeda, E. and H. Murase, 1999. Multi-category classification by kernel based nonlinear subspace method. Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, pp: 1025-1028.

Martinez, A. and A. Kak, 2001. PCA versus LDA. IEEE Trans. Pattern Anal. Machine Intell., 23: 228-233.

Mika, S., B. Scholkopf, A. Smola, K. Muller, M. Scholz and G. Ratsch, 1998. Kernel PCA and De-Noising in Feature Spaces, Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, pp: 536-542.

Oh, H., K. Lee and S. Lee, 2006. Occlusion invariant face recognition using selective LNMF basis images. Lecture Notes in Comput. Sci., 3851: 120-129.

Oja, E., 1983. Subspace Methods of Pattern Recognition, Research Studies Press, Letchworth and J. Wiley.

Park, B., K. Lee and S. Lee, 2005. Face Recognition Using Face-ARG Matching. IEEE. Trans. Pattern Anal. Machine Intell., 27: 1982-1988.

Rosipal, R. and L. Trejo, 2001. Kernel partial least squares regression in reproducing kernel hilbert space. J. Machine Learn. Res., 2: 97-123.

Rosipal, R., M. Girolami, L. Trejo and A. Cichocki, 2001. Kernel PCA for feature extraction and de-noising in non-linear regression. Neural Comput. Applic., 10: 231-243.

Scholkopf, B., A. Smola and K. Muller, 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10: 1299-1319.

Scholkopf, B., A. Smola and K. Muller, 1999. Kernel Principal Component Analysis. In: Scholkopf, B.C., J.C. Burges and A.J. Smola (Eds.). Advances in Kernel Methods-SV Learning, MIT Press, Cambridge, MA. USA, pp: 327-352.

Tsuda, K., 1999. Subspace classifier in the Hilbert Space, Pattern Recog. Lett., 20: 513-519.

Zhang, B., 2005. Kernel Auto-associator from Kernel Principal Component Autoregression with Application to Face Recognition. Proc. Int. Conf. Comput. Int. Modeling, Control and Automation (CIMCA), Vienna, pp: 15-19.

Zhang, B., 2006. Cancer classification by kernel principal component self-regression. Australian Conference on Artificial Intelligence. Horbat, Australia, pp: 719-728.