Chinese Journal of Electronics
Vol.20, No.3, July 2011

# Refined Kernel Principal Component Analysis Based Feature Extraction*

LI Junbao, YU Longjiang and SUN Shenghe

(*Department of Automatic Test and Control, Harbin Institute of Technology, Harbin 150080, China*)

**Abstract — Kernel principal component analysis (KPCA) has been widely applied in pattern recognition areas, but it endures the high store space and time consuming problems on feature extraction in the practical applications. In this paper, we propose a novel Refined kernel principal component analysis (RKPCA) based feature extraction with adaptively choosing the few samples from the training sample set but with less influence on recognition performance in the practical applications. Experimental results on seven datasets show the proposed algorithm achieves the approximate error rates but only about 20%–30% training samples. RKPCA performs well on the conditions of high computation efficiency but not a strict on recognition accuracy.**

**Key words — Kernel method, Kernel principal component analysis (KPCA), Feature extraction, Computation efficiency.**

## I. Introduction

In pattern recognition and in image processing, feature extraction based no dimensionality reduction plays the important role in the relative areas. Feature extraction simplifies the amount of resources required to describe a large set of data accurately for classification and clustering. On the algorithms, when the input data is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information), then the input data will be transformed into a reduced representation set of features also named features vector with linear transformation or the nonlinear transformation. Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input data. In the past research works, there are many applications such as image denoising[1], stochastic complexity regularization[2], Sequence outlier detection[3], face detection[4], facial expressions[5], biometrics recognition[6]. Many algorithms are presented in the previous work. Principal component analysis and linear discriminant analysis[7] are the most popular dimensionality reduction for feature extraction. For many complicated feature extraction applications, recently the nonlinear kernel based dimensionality reduction method are applied into extend the linear method to develop kernel component analysis and kernel discriminant analysis[8,9]. With kernel method in the practical application, all training samples must be saved and computed for feature extraction, which occurs the time consuming and space storing problems. In order to solve these problems, we present a novel feature extraction namely Refined kernel principal component analysis (RKPCA) in this paper. With RKPCA, only a few of training samples are computed in the algorithm procedure.

The rest of this paper is organized as follows. Section II reviews and analyzes KPCA algorithm, and Section III presents the proposed algorithm mainly on theoretical derivations and algorithm procedure. Finally, in Section IV some experiments are implemented to evaluate the performance of the proposed algorithms compared with the previous work. Conclusion is summarized in Section V.

## II. Refined Kernel Principal Component Analysis

### 1. Reviewing of KPCA

Kernel principal component analysis (KPCA) is the extension of Principal component analysis (PCA) as the linear feature extraction. The main idea of KPCA is to project the input data from the linear space into the nonlinear space, and then implement PCA in the nonlinear feature space for feature extraction. By introducing the kernel trick, PCA is extended into KPCA algorithm. The detail theoretical derivation is shown as follows.

$$C = \frac{1}{n}\sum_{i=1}^{n}(\Phi(x_i) - \bar{\Phi})(\Phi(x_i) - \bar{\Phi})^T \tag{1}$$

where $\bar{\Phi} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)$, and let $\tilde{C} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\Phi(x_i)^T$ and $Q = [\Phi(x_1), \cdots, \Phi(x_n)]$, then $\tilde{C} = \frac{1}{n}QQ^T$.

*Manuscript Received Jan. 2011; Accepted Mar. 2011. This work is supported by the National Science Foundation of China (No.61001165), Heilongjiang Provincial Natural Science Foundation of China (No.QC2010066), and HIT Young Scholar Foundation of 985 Project.

According to $\tilde{R} = Q^T Q$, with the kernel function, then

$$\tilde{R}_{ij} = \Phi(x_i)^T \Phi(x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j) \quad (2)$$

Compute the eigenvectors $u_1, u_2, \cdots, u_m$ according to the $m$th eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ of $R$, then $w_1, w_2, \cdots, w_m$ is calculated by

$$w_j = \frac{1}{\sqrt{\lambda_j}} Q u_j, \quad j = 1, 2, \cdots, m \quad (3)$$

Accordingly, $R = \hat{R} - 1_n \hat{R} - \hat{R} 1_n + 1_n \hat{R} 1_n$, where $(1_n)_{ij} = 1/n$ $(i, j = 1, 2, \cdots, n)$, then

$$y_j = w_j^T x = \frac{1}{\sqrt{\lambda_j}} u_j^T [k(x_1, x), k(x_2, x), \cdots, k(x_n, x)] \quad (4)$$

PCA based feature extraction needs to store the $r \times m$ coefficient matrix $\boldsymbol{W}$, where $r$ is the number of principal components, and $m$ is the number of training samples. While KPCA based feature extraction need to store the original sample information owing to computing the kernel matrix with all training samples, which leads to a huge store space and a high computing consuming.

**2. Theory derivation of RKPCA**

In this section, we present a novel learning called Refined kernel principal component analysis (RKPCA) with the viewpoint of Support vector machine (SVM). In SVM, only few support vectors are meaning for classification, and other samples can be ignored for training the classifier. We introduce the idea of SVM into KPCA, and choose the few training samples for KPCA. Firstly we apply a Least squares support vector machine (LS-SVM) formulation to KPCA which is interpreted as a one-class modeling problem with a target value equal to zero around which one maximizes the variance. Then, the objective function can be described as

$$\max_w \sum_{i=1}^{N} [0 - w^T (\phi(x_i) - u^\phi)]^2 \quad (5)$$

where $\phi : \mathbb{R}^N \to \mathbb{R}^l$ denotes the mapping to a high-dimensional feature space and $u^\phi = \dfrac{1}{N} \sum_{i=1}^{N} \phi(x_i)$. We formulate KPCA with direct sparse kernel learning method, and we also use the phase "expansion coefficients" and "expansion vectors" Supposed a matrix $\boldsymbol{Z} = [z_1, z_2, \cdots, z_{N_z}]$, $\boldsymbol{Z} \in \mathbb{R}^{N \times N}$, composed of $N_z$ expansion vectors, and $\beta_i$ $(i = 1, 2, \cdots, N_z)(N_z < N)$ are expansion coefficients, we modify the optimization problem to the following constraint optimization problem:

$$\begin{aligned} \max_{w,e} J(w, e) =& -\frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \\ \text{subject to} \quad e_i =& w^T (\phi(x_i) - u^\phi), \quad i = 1, 2, \cdots, N \\ w =& \sum_{i=1}^{N_z} \phi(z_i) \beta_i \end{aligned} \quad (6)$$

where $\phi(Z) = [\phi(z_1), \phi(z_2), \cdots, \phi(z_{N_z})]$. Now our goal is to solve the above optimization problem. We can divide the above optimization problem into two steps, one is to find the optimal expansion vectors and expansion coefficients; second is to find the optimal projection matrix. When $Z$ is fixed, then we apply the kernel function, that is, $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Given a random $Z$, then the above problem is same to the following problem.

$$\begin{aligned} W(Z) :=& \max_{\beta,e} -\frac{1}{2} \beta^T K_z \beta + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \\ \text{subject to} \quad e_i =& \beta^T g(x_i), \quad i = 1, 2, \cdots, N \end{aligned} \quad (7)$$

where

$$\begin{aligned} g(x_i) =& \Big[ k(z_1, x_i) - \frac{1}{N} \sum_{q=1}^{N} k(z_1, x_q) \cdots k(z_{N_z}, x_i) \\ & - \frac{1}{N} \sum_{q=1}^{N} k(z_{N_z}, x_q) \Big]^T, \\ \beta =& [\beta_1, \beta_2, \cdots, \beta_{N_z}]^T, \\ K_z =& [k(z_i, z_j)] \end{aligned}$$

The solution of the above constrained optimization problem can often be found by using the so-called Lagrangian method. We define the Lagrangian method

$$\begin{aligned} L(\beta, e, \alpha) =& -\frac{1}{2} \beta^T K_z \beta + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \\ & - \sum_{i=1}^{N} \alpha_i (e_i - \beta^T g(x_i)) \end{aligned} \quad (8)$$

with the parameter $\alpha_i$, $i = 1, 2, \cdots, N$. The Lagrangian $L$ must be maximized with respect to $\beta, \alpha_i$, and $e_i$ $(i = 1, 2, \cdots, N)$, and the derivatives of $L$ with respect to them must vanish. We can obtain the optimal solution $\alpha^z$, which is an eigenvector of the $G^T (K_z)^{-1} G$ corresponding to the largest eigenvalue.

$$\beta^z = (K_z)^{-1} G \alpha^z \quad (9)$$

where $G = [g(x_1), g(x_2), \cdots, g(x_N)]$, and now our goal is to find the optimal $Z$ that maximizes the following equation.

$$W(Z) = -\frac{1}{2} (\beta^z)^T K_z (\beta^z) + \frac{\gamma}{2} (\beta^z)^T G G^T (\beta^z) \quad (10)$$

So it is easy to achieve $Z^*$ to maximize the above Eq.(10). After we obtain $Z^*$, and we can obtain $A = [\alpha_1, \alpha_2, \cdots, \alpha_m]$ corresponding to the largest eigenvalue of $G^T (K_z)^{-1} G$. Then we can obtain

$$B = (K_z)^{-1} G A \quad (11)$$

Then, for a input vector $\boldsymbol{x}$, its feature $Y_x$ is calculated with the following equation.

$$Y_x = B K_{ZX} \quad (12)$$

where $K_{ZX}$ is the kernel vector calculated with the input vector $x$ and the refined training set $Z^*$.

As above discussion from the theoretical viewpoints, Refined kernel principal component analysis (RKPCA) algorithm adaptively chooses the few samples from the training sample set but little influence on recognition performance, which saves much space of storing training samples for computing the kernel matrix with lower time consuming. So in the practical applications, RKPCA can solve the limitation from KPCA owing to its high store space and time consuming its ability on feature extraction. So from the theory viewpoint, RKPCA

is adaptive to the applications with the demand of the strict computation efficiency but not strict on recognition.

## III. Simulation Results

Firstly, we use the six UCI datasets popular widely in pattern recognition area to testify the performance of the proposed algorithm compared with the KPCA algorithm using the part of training samples and the whole size of samples. In the experiments, we randomly the one hundred of training samples on each training sample set, especially 20 parts on Image and Splice dataset. In the experiments, we choose the Gaussian kernel with its parameters determined by the training samples. The experimental results are shown in Table 1 and Table 2, and the second column shows the error rate of each algorithm on the corresponding dataset. The third column shows the number of training samples in Table 1, and the number of training samples in the proposed algorithm in Table 2. And in the brackets denote the ratio between the number of training samples of KPCA with the common training method and the proposed training samples. The results show that the proposed algorithm achieves the similar recognition performance, but the proposed algorithm only use the less size of training set. For example, only 8% training samples are used but only error rate 2.8% higher than the common methods. Since only small size of training samples are applied in the proposed algorithm, so it saves store more place and time consuming compared with the traditional KPCA.

Secondly, we elevate the performance on Wisconsin breast cancer database[10] consisting of 569 instances including 357 benign samples and 212 malignant samples. And each one represents FNA test measurements for one diagnosis case. For this dataset each instance has 32 attributes, where the first two attributes correspond to a unique identification number and the diagnosis status (benign or malignant). The rest 30 features are computations for ten real-valued features, along with their mean, standard error and the mean of the three largest values ("worst" value) for each cell nucleus respectively. As shown in Table 3, the recognition accuracy 5.4% and 3.8% are achieved by the common training method and the proposed training method. But only 37% training samples are applied in the training procedure. As shown in the Table 3, only 37% training samples are used but only error rate 1.6% higher than the common methods. Some storing space is saved and high computation efficiency is achieved for the practical applications.

**Table 1. Recognition performance of KPCA**

| Datasets | Error rate (%) | Training samples |
|---|---|---|
| Banana | 13.6±0.1 | 400 |
| Image | $4.8 \pm 0.4$ | 1300 |
| F.Solar | $31.4 \pm 2.1$ | 666 |
| Splice | $8.6 \pm 0.8$ | 1000 |
| Thyroid | $2.1 \pm 1.0$ | 140 |
| Titanic | $22.8 \pm 0.3$ | 150 |

From the above experimental results, we can obtain the similar recognition accuracy but the smaller size of training samples. So, owing to its high store space and time consuming its ability on feature extraction is decreased in the practical applications. Refined kernel principal component analysis (RKPCA) saves much space of storing training samples for computing the kernel matrix with lower time consuming compared with Kernel principal component analysis (KPCA). On the two datasets, the results show that only 8% training samples are used but only error rate 2.8% higher than the common methods on UCI datasets, and only 37% training samples are used but only error rate 1.6% higher than KPCA method.

**Table 2. Recognition performance of RKPCA**

| Datasets | Error rate (%) | Training samples |
|---|---|---|
| Banana | $14.2 \pm 0.1$ | 120 (30%) |
| Image | $5.4 \pm 0.3$ | 180 (14%) |
| F.Solar | $34.2 \pm 2.3$ | 50 (8%) |
| Splice | $9.4 \pm 0.9$ | 280 (28%) |
| Thyroid | $2.2 \pm 1.3$ | 30 (21%) |
| Titanic | $24.4 \pm 0.4$ | 30 (20%) |

**Table 3. Performance comparison on KPCA and RKPCA**

| Algorithms | Error rate (%) | Training samples |
|---|---|---|
| RKPCA | $5.4 \pm 0.3$ | 110 (37%) |
| KPCA | $3.8 \pm 0.4$ | 300 |

## IV. Conclusion

In this paper, we propose a novel feature extraction namely Refined kernel principal component analysis (RKPCA) through adaptively choosing the few samples from the training sample set but little influence on recognition performance. RKPCA saves much space of storing training samples for computing the kernel matrix with lower time consuming in the practical applications. RKPCA overcomes the limitation endured by KPCA including the high store space and time consuming, and has many applications in image classification, face recognition, and speech recognition.

### References


[1] Y. Sun, Z. Wei, M. Wu, L. Xiao, X. Fei, "Image poisson denoising using sparse representations", *Acta Electeonica Sinica*, Vol.39, No.2, pp.285–290, 2011. (in Chinese)

[2] Q. Liu, J. Wang, W. Chen, Z. Qin, "An automatic feature selection algorithm for high dimensional data based on the stochastic complexity regularization", *Acta Electronica Sinica*, Vol.39, No.2, pp.370–374, 2011. (in Chinese)

[3] F. Jiang, J. Du, Y. Ge, Y. Sui, C. Cao, "Sequence outlier detection based on rough set theory", *Acta Electronica Sinica*, Vol.39, No.2, pp.345–350, 2011. (in Chinese)

[4] W.C. Hu, C.Y. Yang, D.Y. Huang, C.H. Huang, "Feature-based face detection against skin-color like backgrounds with varying illumination", *Journal of Information Hiding and Multimedia Signal Processing*, Vol.2, No.2, pp.123–132, 2011.

[5] S. Krinidis, I. Pitas, "Statistical analysis of human facial expressions", *Journal of Information Hiding and Multimedia Signal Processing*, Vol.1, No.3, pp.241–260, 2010.

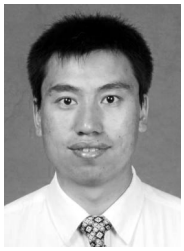
[6] M. Parviz, M. Shahram, "Boosting approach for score level fusion in multimodal biometrics based on AUC maximization", *Journal of Information Hiding and Multimedia Signal Processing*, Vol.2, No.1, pp.51–59, 2011.

[7] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.711–720, 1997.

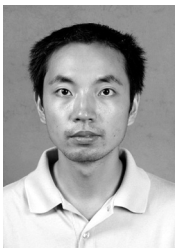
[8] T.V. Gestel, B. Baesens, D. Martens, "From linear to non-linear kernel based classifiers for bankruptcy prediction", *Neurocomputing*, Vol.73, No.16-18, pp.2955–2970, 2010.

[9] Q. Zhua, "Reformative nonlinear feature extraction using kernel MSE", *Neurocomputing*, Vol.73, No.16-18, pp.3334–3337, 2010.

[10] W.H. Wolberg, W.N. Street, D.M. Heisey, O.L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology", *Human Pathology*, Vol.26, pp.792–796, 1995.



**LI Junbao** was born in 1978. He received Ph.D. degree from Harbin Institute of Technology (HIT) in 2008, M.S. and B.S. degrees in 2004 and 2002 respectively from HIT. Now he is working as a lecturer in Department of Automatic Test and Control of HIT. His research interests focus on patter recognition and image processing, especially on kernel learning and its applications. In the past research work, more than 30 papers are published, most of them are indexed by SCI and EI. These papers are cited for more than 50 times. He was invited as the reviewer by many journals including IEEE Transaction on SMC-Part C, Neural Computing and Applications, and other SCI journals. (Email: junbaolihit@gmail.com)



**YU Longjiang** was born in Oct. 1976 in Harbin , China. He received the degrees of B.S., M.S. and Ph.D. in 1999, 2001, and 2006 respectively from Harbin Institute of Technology. Currently he is a lecturer in Harbin Institute of Technology. He is a long-term reviewer of Elsevier journal Information Science. He has issued about twenty papers in overseas and domestic journals and conferences, of which fifteen papers are indexed by SCI and EI. He has seven invention patents, of which four patents are authorized. His research interests include image processing, pattern recognition, information security, and automatic bus and interface.

**SUN Shenghe** was born in Qing Yuan, P.R. Korea in Oct. 1937. He graduated from Harbin Institute of Technology in 1961. He is a professor and a supervisor of Ph.D. candidates in Harbin Institute of Technology. He has received a second prize of National Advancement in Science and Technology Awards, an award in National Science Congress, four first prizes and eight second prizes in ministry level. He has published about 300 papers in overseas and domestic journals and conferences, of which about 200 papers are indexed by SCI and EI. He has written 5 monographs. He has 8 authorized patents. He has supervised 38 Ph.D. candidates, of which three pieces of thesis are awarded as national excellent hundred Ph.D. thesis prize. His research interests include instruments and systems, signal processing, and information security.