Weakly Supervised Learning on Pre-image Problem in Kernel Methods

Wei-Shi Zheng[1,3,5], Jian-Huang Lai[2,3,6] and Pong C. Yuen[4,7]
[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]*Department of Computer Science, Hong Kong Baptist University, Hong Kong*
[5]*wszheng@ieee.org*, [6]*stsljh@mail.sysu.edu.cn*, [7]*pcyuen@comp.hkbu.edu.hk*

Abstract

This paper presents a novel alternative approach, namely weakly supervised learning (WSL), to learn the pre-image of a feature vector in the feature space induced by a kernel. It is known that the exact pre-image may typically seldom exist, since the input space and the feature space are not isomorphic in general, and an approximate solution is required in past. The proposed WSL, however, would find an appropriate rather than only a purely approximate solution. WSL is able to involve some weakly supervised prior knowledge into the study of pre-image. The prior knowledge is weak and no class label of the sample is required, providing only information of positive class and negative class which should properly depend on applications. The proposed algorithm is demonstrated on kernel principal component analysis (KPCA) with application to illumination normalization and image denoising on faces. Evaluations of the performance of the proposed algorithm show notable improvement as comparing with some well-known existing approaches.

1. Introduction

In the last decade, kernel learning has been attractive in pattern recognition and machine learning [1]. Utilizing the kernel trick, a nonlinear mapping $\phi(\cdot)$ that maps data from input space $\mathbf{X}$ into a feature space $\mathbf{H}$ is implicitly induced by the kernel function $k(\cdot,\cdot)$ satisfies $k(\mathbf{x},\mathbf{y}) = <\phi(\mathbf{x}),\phi(\mathbf{y})>$, where $\mathbf{x},\mathbf{y} \in \mathbf{X}$ and $<\cdot,\cdot>$ is the inner product. It gives rise to drive nonlinear analysis on patterns without explicitly defining ϕ.

The pre-image problem [2] in kernel methods is interesting and is useful in some applications, such as image denoising [2] and image compression [5]. And what we are interested in is to find a pre-image in $\mathbf{X}$ for $\mathbf{Pro}(\phi(\mathbf{x}))$, where $\mathbf{x} \in \mathbf{X}$ and $\mathbf{Pro}(\phi(\mathbf{x}))$ is the projection of $\phi(\mathbf{x})$ onto some linear subspace defined in $\mathbf{H}$. Take KPCA and denoising for example. For some noisy pattern $\mathbf{x} \in \mathbf{X}$, $\mathbf{Pro}(\phi(\mathbf{x}))$ is the projection of $\phi(\mathbf{x})$ onto the kernel principal component subspace for denoising. However, $\mathbf{Pro}(\phi(\mathbf{x}))$ is defined in $\mathbf{H}$ so we want to recover its denoised pattern in $\mathbf{X}$. Ideally, the pre-image problem finds an exact pattern $\hat{\mathbf{x}}_0 \in \mathbf{X}$ such that $\phi(\hat{\mathbf{x}}_0) = \mathbf{Pro}(\phi(\mathbf{x}))$. The most difficulty of the pre-image problem, however, is that the input space and the feature space are not isomorphic in general and the feature space always holds higher dimensionality. So from the mathematical aspect, an exact value $\hat{\mathbf{x}}_0$ as the pre-image of $\mathbf{Pro}(\phi(\mathbf{x}))$ would not exist. Some existing efforts settle for an approximate solution, such as the least square distance minimization (LSDM) [2] scheme (or more general technique in [8]) and the distance constraint (DC) scheme [4] which has similar idea with multidimensional scaling (MDS)[4], etc.

In essence, we always hold the question: “What is the best pre-image?” Actually, approximation should not be the sole strategy [5]. Rather than only achieving a purely approximate solution of the pre-image, for pattern recognition and image processing, we may be more interested in learning an appropriate pre-image, resulting in better classification performance or more satisfactory visual result. Moreover, previous methods are unsupervised, and may not achieve good performance if the application is challenging, e.g. the illumination normalization shown in our experiments.

Our idea is to integrate some weakly supervised prior knowledge to find a more appropriate solution of the pre-image depending on the application. Based on this idea, we would propose a novel approach, namely weakly supervised learning (WSL), to learn the pre-image that is appropriate but also approximate by integrating some weak prior knowledge, including only positive class and negative class. Definitions of them depend on different applications. No class label of any sample is in fact required. In the experiment, we extend the application of the pre-image learning in KPCA to face illumination normalization, where no physical models of the illumination and shape of the face are required. Moreover, a denoising experiment has also been presented. The experimental results show the proposed algorithm achieves notable improvement against the previous approaches on pre-image learning. The exposition would focus on the RBF kernel, and the idea can be also well extended to other kernels.

2. Weakly Supervised Learning on Pre-image

Before demonstrating our idea, suppose $\mathbf{x}_1,\cdots,\mathbf{x}_N \in \mathbf{X}$ are N training samples and $\mathbf{\Phi}=(\phi(\mathbf{x}_1),\cdots,\phi(\mathbf{x}_N))$. In this paper we mainly consider the RBF kernel, $k(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/c), c>0$. We denote any linear transform matrix by $\mathbf{U}^{\phi^T}=(\mathbf{u}_1^\phi,\cdots,\mathbf{u}_q^\phi)^T$ defined in the Reproducing Kernel Hilbert Space (RKHS) $\mathbf{H}$ by:

$$\mathbf{U}^\phi=\mathbf{\Phi}(\mathbf{p}_1,\cdots,\mathbf{p}_q)=\mathbf{\Phi}\mathbf{P}\,,\ \mathbf{P}=(\mathbf{p}_1,\cdots,\mathbf{p}_q)\in\mathbf{R}^{N\times q} \qquad (1)$$

It is valid with the Representer Theorem of RKHS [1]. For KPCA, $\mathbf{U}^{\phi^T}$ is the transform of the principal component subspace in RKHS [2-4]. Then $\forall \mathbf{x}\in\mathbf{X}$, the projection of $\phi(\mathbf{x})$ onto the space $span\{\mathbf{u}_1^\phi,\cdots,\mathbf{u}_q^\phi\}$ is:

$$\mathbf{Pro}(\phi(\mathbf{x}))=\mathbf{U}^\phi\mathbf{U}^{\phi^T}\phi(\mathbf{x})=\mathbf{\Phi}\mathbf{P}\mathbf{P}^T\mathbf{\Phi}^T\phi(\mathbf{x})=\mathbf{\Phi}\boldsymbol{\gamma}^{\mathbf{x}} \qquad (2)$$

where $\boldsymbol{\gamma}^{\mathbf{x}}=(\gamma_1^{\mathbf{x}},\cdots,\gamma_N^{\mathbf{x}})^T=\mathbf{P}\mathbf{P}^T\mathbf{\Phi}^T\phi(\mathbf{x})$.

Now, we introduce the weakly supervised learning (WSL) to learn an appropriate pre-image of $\mathbf{Pro}(\phi(\mathbf{x}))$ by integrating weakly supervised prior knowledge. We at the moment assume we have such knowledge and establish the theoretical fundamental. We define the positive class which the pre-image is expected close to and the negative class which the pre-image is expected far away from. Denote the positive class by $C^+=\{\mathbf{z}_1^+,\cdots,\mathbf{z}_{N_+}^+\}$ and the negative class by $C^-=\{\mathbf{z}_1^-,\cdots,\mathbf{z}_{N_-}^-\}$, where $\mathbf{z}_i^+$ and $\mathbf{z}_j^-$ are samples in the positive class and negative class respectively. We do not restrict $\mathbf{z}_i^+$ or $\mathbf{z}_j^-$ to be out of the samples used to produce $\mathbf{U}^{\phi^T}$, such as the transform of KPCA. Specific values of them would be designed properly depending on applications later.

The key idea of our learning would drive the pre-image close to the local positive class information and far away from the local negative class information as a constraint added to the least square distance minimization criteria. Take illumination normalization for example. $\mathbf{Pro}(\phi(\mathbf{x}))$ is the projection of $\phi(\mathbf{x})$ onto the kernel principal component space to attenuate illumination. If the positive class holds faces with nice illumination and the negative class is with the bad ones, then $\mathbf{Pro}(\phi(\mathbf{x}))$ should be expected towards the faces with nice illumination and away from the bad ones in the feature space $\mathbf{H}$. So if $\mathbf{Pro}(\phi(\mathbf{x}))$ has θ^+ nearest positive samples $\phi(\mathbf{z}_{i_1}^+),\cdots,\phi(\mathbf{z}_{i_{\theta^+}}^+)$ and θ^- nearest negative samples $\phi(\mathbf{z}_{i_1}^-),\cdots,\phi(\mathbf{z}_{i_{\theta^-}}^-)$ in $\mathbf{H}$, then our learning aims to learn an appropriate solution $\hat{\mathbf{x}}_0$, the pre-image of $\mathbf{Pro}(\phi(\mathbf{x}))$ drawn towards $\mathbf{z}_{i_1}^+,\cdots,\mathbf{z}_{i_{\theta^+}}^+$ and far away from $\mathbf{z}_{i_1}^-,\cdots,\mathbf{z}_{i_{\theta^-}}^-$ in the input space as a constraint added to least square distance minimization criteria such that both approximate and appropriate solution could be approached meanwhile. The idea is valid in the ideal case that $\mathbf{Pro}(\phi(\mathbf{x}))$ has its exact pre-image $\hat{\mathbf{x}}_0$. Since for RBF kernel (or more general isotropic kernel [4]), $\forall \mathbf{z}\in\mathbf{X}$, $\phi(\mathbf{z})$ is close to (far from) $\mathbf{Pro}(\phi(\mathbf{x}))$ if and only if $\mathbf{z}$ is close to (far from) $\hat{\mathbf{x}}_0$. However an exact pre-image seldom exists. So WSL drives a constraint to find $\hat{\mathbf{x}}_0$ that preserves the local relationship with the positive and negative class in the input space as $\mathbf{Pro}(\phi(\mathbf{x}))$ does in the feature space. In summary, WSL learns the criterion below:

$$\begin{cases} \hat{\mathbf{x}}_0=\arg\min_{\hat{\mathbf{x}}\in\mathbf{X}}\{F_0+F^+-F^-\} \\ \quad F_0=\left\|\phi(\hat{\mathbf{x}})-\mathbf{Pro}(\phi(\mathbf{x}))\right\|^2 \\ \quad F^+=\eta^+\left|H_\theta^+(\mathbf{x})\right|^{-1}\ (\sum_{\mathbf{z}^+\in H_\theta^+(\mathbf{x})}\left\|\hat{\mathbf{x}}-\mathbf{z}^+\right\|^2) \\ \quad F^-=\eta^-\left|H_\theta^-(\mathbf{x})\right|^{-1}\ (\sum_{\mathbf{z}^-\in H_\theta^-(\mathbf{x})}\left\|\hat{\mathbf{x}}-\mathbf{z}^-\right\|^2) \end{cases} \qquad (3)$$

where $H_\theta^+(\mathbf{x})=\{\mathbf{z}^+ \mid \mathbf{z}^+\in C^+,\phi(\mathbf{z}^+)$ is one of the θ^+ nearest positive samples of $\phi(\mathbf{x})$ } and $H_\theta^-(\mathbf{x})=\{\mathbf{z}^- \mid \mathbf{z}^-\in C^-$, $\phi(\mathbf{z}^-)$ is one of the θ^- nearest negative samples of $\phi(\mathbf{x})$ }. And $\eta^+>0$, $\eta^->0$. $|H_\theta^+(\mathbf{x})|$ and $|H_\theta^-(\mathbf{x})|$ are the size of $H_\theta^+(\mathbf{x})$ and $H_\theta^-(\mathbf{x})$ respectively.

To get optimization, we first note that:

$$F_0=\left\|\phi(\hat{\mathbf{x}})\right\|^2+\left\|\mathbf{Pro}(\phi(\mathbf{x}))\right\|^2-2<\mathbf{Pro}(\phi(\mathbf{x})),\phi(\hat{\mathbf{x}})> \qquad (4)$$

As declared, we consider the RBF kernel such that $k(\mathbf{x}_i,\mathbf{x}_j)=<\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)>=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/c)$, then

$$\partial F_0/\partial\hat{\mathbf{x}}=\partial(-2<\mathbf{Pro}(\phi(\mathbf{x})),\phi(\hat{\mathbf{x}})>)/\partial\hat{\mathbf{x}}=\partial(-2<\mathbf{\Phi}\boldsymbol{\gamma}^{\mathbf{x}},\phi(\hat{\mathbf{x}})>)/\partial\hat{\mathbf{x}}$$

$$<\mathbf{\Phi}\boldsymbol{\gamma}^{\mathbf{x}},\phi(\hat{\mathbf{x}})>=\sum_{i=1}^N\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}})=\sum_{i=1}^N\gamma_i^{\mathbf{x}}\exp(-\|\mathbf{x}_i-\hat{\mathbf{x}}\|^2/c)$$

So $\partial F_0/\partial\hat{\mathbf{x}}=4c^{-1}\sum_{i=1}^N\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}})(\hat{\mathbf{x}}-\mathbf{x}_i)$. Also, we have:

$$\partial F^+/\partial\hat{\mathbf{x}}=2\eta^+(\hat{\mathbf{x}}-\tilde{\mathbf{z}}_{H_\theta^+(\mathbf{x})})\,,\partial F^-/\partial\hat{\mathbf{x}}=2\eta^-(\hat{\mathbf{x}}-\tilde{\mathbf{z}}_{H_\theta^-(\mathbf{x})}) \quad (5)$$

$$\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t - \frac{2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}}_t)(\hat{\mathbf{x}}_t - \mathbf{x}_i) + \eta^+(\hat{\mathbf{x}}_t - \tilde{\mathbf{z}}_{H_\theta^+(\mathbf{x})}) - \eta^-(\hat{\mathbf{x}}_t - \tilde{\mathbf{z}}_{H_\theta^-(\mathbf{x})})}{2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}}_t) + \eta^+ - \eta^-} = \hat{\mathbf{x}}_t - \rho_t \frac{\partial(F_0 + F^+ - F^-)}{\partial\hat{\mathbf{x}}}\Big|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_t} \quad (8)$$

where $\tilde{\mathbf{z}}_{H_\theta^+(\mathbf{x})} =| H_\theta^+(\mathbf{x}) |^{-1} \sum_{\mathbf{z}^+ \in H_\theta^+(\mathbf{x})} \mathbf{z}^+$ is the mean of local positive samples, $\tilde{\mathbf{z}}_{H_\theta^-(\mathbf{x})} =| H_\theta^-(\mathbf{x}) |^{-1} \sum_{\mathbf{z}^- \in H_\theta^-(\mathbf{x})} \mathbf{z}^-$ is the mean of local negative samples. Hence we have

$$\partial(F_0 + F^+ - F^-)/\partial\hat{\mathbf{x}} = \frac{4}{c}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}})(\hat{\mathbf{x}} - \mathbf{x}_i) + 2\eta^+(\hat{\mathbf{x}} - \tilde{\mathbf{z}}_{H_\theta^+(\mathbf{x})}) - 2\eta^-(\hat{\mathbf{x}} - \tilde{\mathbf{z}}_{H_\theta^-(\mathbf{x})})$$

To optimize, let $\partial(F_0 + F^+ - F^-)/\partial\hat{\mathbf{x}} = \mathbf{0}$. We have:

$$\hat{\mathbf{x}} = \frac{2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}})\mathbf{x}_i + \eta^+\tilde{\mathbf{z}}_{H_\theta^+(\mathbf{x})} - \eta^-\tilde{\mathbf{z}}_{H_\theta^-(\mathbf{x})}}{2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}}) + \eta^+ - \eta^-} \quad (6)$$

Like [1], one can solve $\hat{\mathbf{x}}$ by the fixed-point iteration

$$\hat{\mathbf{x}}_{t+1} = \frac{2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}}_t)\mathbf{x}_i + \eta^+\tilde{\mathbf{z}}_{H_\theta^+(\mathbf{x})} - \eta^-\tilde{\mathbf{z}}_{H_\theta^-(\mathbf{x})}}{2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}}_t) + \eta^+ - \eta^-} \quad (7)$$

Or we have an essence view from (8) on the top, where $\rho_t = 0.5\times(2c^{-1}\sum_{i=1}^{N}\gamma_i^{\mathbf{x}}k(\mathbf{x}_i,\hat{\mathbf{x}}_t) + \eta^+ - \eta^-)^{-1}$, so that the iteration is a special case of the gradient descent optimization scheme in essence. Therefore, a more general practical solution to get the local optimization could be drawn by the gradient descent scheme like

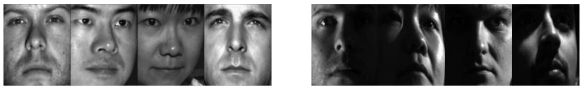
$$\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t - \rho_t^* \frac{\partial(F_0 + F^+ - F^-)}{\partial\hat{\mathbf{x}}}\Big|_{\hat{\mathbf{x}}=\hat{\mathbf{x}}_t} \quad (9)$$

for some suitable $\rho_t^* > 0$ at each step. Actually, (9) is a general scheme suitable for any kernel but not only for RBF and (7) is also enough in our experiment. In the experiment, the mean value of all training samples is set to be the initial value of $\hat{\mathbf{x}}_t$ for the iteration. This setting is the same for LSDM [2] for a fair comparison.

From (7), we see that, as long as given the appropriate prior knowledge, $\hat{\mathbf{x}}_{t+1}$ is drawn towards the appropriate local prior information of positive class and pushed away some local prior information of negative class at each step. Finally we would get an appropriate pre-image depends on the application meanwhile achieving the approximation. However, θ^+ , θ^-, η^+ and η^- should be carefully set so that the prior knowledge helps learn an appropriate pre-image since approximation is also important in our algorithm.

Finally, the learning is weakly supervised since it only integrates the prior information involving the positive class and the negative class. In fact, the proper class label of each sample is not required. It provides a potentially wide application of the algorithm.

Although we have driven an iterative scheme for the solution, however, experiment results show much improvement of our approach comparing with well-known learning methods on pre-image problem.



(a) positive class (b) negative class

Figure 1. Examples (images are resized to present)

3. Applications

Pre-image learning could provide wide applications. We herein present two applications, namely the illumination normalization and face image denoising. All experiments are based on YALEB [7] database. It has 10 persons with 9 different poses. Each pose of each person owns 65 faces with different illumination conditions that are always divided into 5 subsets [7]. We select face images from all persons with all poses in the first 4 subsets, totally 4050 images, to evaluate the performance of the algorithms. All images are aligned with the size 92×112 . The gray value of pixel of images is finally stretched ranging from 0 to 1. The parameters in the proposed algorithm are manually and experimentally set to get better performance.

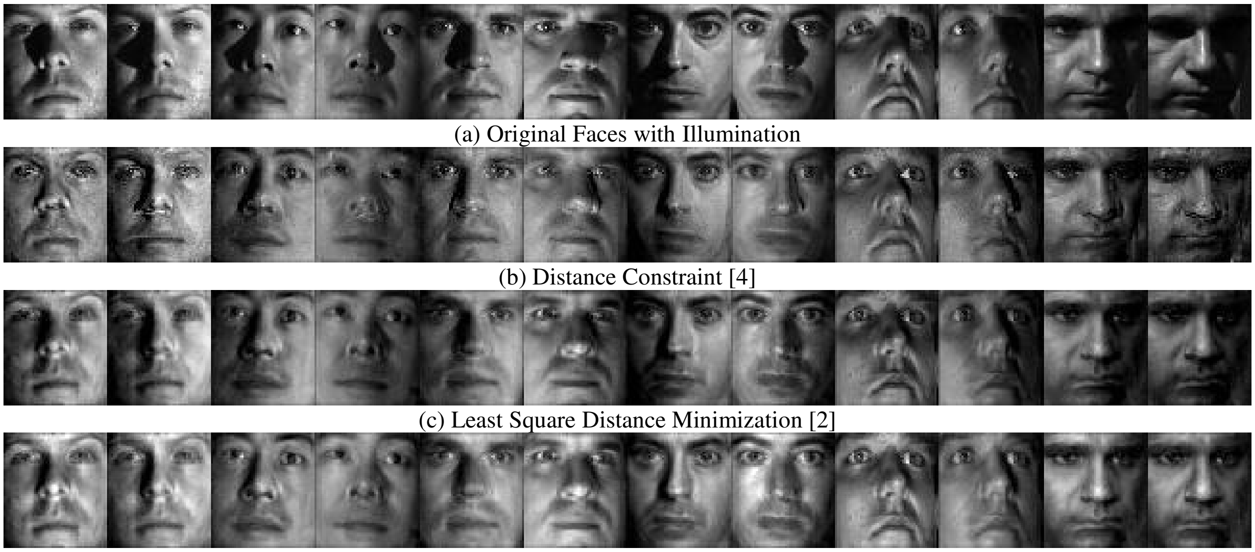
3.1. Illumination Normalization

For illumination normalization, as aforementioned in section 2, we define the positive class holds face images with nice illumination while the negative class holds face images with bad one. In the experiment, we let the positive class be all images of all persons with different poses in the subset 1. Technically speaking, faces with bad illumination could be simulated. For convenience, the negative class is produced from the subset 5 (Fig. 1), where 7 images of each pose of each person are randomly selected. Therefore the total size of the positive class and negative class are 630 respectively. For all algorithms, the kernel principal component subspace is produced from all images in subset 1 that holds the face images with nice illumination with 95% energy maintained. As declared, we do not restrict samples in the positive class or negative class to be out of the samples for training to produce the kernel principal component subspace.

Utilizing the kernel trick, each testing image in subsets 2~4 is first projected onto the kernel principal component subspace to attenuate illumination. Then

Table 1. Average recognition accuracy ($c = 10^5$)

Method	Subset2	Subset3	Subset4
WSL	99.81%	89.63%	48.73%
DC [4]	99.35%	76.02%	30.79%
LSDM [2]	99.44%	81.20%	37.22%



(a) Original Faces with Illumination

(b) Distance Constraint [4]

(c) Least Square Distance Minimization [2]

(d) Weakly Supervised Learning

Figure 2. Illustration of illumination normalization of subset 3 (images have been resized to present and $c=10^5$)

pre-image learning algorithms are used to learn the pre-image of that projection respectively. In the experiment, we set $c=10^5$ in the RBF kernel and set $\theta^+=\theta^-$=50, $\eta^+=\eta^-$=0.00001. The reason why η^+ and η^- are small is because $c\times\eta^+$ and $c\times\eta^-$ should be appropriately small in (7). And 30 neighbors are used to find the pre-image in distance constraint method. Fig. 2 shows some results of the pre-images.

Table 1 gives the average recognition rate on the reconstructed pre-images of different subsets respectively. For each subset, the recognition rate is obtained by averaging the recognition rates over all poses in this subset. For the recognition of each pose, LDA [6] is implemented where all the first three faces of the same pose from subset 1 are used for training, since they are original images under almost normal illumination condition without any transformation.

We see that, in subsets 3~4, WSL achieves notable improvement against the compared pre-image learning algorithms. For subset 3, the recognition rates of the other two algorithms sharply reduce. The illumination condition in subset 4 is quite challenging. Though all algorithms do not achieve satisfied results, WSL still gets significant improvement against the others. Note that KPCA is not a specific algorithm for illumination normalization. Moreover our learning does not depend on any physical model and any information of the shape of the face. However, it may be an interesting task to improve KPCA for illumination normalization.

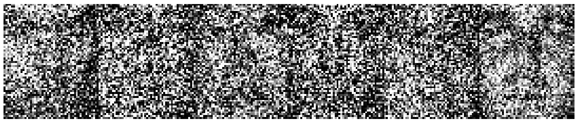
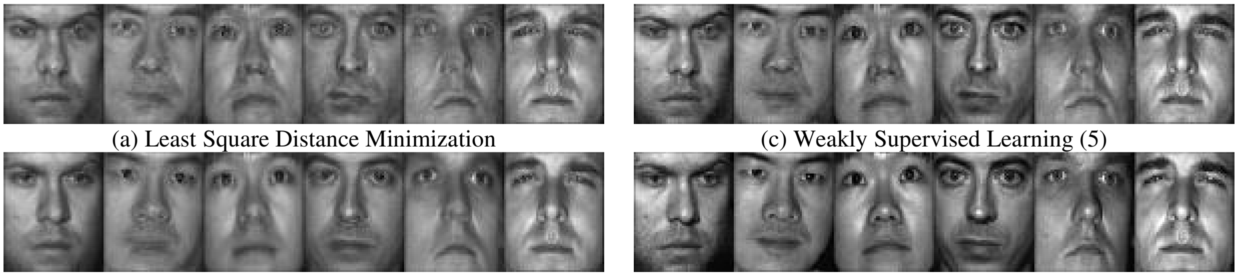


Figure 3. Illustration of noisy faces

3.2. Face Image Denoising

First, the positive class is the same as that in section 3.1. Secondly we produce 1890 (=7 × 9 × 10 × 3) noisy face images by producing 3 different noisy images from each face image in subset 1, where the noise type is Gaussian with mean 0 and random variance falling in (0,0.5] (Fig. 3). Thirdly the negative class is established by selecting 630 noisy images produced, where each noisy image has one-one correspondence with its original image in the positive class. Then the other 1260 images are tested for denoising. Note that all noisy images are also linearly stretched to full range of pixel values of [0, 1] before the experiments.

For all algorithms, the kernel principal component subspace is developed the same as the previous application. And all testing noisy images are projected onto the kernel principal component subspace for denoising. Then the pre-images of the projections are found by the algorithms respectively. In the experiment the parameter settings are: for $c=10^5$, set η^+ =0.00002 and η^- =0.00001; for $c=10^4$, set η^+ =0.0002 and η^- =0.0001. Tables 2~3 show the mean square error (MSE) between the denoised images and the original images. The notations in the tables are explained as: "WSL(n_0)" means that $\theta^+=\theta^-=n_0$ and "DC(n_0)" indicates that the number of neighbors used is set to be n_0 using distance constraint [4]. Due to the limited length, table 3 only gives main results. Fig. 4 shows some denoised images. Apparently, WSL and DC are able to get more excellent improvement than LSDM. It also coincides with [4] that reported DC was



(a) Least Square Distance Minimization

(c) Weakly Supervised Learning (5)

(b) Distance Constraint (15)

(d) Original Faces without Noise

Figure 4. Illustration of denoised faces shown in figure 3, (images have been resized to present and $c = 10^5$)

superior to LSDM for denoising. Here we also find interesting results that WSL would become better if θ^+ and θ^- are set to be small, while DC would fail if the number of neighbors set is quite small. This is also intuitive in the extreme case that DC is unstable if only one neighbor is used. Hence, WSL achieves the best.

4. Further Discussion

Parameters. Parameters in WSL have their own meanings. Basically, θ^+ and θ^-, which indicate how many local nearest positive and negative samples would be considered, could actually be treated as regularization parameters. As θ^+ or θ^- becomes large, the information of the positive or negative class becomes global, vice versa becomes local. This is well observed during the denoising experiment. As more neighbors are considered, the denoised image would become smoother, but may be away from the original image. That is why for denoising fewer neighbors are suggested. However, for illumination, since the reflection of illumination is more challenging and more complex than noise, here Gaussian noise, hence a properly large value of θ^+ and θ^- should be significant for the performance. Besides θ^+ and θ^-, values of η^+ and η^- are also important and they have been discussed in section 2. If η^- and θ^- become zero and the positive class is set to be all training samples, then WSL could be understood as a locality preserving constraint. We believe adaptive estimations of them would further enhance the performance of WSL.

Prior knowledge. Prior knowledge is important and could be simulated. So it may be possible to achieve as better performance as given more prior information.

5. Conclusion and Future Work

This paper proposes a novel alternative approach for the pre-image learning in kernel methods, namely weakly supervised learning. Rather than only finding a purely approximate solution in past, we integrate weakly supervised information to find a solution that is appropriate for the application and also approximate. Therefore, this paper actually suggests a way to integrate prior information into pre-image learning in kernel methods. Our feature work would attempt to find a non-iterative scheme for the idea of WSL.

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]. B. Schölkopf and A. J. Smola, Learning with Kernels. Cambridge, MA: MIT Press, 2002.

[2]. S. Mika, B. Schölkopf, A. Smola, K. R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," NIPS, 1998.

[3]. B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comp., vol. 10, pp.1299–1319, 1998.

[4]. J. T. Kwok and I. W. Tsang, "The Pre-Image Problem in Kernel Methods," IEEE TNN, 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]. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs.fisherfaces: Recognition using class specific

Table 2. Mean square error (MSE) between original image and denoised image ($c = 10^5$)

Method	MSE	Method	MSE
WSL (3)	**42.2881**	DC (5)	987.0362
WSL (5)	42.5065	DC (10)	874.7932
WSL (10)	45.0719	DC (15)	**43.8077**
WSL (20)	50.7291	DC (20)	44.1442
WSL (30)	52.723	DC (30)	46.5774
WSL (50)	55.3098	DC (50)	49.7544
LSDM	**62.2306**		

Table 3. Mean square error between original image and denoised image ($c = 10^4$)

Method	MSE
Weakly supervised Learning (WSL) (5)	43.0745
Distance Constraint (DC) (15)	46.2966
Least Square Distance Minimization (LSDM)	66.3746

linear projection," IEEE TPAMI, vol. 19, no. 7, pp. 711–720, July 1997.

[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.