Regularized Locality Preserving Learning of Pre-Image Problem in Kernel Principal Component Analysis

Wei-Shi Zheng[1,3,4] and Jian-huang Lai[2,3,5,*]
[1]*Mathematics Department, Sun Yat-sen University, Guangzhou, P. R. China*
[2]*School of Information Science & Technology, Sun Yat-sen University, Guangzhou, P. R. China*
[3]*Guangdong Province Key Laboratory of Information Security, P. R. China*
[4]*wszheng@ieee.org,* [5]*stsljh@mail.sysu.edu.cn*

Abstract

In this paper, we address the pre-image problem in kernel principal component analysis (KPCA). The pre-image problem finds a pattern as the pre-image of a feature vector defined in the nonlinear principal component space produced by KPCA. Since the pre-image typically seldom exists in general, an approximate solution is appreciated. By posing a novel perspective, we find the pre-image with regularized locality preserving learning. Our approach achieves a unique solution, avoiding iteration and numerical instability. Significant superiority of the proposed novel algorithm is demonstrated by driving two applications, namely face denoising and occluded face reconstruction, as comparing with some existing well-known methods on pre-image learning.

1. State of The Art of Pre-Image Learning

Principal Component Analysis (PCA) [1] is a well-known technique which has been widely applied to unsupervised learning, dimension reduction, and image analysis etc. However linear PCA cannot handle data with nonlinear structure well while explicitly finding a nonlinear transform is always hard. One popular technique tackles this problem is called Kernel Principal Component Analysis (KPCA) [2].

Let $\mathbf{X}$ be the input space and $\mathbf{H}_k$ be the Reproducing Kernel Hilbert Space (RKHS) associated with the kernel $k(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$, where $\mathbf{x},\mathbf{y} \in \mathbf{X}$ and $\phi(\cdot)$ is an implicit mapping induced by kernel $k(\cdot,\cdot)$ such that $\phi(\mathbf{x}) : \mathbf{X} \to \mathbf{H}_k$. Take denoising utilizing KPCA for example. For any noisy pattern $\mathbf{x} \in \mathbf{X}$, to perform denoising, $\phi(\mathbf{x})$ is projected onto the subspace produced by linear PCA in the feature space $\mathbf{H}_k$. Let $P_k\phi(\mathbf{x})$ be the projection of $\phi(\mathbf{x})$ onto such kernel principal component subspace. However, $P_k\phi(\mathbf{x})$ is still defined in $\mathbf{H}_k$ and we would want to have its pre-image (the denoised pattern) in the input space $\mathbf{X}$.

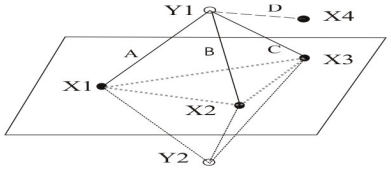


Figure 1. Illustration of the problem in the distance constraint scheme (see text for details)

The pre-image problem in KPCA therefore attempts to find a pattern $\tilde{\mathbf{x}} \in \mathbf{X}$ such that $\phi(\tilde{\mathbf{x}}) = P_k\phi(\mathbf{x})$ [3]. Unfortunately, it is ideal. It is possible that $\mathbf{H}_k$ holds higher or infinite dimensionality in general while $\mathbf{X}$ is a finite dimensional input space. Hence, they are not isomorphic and an exact pre-image $\tilde{\mathbf{x}}$ always seldom exists. To tackle this problem, as a special case of [8], S. Mika et al. found approximate pre-image with least square minimization: $\tilde{\mathbf{x}} = \arg\min_{\hat{\mathbf{x}} \in \mathbf{X}} \| \phi(\hat{\mathbf{x}}) - P_k\phi(\mathbf{x}) \|^2$ [3]. However, it is iterative. J. T. Kwok et al. proposed to find the pre-image via distance constraint [4]. It would be a nice idea and is non-iterative. However, when it is applied to more challenging data, such as faces, we surprisingly find if the number of neighbors used in their algorithm is small, it would fail. A simple example to explain such scenario is illustrated in Fig. 1. In three-dimensional space, if there exists a point $\mathbf{z}$ that has the distance constraint with neighbors $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ such that $|\mathbf{z}\mathbf{x}_1| = \mathrm{A}$, $|\mathbf{z}\mathbf{x}_2| = \mathrm{B}$ and $|\mathbf{z}\mathbf{x}_3| = \mathrm{C}$, then it is possible to find only two symmetric solutions for $\mathbf{z}$, i.e. $\mathbf{Y}_1$ and $\mathbf{Y}_2$. But if one more neighbor $\mathbf{x}_4$ and constraint $|\mathbf{z}\mathbf{x}_4| = \mathrm{D}$ are provided, then only $\mathbf{Y}_1$ is

* correspondence author

required. Though it is just a simple counterexample, however, it potentially shows the solution of the pre-image problem is always not unique under the distance constraint if few neighbors are used in contrast to high dimension of data. Recently, G. H. Bakır et al. learnt pre-image with regression [5]. As shown in their experiment [5], their method achieved better visual result but the Mean Square Error was lower than PCA.

All former work [3-5] have suggested that for image processing and pattern recognition, finding an appropriate pre-image that achieves smaller MSE, better visual result and sometime better classification may be more important and appreciated rather than only finding the purely approximate pre-image.

In this paper, we would give a novel perspective to the pre-image and propose a novel approach to learn a pre-image with regularized locality preserving. Our motivation is inspired by the manifold learning, like LLE [6]. The solution of pre-image is unique. Iteration and numerical instability are avoided. Our approach is not complex but achieves significant improvement against some well-known methods. Our algorithm is implemented for KPCA. However, it is feasible to be extended to other kernel methods.

2. Brief Review of KPCA

Suppose $\mathbf{x}_1,\cdots,\mathbf{x}_N \in \mathbf{X}$ are N training samples. Let $\mathbf{\Phi} = (\phi(\mathbf{x}_1),\cdots,\phi(\mathbf{x}_N))$, then $\boldsymbol{\mu}^\phi = \frac{1}{N}\sum_{i=1}^N \phi(\mathbf{x}_i) = \frac{1}{N}\mathbf{\Phi}\mathbf{e}$ is the mean of all samples, where $\mathbf{e} = (1,\cdots,1)^T \in R^N$. Let $\mathbf{O}_t^\phi = \frac{1}{\sqrt{N}}(\phi(\mathbf{x}_1)-\boldsymbol{\mu}^\phi,\cdots,\phi(\mathbf{x}_N)-\boldsymbol{\mu}^\phi)$, then

$$\mathbf{O}_t^\phi = (1/\sqrt{N})(\mathbf{\Phi}-\boldsymbol{\mu}^\phi\mathbf{e}^T) = (1/\sqrt{N})\mathbf{\Phi}(\mathbf{I}-\mathbf{e}\mathbf{e}^T/N) \quad (1)$$

and the total scatter matrix is defined as

$$\mathbf{S}_t^\phi = \mathbf{O}_t^\phi\mathbf{O}_t^{\phi^T} = (1/N)\mathbf{\Phi}(\mathbf{I}-\mathbf{e}\mathbf{e}^T/N)\mathbf{\Phi}^T$$

where $(\mathbf{I}-\mathbf{e}\mathbf{e}^T/N)(\mathbf{I}-\mathbf{e}\mathbf{e}^T/N)^T = (\mathbf{I}-\mathbf{e}\mathbf{e}^T/N)$. In essence, KPCA performs linear PCA in the feature space and also solves the eigenvalue problem below:

$$\mathbf{S}_t^\phi\mathbf{U}^\phi = \mathbf{U}^\phi\mathbf{\Lambda}^\phi \quad (2)$$

where $\mathbf{U}^\phi = (\mathbf{u}_1^\phi,\cdots,\mathbf{u}_q^\phi)$, $\mathbf{\Lambda}^\phi = diag(\lambda_1^\phi,\cdots,\lambda_q^\phi)$, $\lambda_1^\phi \geq \cdots \geq \lambda_q^\phi > 0$. From (2), for each $\mathbf{u}_i^\phi$, we could have some $\boldsymbol{\alpha}_i = (\alpha_1^i,...,\alpha_N^i)^T$ [2] such that

$$\mathbf{u}_i^\phi = \sum_{j=1}^N \alpha_j^i \frac{1}{\sqrt{N}}(\phi(\mathbf{x}_j)-\boldsymbol{\mu}^\phi) = \mathbf{O}_t^\phi\boldsymbol{\alpha}_i = \mathbf{\Phi}\mathbf{p}_i \quad (3)$$

It is also valid because of the representer theorem of Reproducing Kernel Hilbert Space [2]. So $\mathbf{U}^\phi = \mathbf{\Phi}\mathbf{P}$, $\mathbf{P} = (\mathbf{p}_1,\cdots,\mathbf{p}_q)$, $\mathbf{p}_i = (1/\sqrt{N})(\mathbf{I}-\mathbf{e}\mathbf{e}^T/N)\boldsymbol{\alpha}_i$. Then, for a given pattern $\mathbf{x}$, the projection $P_k\phi(\mathbf{x})$ of $\phi(\mathbf{x})$ onto the subspace spanned by the first q_0 largest kernel principal components $\mathbf{U}_{q_0}^\phi = \mathbf{\Phi}\mathbf{P}_{q_0}$, $\mathbf{P}_{q_0} = (\mathbf{p}_1,\cdots,\mathbf{p}_{q_0})$, is:

$$\begin{aligned} P_k\phi(\mathbf{x}) &= \mathbf{U}_{q_0}^\phi\,\mathbf{U}_{q_0}^{\phi\ T}(\phi(\mathbf{x})-\boldsymbol{\mu}^\phi)+\boldsymbol{\mu}^\phi \\ &= \mathbf{\Phi}\mathbf{P}_{q_0}\mathbf{P}_{q_0}^T\mathbf{\Phi}^T(\phi(\mathbf{x})-\tfrac{1}{N}\mathbf{\Phi}\mathbf{e})+\tfrac{1}{N}\mathbf{\Phi}\mathbf{e} \\ &= \mathbf{\Phi}\boldsymbol{\gamma}^{\mathbf{x}} \end{aligned} \quad (4)$$

where $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$ is the kernel matrix of training data and

$$\boldsymbol{\gamma}^{\mathbf{x}} = (\gamma_1^{\mathbf{x}},\cdots,\gamma_N^{\mathbf{x}})^T = \mathbf{P}_{q_0}\mathbf{P}_{q_0}^{\ T}\mathbf{\Phi}^T\phi(\mathbf{x}) - \tfrac{1}{N}\mathbf{P}_{q_0}\mathbf{P}_{q_0}^{\ T}\mathbf{K}\mathbf{e} + \tfrac{1}{N}\mathbf{e} \quad (5)$$

Our pre-image learning task is to find the appropriate solution $\tilde{\mathbf{x}}$ as the pre-image of $P_k\phi(\mathbf{x})$.

3. Regularized Locality Preserving Learning of Pre-image Problem

Motivation. We observe that, the dimension of the feature space $\mathbf{H}_k$ is always higher than the input space $\mathbf{X}$. To address the pre-image problem in KPCA, inspired by the spirit of manifold learning, our idea is to pose $\mathbf{H}_k$ as a high dimensional space and $\mathbf{X}$ as its low dimensional space. Then we could treat the approximate pre-image $\tilde{\mathbf{x}}$ of $P_k\phi(\mathbf{x})$ as an embedding point found for $P_k\phi(\mathbf{x})$ in $\mathbf{X}$. Based on this idea we learn such embedding point for $P_k\phi(\mathbf{x})$ as the pre-image by preserving the local relationship, i.e. the local linear reconstruction relationship of $P_k\phi(\mathbf{x})$ with $\phi(\mathbf{x}_1)$,…, $\phi(\mathbf{x}_N)$. Since the exact pre-image seldom exists, as a novel approach, our learning model is also just an approximation scheme with locality preserving relationship. We therefore perform regularization for the reconstruction weights in order to prevent overfitting. Details are as follows.

Modeling. For given $P_k\phi(\mathbf{x})$, let $\phi(\hat{\mathbf{x}}_1),\cdots,\phi(\hat{\mathbf{x}}_s)$ be its s distinct nearest neighbors from the training set $\{\phi(\mathbf{x}_1),\cdots,\phi(\mathbf{x}_N)\}$ in $\mathbf{H}_k$. Then we learn the local linear reconstruction weights $\mathbf{W}_{\min}^{\mathbf{x}} = (w_1^{\mathbf{x}},\cdots,w_s^{\mathbf{x}})^T$ by

$$\mathbf{W}_{\min}^{\mathbf{x}} = \underset{\mathbf{W}^{\mathbf{x}}=(w_1^{\mathbf{x}},..,w_s^{\mathbf{x}})^T}{\arg\min} \| \textstyle\sum_{i=1}^s w_i^{\mathbf{x}}\phi(\hat{\mathbf{x}}_i) - P_k\phi(\mathbf{x}) \|^2 \quad (6)$$

To avoid overfitting, we do regularization and have:

$$\mathbf{W}_{\min}^{\mathbf{x}} = \underset{\mathbf{W}^{\mathbf{x}}=(w_1^{\mathbf{x}},..,w_s^{\mathbf{x}})^T}{\operatorname{argmin}} \| \textstyle\sum_{i=1}^s w_i^{\mathbf{x}}\phi(\hat{\mathbf{x}}_i) - P_k\phi(\mathbf{x}) \|^2 + \lambda \| \mathbf{W}^{\mathbf{x}} \|^2 \quad (7)$$

where $\lambda>0$ is a regularized parameter. We solve (7) by

$$\mathbf{W}_{\min}^{\mathbf{x}} = (\hat{\mathbf{\Phi}}^T\hat{\mathbf{\Phi}} + \lambda\times\mathbf{I})^{-1}\hat{\mathbf{\Phi}}^T P_k\phi(\mathbf{x}) \quad (8)$$

where $\hat{\mathbf{\Phi}} = (\phi(\hat{\mathbf{x}}_1),\cdots,\phi(\hat{\mathbf{x}}_s))$. Integrating equality (4) that $P_k\phi(\mathbf{x}) = \mathbf{\Phi}\boldsymbol{\gamma}^{\mathbf{x}}$, we then obtain:

$$\mathbf{W}_{\min}^{\mathbf{x}} = (\hat{\mathbf{\Phi}}^T\hat{\mathbf{\Phi}} + \lambda\times\mathbf{I})^{-1}\hat{\mathbf{\Phi}}^T\mathbf{\Phi}\boldsymbol{\gamma}^{\mathbf{x}} \quad (9)$$

Note that $\phi(\hat{\mathbf{x}}_1),\cdots,\phi(\hat{\mathbf{x}}_s)$ have their exact pre-images $\hat{\mathbf{x}}_1,\cdots,\hat{\mathbf{x}}_s$ in $\mathbf{X}$ respectively. We aim to find the pre-image of $P_k\phi(\mathbf{x})$ as an embedding point by preserving the regularized local linear reconstruction relationship with $\hat{\mathbf{x}}_1,\cdots,\hat{\mathbf{x}}_s$ as $P_k\phi(\mathbf{x})$ does in the feature space $\mathbf{H}_k$

developed by equalities (7) and (9). Based on this idea, we find the pre-image of $P_k\phi(\mathbf{x})$ by

$$\tilde{\mathbf{x}} = (\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_s)\mathbf{W}^{\mathbf{x}}_{\min} \quad (10)$$

Our learning absorbs the spirit of manifold learning by treating $\mathbf{H}_k$ as a high dimensional space and $\mathbf{X}$ as its low dimensional space, and finds the approximate pre-image by learning an embedding point with regularized locality preserving.

Considering that when s=N, then the embedding pre-image $\tilde{\mathbf{x}}$ could be reconstructed with all samples $\mathbf{x}_1, \cdots, \mathbf{x}_N$, where the reconstruction coefficients are determined by the KPCA with regularization in (9). When s=1, our approach would likely almost degrade as a simple nearest neighbor classifier (NN) regardless of scaling in the feature space. However, we do not recommend setting s=N or s=1. Because $\tilde{\mathbf{x}}$ would intuitively become smoother as s is larger and when s=1 such simple NN classifier could not handle the problem well. Also the parameter λ is important, and we would show the performances with different values of it in our experiments. The experiment results would support the feasibility and superiority of our approach.

4. Applications

All experiments are based on YALEB [7] database. To produce noisy images and occluded images, we use the subset 1 in YALEB. It contains 10 persons with 9 different poses. Each pose of each person contains 7 faces with nice illumination. Totally 630 images are selected. All image are aligned with size 92×112.

Noisy Images. For each face image in the subset 1, we produce 2 noisy images, where the noise type is Gaussian with mean 0 and variance with 0.5. Therefore there are 1260 noisy face images produced.

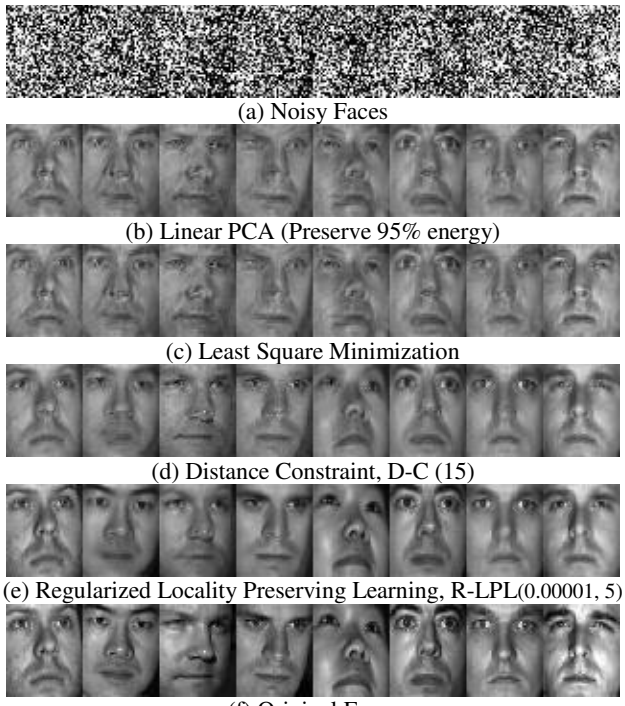
Occluded Images. Similarly, for each face image in subset 1, we produce 2 occluded faces. The occlusion is simulated by a rectangle black patch at a random coordinate, where the width and the height of the patch are randomly determined such that both width and height are 20 pixels at least and 50 pixels at most.

It is noted that all images, including occluded and noisy images, are linearly stretched to full range of pixel values of [0, 1].

KPCA Subspace. The kernel principal component subspace is trained by all images in subset 1, totally 630 images from 9 poses. The largest kernel principal components are selected to preserve 95% energy.

Kernel Function. Due to limited length, we mainly use the RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / c)$, $c = 10^5$.

Notations. "R-LPL(a,b)" means λ=a,s=b in regularized locality preserving learning of pre-image. "D-C(n)" means n neighbors are used for distance constraint [4].



(a) Noisy Faces
(b) Linear PCA (Preserve 95% energy)
(c) Least Square Minimization
(d) Distance Constraint, D-C (15)
(e) Regularized Locality Preserving Learning, R-LPL(0.00001, 5)
(f) Original Faces

Figure 2. Illustration of denoised faces

4.1. Face Denoising

To denoise each noisy face image, we project it onto the KPCA subspace by utilizing the kernel trick in (4). Then we find the pre-image of that projection with different models. Fig. 2 shows some results of the denoised faces. Table 1 shows the mean square error (MSE) of the denoised faces. Our approach performs the best, especially when the number of neighbors used is appropriately small. And it is indicated why we do not recommend $s=1$. We also see that the distance constraint scheme would fail if few neighbors are used. As more neighbors are used, it also plays superior to

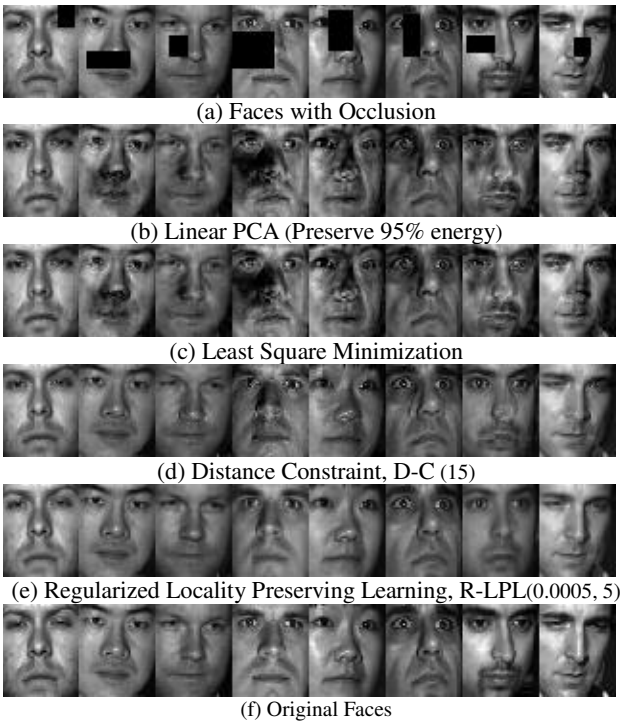
Table 1. Mean square error (MSE) between the original faces and the denoised faces

Method	MSE	Method	MSE
R-LPL(0.00001, 1)	96.3359	D-C (3)	27136.795
R-LPL(0.00001, 3)	80.2613	D-C (5)	15458.2593
R-LPL(0.00001, 5)	**80.2095**	D-C (10)	9735.745
R-LPL(0.00001, 10)	82.9539	D-C (15)	**86.2367**
R-LPL(0.00001, 15)	85.1357	D-C (20)	88.0076
R-LPL(0.00001, 20)	86.7654	D-C (30)	90.6767
R-LPL(0.00001, 30)	89.2883		

Method	MSE
Linear PCA (Preserve 95% energy)	**111.3391**
Least Square Minimization [3]	**112.4266**

Table 2. MSE of denoising with Regularized Locality Preserving Learning (s=5)

Value of λ	MSE	Value of λ	MSE
0.01	81.2285	0.0001	79.7861
0.001	79.8802	0.000001	80.3279
0.0005	**79.6885**	0.0000001	80.3422



(a) Faces with Occlusion
(b) Linear PCA (Preserve 95% energy)
(c) Least Square Minimization
(d) Distance Constraint, D-C (15)
(e) Regularized Locality Preserving Learning, R-LPL(0.0005, 5)
(f) Original Faces

Figure 3. Illustration of reconstructed occluded faces

the least square minimization scheme. However, our approach still achieves the notably smallest MSE. Interestingly, it seems Linear PCA performs a little better than Least Square Minimization scheme. However, Least Square Minimization scheme may achieve local optimization. And it was also declared sometime Linear PCA did better [3]. Finally, table 2 shows the performances if λ equals different values. We see setting λ appropriately small is recommended.

4.2. Reconstruction of Occluded Face

Similarly, each occluded face is projected onto the KPCA subspace for reconstruction. Then the pre-image is found respectively with different models. Fig. 3 shows some results. Table 3 shows the MSE results, and Table 4 shows how different values of λ have effect on MSE. We see that our approach still obtains the smallest MSE and achieves better visual result.

5. Conclusion and Feature Work

This paper poses a novel perspective to the pre-image learning in KPCA and has demonstrated a novel approach, called regularized locality preserving learning that absorbs the spirit of manifold learning with regularization technique. The proposed approach requires no iteration and avoids numerical instability with unique solution. Experimental results show much improvement under the measurement of MSE. In future, we attempt to improve the technique when *s* is small since we experimentally find some reconstructed occluded faces are somewhat not (similar to) the real persons. It could be seen by the 7th person from the left side in fig. 3. It may be mainly because KPCA is unsupervised. Nonetheless, our approach has been indicated as a effective way for pre-image learning.

Acknowledgement

This project was supported by the National Natural Science Foundation of China under Grant No. 60373082, the Key (Key grant) Project of Chinese Ministry of Education under Grant No. 105134

References

[1]. M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," IEEE TPAMI, vol. 12, no. 1, pp. 103–108, Jan. 1990.
[2]. B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comp., vol. 10, pp.1299–1319, 1998.
[3]. S. Mika, B. Schölkopf, A. Smola, K. R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," NIPS, 1998.
[4]. J. T. Kwok and I. W. Tsang, "The Pre-Image Problem in Kernel Methods," IEEE Trans. on Neural Networks, vol. 15 no. 6, pp. 1517-1525, Nov. 2004.
[5]. G. H. Bakır, J. Weston and B. Schölkopf, "Learning to Find Pre-Images", NIPS, 2004
[6]. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, pp. 2323–2326, 2000.
[7]. A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose", IEEE TPAMI, vol. 23, no 6, pp. 643-660, June 2001.
[8]. C. J. C. Burges, "Simplified support vector decision rules," ICML, 1996, pp. 71–77.

Table 3. Mean square error (MSE) between the original faces and the reconstructed occluded faces

Method	MSE	Method	MSE
R-LPL (0.0005 , 1)	50.6897	D-C (3)	8982.5537
R-LPL (0.0005 , 3)	**45.3118**	D-C (5)	4596.9681
R-LPL (0.0005 , 5)	46.4477	D-C (10)	1243.0809
R-LPL (0.0005 , 10)	49.7993	D-C (15)	**63.6917**
R-LPL (0.0005 , 15)	52.3346	D-C (20)	66.8337
R-LPL (0.0005 , 20)	54.3451	D-C (30)	73.6555
R-LPL (0.0005 , 30)	58.0042		

Method	MSE
Linear PCA (Preserve 95% energy)	**140.2855**
Least Square Minimization	**137.906**

Table 4. MSE of reconstructed occluded faces with Regularized Locality Preserving Learning, (*s*=5)

Value of λ	MSE	Value of λ	MSE
0.01	50.8667	0.005	49.7615
0.001	47.1038	0.0001	**46.8279**