Perception Preserving Projections

Saining Xie[1]
xiesaining@gmail.com
Jiashi Feng[2]
a0066331@nus.edu.sg
Shuicheng Yan[2]
eleyans@nus.edu.sg
Hongtao Lu[1]
htlu@sjtu.edu.cn

[1] Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China

[2] Department of Electrical and Computer Engineering
National University of Singapore
Singapore

Abstract

Linear projection for reducing data dimensionality is a common practice in various data processing applications. Among the existing projection methods, Principal Component Analysis (PCA) is arguably the most popular one. Standard PCA used in image preprocessing pursues the projection directions by minimizing the reconstruction error in a least square sense. However, since PCA does not adapt to the data or any specific domains, it may lead to severe loss of certain discriminative features during the projection, and damage the performance of either human perception (e.g. stimulus in the visual cortex, as modeled by Gabor wavelets), or machine perceptions (e.g. recognizing the images based on a certain type of visual features), or both. In this paper, we propose a novel Perception Preserving Projections (PPP) method to preserve the information for specific perception systems. In particular, PPP incorporates domain-specific feature extractor into the standard PCA formulation for the projection learning procedure. This enables PPP to make more sensible projections for feature based perception systems while retaining the simplicity and unsupervised manner of PCA. In experimental studies, PPP shows clear effectiveness and improvement over PCA in terms of two performance metrics: feature extraction deviation and the pattern recognition accuracy.

1 Introduction

Unsupervised learning of feature projections to low-dimensional linear subspaces from training data has become a standard paradigm in the areas of pattern recognition and computer vision. Principle component analysis (PCA) [9] in particular is one of the most popular projection learning methods for dimensionality reduction and feature extraction. PCA in the image preprocessing procedure allows one to project the images into a lower dimensional subspace where the reconstruction loss is minimized in a least square sense. The new representations obtained from the projection can then be exploited as descriptive features. These learned PCA representations have proven useful for solving problems such as face and object recognition, tracking, detection, and background modeling. For example, PCA has become one of the most successful approaches in face recognition [22]. Classifiers (e.g., kNN) or linear discriminators (e.g., LDA [2] and MFA [28]) are performed in the PCA-projected spaces to recognize the images.

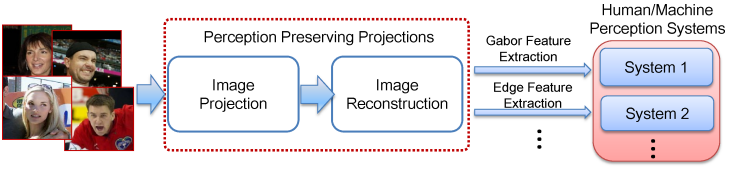


Figure 1: The red rectangle highlights the problem we aim to solve.

Compared with the descriptive features learned from the raw intensity domain, many advanced features work better either for machine perception (pattern recognition accuracy) [23], or human perception (physical interpretability) [6], or both [15] [30]. In those cases, PCA still serves as an indispensable component in data pre-processing, due to the considerable advantages gain including background noise reduction, data de-correlation and storage/computational efficiency [21] [8].

However, the standard PCA is not tuned to any domain-specific features, thus the discriminative features for certain perception system (such as feature-based automatic face recognition system) may be lost in the projection process. For example, the discriminative edges of objects in the images may be destroyed after the PCA projection and reconstruction. Such loss of discriminative information is dependent on data and specific features used in different perception systems. For example, for face recognition, the Gabor feature is important, while for object detection, the gradient feature is more crucial. This required adaptation in feature domain makes PCA hardly give the sensible projections and preserve informative features. In this work, we investigate how we can pursue a linear projection method that is able to prevent significant machine/human perception loss (i.e. feature extraction deviations before and after projection and reconstruction).

In particular, to preserve the desired feature characteristic for projected images, we propose a perception preserving projection (PPP) method. PPP pursues a set of suitable projection basis which minimize the reconstruction loss for both the original images and the extracted features simultaneously. Here the feature extractor is formulated as a general linear operator and integrated in the loss function explicitly.

In this way, by minimizing such reconstruction loss, we can obtain a set of projection basis preserving certain type of perceptions(features). To efficiently solve the induced optimization problem, we further introduce a low-rank relaxation to the original orthogonal constraint on the projection basis, which offers a more efficient and robust solution.

To quantitatively evaluate the machine perception preserving capability of PPP, we experimentally evaluate the performance of face recognition on the reconstructed face images from both PCA and PPP. It is shown that on the face images reconstructed by PPP, face recognition can achieve significantly improved performance, even in much lower dimensions compared to PCA. The results demonstrate that the proposed PPP method can indeed preserve the feature characteristic of the data.

The remainder of the paper is organized as follows. The formulation of the proposed PPP method is given in Section 2. In Section 3 we introduce the optimization algorithm. Section 4 gives the experiment results. In Section 5, we discuss several related works. Finally we conclude the work in Section 6.

2 Problem Formulation

In this section, we introduce the description of our problem formally and then provide corresponding solutions in the next section. Through out the paper, we assume the data points are stacked in the matrix $X = [x_1, \cdots, x_n] \in \mathbb{R}^{d \times n}$ column-wisely. Here n is the number of

samples. And d is the ambient dimension of the observed data points. Denote the projection basis as $U = [u_1, \cdots, u_r] \in \mathcal{R}^{d \times r}$. Here r is the dimension of subspace.

2.1 PCA Revisiting

Principal Component Analysis (PCA) is a widely used data projection method, which projects a set of data vectors onto a set of uncorrelated principal components computed from the data observations. More specifically, the objective of PCA is to minimize following loss function:

$$\min_{U^T U = I_r} \mathcal{L}(U) = \|X - UU^T X\|_F^2. \tag{1}$$

Namely we want to find a set of r orthogonal basis along which the reconstructed data approximate the original data well. Though the optimization problem (1) is not convex, a global optimal solution is obtained in closed-form by performing eigen-decomposition on the data covariance matrix $XX^T = V\Sigma V^T$, and the solution is given by selecting the leading r eigenvectors from the decomposition. Though PCA is able to minimize the data reconstruction loss in quadratic measure, PCA does not take the following feature extraction step into consideration and may incur severe information loss in specific perception systems.

2.2 PPP Formulation

There are various feature extractors can be seen as linear operators $\mathcal{P}$ over the data vector $\mathbf{x} \in \mathbb{R}^d$ [27]. One example is the convolution of data with linear filter f: $\mathcal{P}(\mathbf{x}) = P_f \mathbf{x} = f * \mathbf{x}$. Widely used linear filters include Gabor filter, Laplacian of Gaussian (LoG) filter. Here P_f denotes the corresponding convolution matrix for certain filter f. Another popular example is pixel-wise "masking": $P\mathbf{x}$, where P is a $d \times d$ diagonal matrix. The third example is the sum of filters $\sum_{k=1}^{K} P_{f_k}\mathbf{x}$, where P_{f_k} represents a matrix constructed for the k-th linear filter f_k. Thus we can unify the above feature extractors as a linear operator $\mathcal{P}$ and directly add it into the PCA objective function. The objective function of PPP can then be formulated as:

$$\min_{U^T U = I_r} \mathcal{L}(U) = \|\mathcal{P}'(X) - \mathcal{P}'(UU^T X)\|_F^2, \tag{2}$$

Where $P' = [(1-\alpha)P, \alpha I_d]$, and α is a trade-off parameter between the original data space and feature space. Note that this formulation is a more general form compared to PCA. When $\alpha = 1$, the objective function degenerates to standard PCA.

The above objective function states that we aim to find a set of projection basis $U \in \mathbb{R}^{d \times r}$, such that the extracted features from the reconstructed data $\mathcal{P}(UU^T X)$ will not deviate from the features extracted from the original data $\mathcal{P}(X)$ too much. Thus the critical feature information specializes to a certain perception system represented by $\mathcal{P}$ is preserved.

3 Optimization

3.1 Optimization on the Stiefel Manifold

The most straightforward method to solve the problem (2) is performing gradient descent on the Stiefel manifold defined by $U^T U = I$ [1]. One off-the-shelf solver using Cayley transformation during the iterative optimization process is proposed recently [25]. We test this method as a baseline, the detailed algorithm for our particular problem and parameter settings are presented in the supplementary material for reference.

We notice that though state-of-the-art algorithms like [25] can be directly exploited, the computational cost is quite high because 1) the gradient computation involves calculating the matrix inverse and 2) the convergence of gradient descent is usually slow. Based on above observations, instead of directly employing the above gradient descent method on the Stiefel manifold, we propose our new objective function for PPP.

3.2 Low-rank Relaxation

Inspired by the Robust PCA work [4], which seeks a low rank matrix to approximate the original data matrix, here we also relax the orthogonal constraint in the objective function and just seek a low rank matrix as the transformation matrix.

Therefore, the objective function in (2) can be relaxed as follows:

$$\min_{W} \|\mathcal{P}(X) - \mathcal{P}(WX)\|_F^2, \text{ s.t. } \operatorname{rank}(W) \leq r, \tag{3}$$

where r is also the pre-defined number of dimension for the data dimension reduction. To see this, we can first perform skinny SVD on the obtained reconstruction matrix $W = U\Sigma V^T$. Since the rank of W is r, the sizes of U, Σ, V are $d \times r, r \times r, d \times r$ respectively. Then in the data dimension reduction, V can serve as the projection matrix which projects the d-dimension data to r-dimension ones. And in the data reconstruction, $U\Sigma$ can be applied.

Since the nuclear norm is the convex envelope of the rank function [17], the above objective function can be further relaxed as:

$$\min_{W} \|W\|_* + \lambda \|E\|_F^2, \text{ s.t. } \mathcal{P}(X) - \mathcal{P}(WX) = E, \tag{4}$$

where E explicitly accommodates the reconstruction errors. Here the Frobenius norm can be replaced by other norms if the prior structure information of E is available. For example, if we know there are some gross noise on the data, we can use ℓ_1-norm to enforce a sparse structure. If there are some outliers, we can use $\ell_{2,1}$-norm to isolate the outliers to some extent. Here, since such prior information is not available, we use the Frobenius norm to ensure the reconstruction error of each data dimension to be small.

The optimization problem in (4) can be solved by the Alternating Direction Method (ADM) [11] efficiently. In particular, we first define the following augmented Lagrangian function:

$$\mathcal{L}(W,E,Y) = \|W\|_* + \lambda \|E\|_F^2 + \langle Y, \mathcal{P}(X) - \mathcal{P}(WX) - E \rangle + \frac{\mu}{2} \|\mathcal{P}(X) - \mathcal{P}(WX) - E\|_F^2 \tag{5}$$

where Y is the Laplacian multiplier accounting for the hard constraint $\mathcal{P}(X) - \mathcal{P}(WX) = E$ and μ is an adaptive penalty parameter. The larger value of μ, the greater penalty imposed on the constraint.

One of the advantages of ADM is that the original optimization problem can be decomposed into several subproblems which are relatively easier to solve. However, since in the current objective function there exists a linear operator $\mathcal{P}(\cdot)$ imposing on variable W, it is difficult to solve the sub-problem for optimizing the function w.r.t. W:

$$\mathcal{L}(W) = \|W\|_* + \langle Y, \mathcal{P}(X) - \mathcal{P}(WX) - E \rangle + \frac{\mu}{2} \|\mathcal{P}(X) - \mathcal{P}(WX) - E\|_F^2 \tag{6}$$

We are confronted with a problem that directly minimizing the subproblem above will lead to solving a discrete-time Sylvester equation $\mathcal{P}(WX) + W = C$ (C is a constant matrix) in each iteration. The computing complexity for solving the equation can be as high as $O(n^6)$, which is infeasible for the ADM algorithm and making the low-rank relaxation rewardless.

3.3 Linearization of the Objective Function

To alleviate such difficulty, we adopt the recently developed Linearized Alternating Direction Method (LADM) [12] and linearize the quadratic term in the above Lagrangian function at the point W^k:

$$\mathcal{L}(W,W_k) = \|W\|_* + \langle Y, -\mathcal{P}(W_k X) \rangle + \mu \left\langle \mathcal{P}^* \left(\mathcal{P}(X) - \mathcal{P}(W^k X) - E \right) X^T, W - W_k \right\rangle + \frac{\mu\eta}{2} \|W - W_k\|_F^2. \tag{7}$$

Here $\eta > (\|\mathcal{P}\|\|X\|)^2$ is the Lipschitz constant of the linear operators imposed on variable W. After several algebra computation, the above objective function can be written as:

$$\mathcal{L}(W,Y,W_k) = \|W\|_* + \frac{\mu\eta}{2} \|W - M_k\|_F^2, \tag{8}$$

where $M_k = W_k - \mathcal{P}^*(\mathcal{P}(X) - \mathcal{P}(W_k X) - E)X^T/\eta + \mathcal{P}^* Y X^T/\mu\eta$. $\mathcal{P}^*$ denotes the adjoint of the operator $\mathcal{P}$, which is defined as $\langle \mathcal{P}(X), Y\rangle = \langle X, \mathcal{P}^*(Y)\rangle$. It is well known that the above objective function has following closed form solution:

$$W_{k+1} = U\mathcal{S}_{\frac{1}{\mu\eta}}(\Sigma)V^T, \tag{9}$$

where U, Σ, V are from SVD on the matrix M_k. And $\mathcal{S}(\cdot)$ is a shrinkage operator defined as $\mathcal{S}_\varepsilon[x] = sgn(x)max(|x| - \varepsilon, 0)$. Here the shrinkage operator is performed element-wisely for the involved matrix. To guarantee a good convergence rate, we adopt following adaptive penalty strategy [12]:

$$\mu_{k+1} = \begin{cases} \rho_0\mu_k, & \text{if } \mu_k \max(\sqrt{\eta}\varepsilon_W, \varepsilon_E)/\|\mathcal{P}(X)\| < \varepsilon_2, \\ \mu_k, & \text{otherwise.} \end{cases} \tag{10}$$

Here $\varepsilon_W = \|W_{k+1} - W_k\|$ and $\varepsilon_E = \|E_{k+1} - E_k\|$. And the stopping criterion is:

$$\|\mathcal{P}(X) - \mathcal{P}(W^k X) - E\|/\|\mathcal{P}(X)\| < \varepsilon_1. \tag{11}$$

The implementation details of the above linearized ADM for PPP is given in Algorithm 1. Here the convergence parameters are fixed as $\varepsilon_1 = \varepsilon_2 = 1 \times 10^{-6}$. The optimization algorithm is guaranteed to be convergent by the following theorem 1. In our implementation, we also adopt partial SVD and rank prediction techniques using PROPACK [10] and represent W as its skinny SVD to avoid full matrix multiplications and thus yielding a complexity of $\mathcal{O}(nd^2)$.

Theorem 1 *If $\{\mu_k\}$ is non-decreasing and upper bounded, $\eta > \|\mathcal{P}\|^2\|X\|^2$, then the sequence $\{(W_k, E_k, Y_k)\}$ generated by Algorithm 1 converges to a KKT point of problem (4).*

Algorithm 1: Linearized ADM for PPP optimization.

Input : X, $\mathcal{P}$, λ, $\mu^0 = 1 \times 10^{-6}$, $\mu_{\max} = 1 \times 10^{10}$, $\rho_0 = 1.1$.
Output: Reconstruction matrix W.

$k = 0$.
repeat
  (S.1) Compute W_k as in (9);
  (S.2) Update $E_{k+1} = (Y_k - \mu(\mathcal{P}(X) - \mathcal{P}(W_k X)))/(2\lambda - \mu_k)$;
  (S.3) Update the multiplier $Y_{k+1} = Y_k + \mu_k(\mathcal{P}(X) - \mathcal{P}(WX) - E)$;
  (S.4) Update μ_k to μ_{k+1} as in (10);
  (S.5) $k \leftarrow k + 1$.
until *Convergence*;

4 Experiments

4.1 Experiment I: Synthetic Data

In this experiment, we investigate whether the proposed PPP method could preserve simple line patterns along one direction in the image reconstruction and immune to the distracter lines along other directions.

Data: We generate 100 gray level patches with size of 16×16 pixels. Among these patches, 70 patches contain only vertical lines and the other 30 patches contain only horizontal lines. The position and intensity of these vertical and horizontal lines are generated randomly. The patches with horizontal lines have distinguished difference from most patches with vertical lines and they can be seen as outliers. Intuitively, due to the sensitiveness of standard PCA to outliers, the reconstructed patches will be contaminated severely by these horizontal outliers. While for PPP projection, we instantiate the linear operator $\mathcal{P}$ as vertical edge detector formed by the convolution kernel $f = [-1, 0, 1]$. And explicitly minimize the loss of such vertical line patterns in the patch reconstruction.

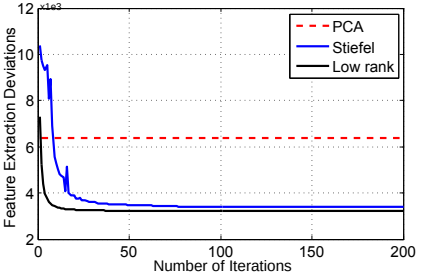


Figure 2: Feature deviation descent curve on the synthetic data along with iterations.

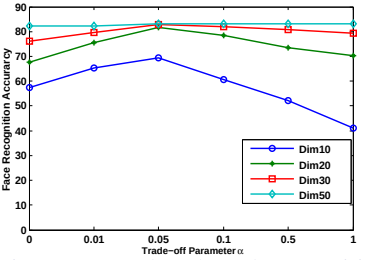
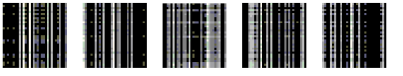
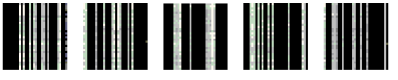


Figure 3: Gabor based recognition performance under different trade-off parameter α and different reconstruction dimensions.

(a) Original images



(b) PCA reconstruction



(c) PPP reconstruction based on gradient descent over Stiefel manifold

(d) PPP reconstruction based on low rank optimization

Figure 4: Examples of the synthetic data and reconstruction results from different methods.

Results: The feature deviation value (i.e. $\|\mathcal{P}(X)-\mathcal{P}(WX)\|_F^2$) is plot versus the number of iterations in Figure 2. From the plot, we can see that both Stiefel manifold and low-rank relaxation can achieve much smaller feature deviations than PCA. And the convergence of low-rank is faster and more smooth than Stiefel manifold gradient descent. The reconstruction results are shown in Figure 4. From the figure, it can be observed that standard PCA completely fails to reconstruct the original patches after projecting the patch vector onto low-dimensional subspace. And the characteristic vertical line patterns of the patch are destroyed completely. However, for the proposed PPP method, we can see that it can successfully preserve the vertical line pattern in the reconstructed data. And the outliers (horizontal line patterns) are suppressed. From the results, we can see clearly that PPP is able to preserve critical perception information. It is worth noting that in this case, even if the outlier lines are discarded mainly because the response of $\mathcal{P}$ over horizontal lines is nearly zero, the perceptual results generated by low-rank approximation are better than those given by Stifel manifold gradient descent. We believe it is because the low-rank relaxation formulation inherits good characteristics such as insensitivity to outliers(at least to some extend), from robust subspace learning algorithms such as LRR [14], which shares a very similar formulation as ours. Thus PPP can produce a much more robust reconstruction results, with significantly less loss of specialized information and contamination of outliers.

4.2 Experiment II: Gradient Preserving

In this experiment, we investigate the capability of the proposed PPP method on gradient feature preserving in the procedure of facial image reconstruction. For comparison, we employ standard PCA as our baseline. To quantitatively evaluate the superiority of PPP over PCA, we conduct face recognition tasks on following two benchmark datasets. The gradient features are extracted from the reconstructed images and then fed into classifiers for classification.

Dataset: In this and following experiments, we employ two benchmark datasets for

evaluation. The first one is the Extended Yale-B face dataset [7]. It contains $16,128$ images of 38 human subjects under 9 poses and 64 illumination conditions. In this experiment, we choose the frontal pose and use all the images under different illumination, thus we get 64 images for each person. All the face images are manually aligned and cropped. The size of each cropped image is 32×32 pixels with 256 gray levels per pixel. The pixel values are scaled to $[0,1]$.

The second dataset is the FRGC face dataset [16]. The dataset consists of $5,658$ face images from 275 persons. The size of the original images is 100×100 pixels. Here, we resize all the images to 32×32 pixels for efficiency. For the training and test split, we randomly divide the images into two equal parts. Namely, the training and test sets contain $2,829$ images respectively.

Experiment settings: In recent computer vision research, features based on gradient are shown to be important for face recognition [23] and object recognition [5]. Thus how to preserve the gradient information in data projection and reconstruction process is also important in real applications. In this experiment, we employ the Laplacian of Gaussian (LoG) as the gradient feature extractor. The kernel size of LoG is set as 15×15 pixels. And the standard deviation σ of the Gaussian kernel is fixed as 1.

As shown in Figure 3, for different projected subspace dimensions, while embedding original intensity domain of PCA is useful to avoid possible overfitting (for extracting meaningful features in face recognition tasks), if alpha is too large, the performance will decrease and finally degenerate to standard PCA results. We empirically find that $\alpha = 0.05$ is good choice in this and following experiments.

For the classification, we adopt two popular classifiers in face recognition. The first one is based on k-NN and here the parameter k is fixed as 1. And the classification result for each test is determined by the voting from its neighboring training images. The second classifier is based on the widely used discriminator Linear Discriminant Analysis (LDA) [2].

Since the discriminative capability of LoG on face recognition is limited. Directly p-reserving LoG feature may not improve the recognition performance too much. Therefore, in the experiments, we adopt a strategy to augment the capability of the linear operator in capturing the discriminative features. In particular, we learn a discriminative feature transformation based on LDA on an extra face dataset. Here we use CMU PIE [19]. Note that such extra knowledge can be obtained from public and thus can be embedded into our proposed PPP projection method.

Results: The experimental results are shown in Figure 6 and Table 1. From the results, we can observe that with the increasing of reconstructed data rank (*i.e.*, the reduced dimension), both the recognition accuracy of PCA and PPP reconstruction data increase. This is because reconstructing data with higher rank, the information loss decreases. On both the Extended Yale-B and FRGC datasets, across all of the data rank setting, PPP consistently outperforms PCA. For example, on the Extended Yale-B dataset, when reduced dimension is 5, the recognition accuracy based on PCA reconstructed data is only 5.28%. While for PPP, the performance is increased to 42.30%. The reason is that for the extremely low rank scenario, the information loss is quite severe. And thus the performance of PCA decreases significantly since it only considers the loss w.r.t. the original data. However, in the reconstruction process, PPP is able to preserve the critical gradient features, even if a considerable amount of (inessential) information is inevitably lost due to dimensionality reduction.. Therefore, even at the same rank (*i.e.*, with roughly the same amount of information loss), PPP can achieve much better performance than PCA. Note that when the rank of reconstructed data increases to 50, the information loss is not significant. And the performance margin of PPP

Dim	PCA+kNN	PPP-S+kNN	PPP-L+kNN	PCA+LDA	PPP-S+LDA	PPP-L+LDA
5	5.82 ± 1.64	$\mathbf{43.51 \pm 2.63}$	42.30 ± 2.17	10.40 ± 1.53	$\mathbf{44.30 \pm 1.17}$	44.41 ± 2.68
10	33.56 ± 1.24	68.34 ± 2.81	$\mathbf{70.58 \pm 2.29}$	49.89 ± 2.17	69.57 ± 2.60	$\mathbf{71.14 \pm 2.73}$
15	48.66 ± 1.95	$\mathbf{77.63 \pm 1.25}$	76.52 ± 1.09	70.13 ± 1.07	$\mathbf{78.75 \pm 2.73}$	77.40 ± 1.66
20	58.39 ± 2.49	$\mathbf{81.66 \pm 2.83}$	80.21 ± 2.10	80.54 ± 2.79	83.00 ± 1.69	$\mathbf{83.54 \pm 2.96}$
30	71.36 ± 2.07	84.23 ± 2.26	$\mathbf{84.34 \pm 3.03}$	85.12 ± 1.35	85.12 ± 2.85	$\mathbf{86.92 \pm 1.09}$
50	79.08 ± 2.27	82.66 ± 1.20	$\mathbf{86.13 \pm 0.81}$	89.15 ± 2.44	86.02 ± 1.75	$\mathbf{89.58 \pm 1.16}$

Table 1: Face recognition accuracy(%) on the Extended Yale-B database based on gradient feature. Here PPP-L denotes the PPP with low rank relaxation and PPP-S denotes the PPP optimized by gradient descent over Stiefel Manifold.

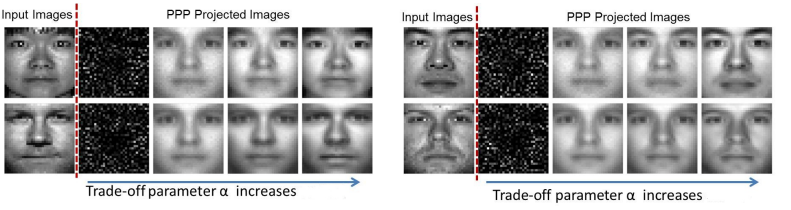


Figure 5: The reconstruction results of PPP under different values of trade-off parameter α.

and PCA also decreases. But it can be seen that PPP still outperforms PCA with around 7% performance margin.

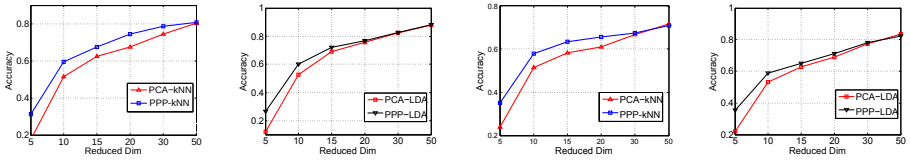
4.3 Experiment III: Gabor Feature Preserving

In this experiment, we specifically investigate the capability of PPP for Gabor feature p-reserving in the data projection. Besides Gabor feature is useful for machine recognition [15] [30], there are also findings suggesting that Gabor filter is close to human visual perception systems [6]. To evaluate the performance quantitatively, we also conduct the face recognition experiments on the Extended Yale-B and FRGC datasets, following the same experiment setting as above for Gradient feature.

Experiment settings: For the Gabor feature extraction, here we adopt a bank of Gabor filters [13] consisting of 8 different orientations (with orientation parameter $\phi \in \{0,\ldots,7\}$) and 5 different scales (with scale parameter $\nu = \{0,\ldots,4\}$). Namely there are in total 40 different Gabor filters and 40 different corresponding linear operators $\mathcal{P}_i, i = 1,\ldots,40$. Since the Gabor filter is a complex filter. We isolate the real part $P_i^{\mathcal{R}}$ and imaginary part $P_i^{\mathcal{I}}$ and concatenate them as $P_i = [P_i^{\mathcal{R}}; P_i^{\mathcal{I}}]$ to form the used linear operator. This projection methodology is reasonable because in the following feature extraction process, magnitude of Gabor feature is used and the real/imaginary part can be separably projected in our algorithm.

Results: The experimental results are shown in Figure 3, Figure 6 and Table 2. We can observe the similar performance trend as the above gradient feature based experiments. In particular, on the Extended Yale-B dataset, when the reduced dimension is equal to 5, the performance of PCA is 4.47% while PPP improves the performance to 40.16% for the low-rank based optimization. The performance improvement brought by PPP is quite significant. And when the reduced dimension is 50, the performance of PCA is 83% while PPP achieves a little higher accuracy, namely 83.11%. To investigate how the PPP can preserve specialized perception in the image projection, we provide some exemplar results in Figure 5. Here when $\alpha = 0$, the results are from purely feature preserving. And only some discriminative points are preserved in the reconstruction. When the value of α increases, the results become more similar to the original images. An interesting result is that when α is small, the reconstructed images of different persons look quite similar. However, by extracting the Gabor features, these images can be distinguished correctly.

We should also explain that the experiments in this section are *not* to achieve state-of-the-



(a) Gabor + kNN (b) Gabor + LDA (c) LoG + kNN (d) LoG + LDA

Figure 6: The performance comparison for reconstructed face recognition from (a) PCA and (b) PPP on FRGC dataset based on Gabor/LoG feature.

Dim	PCA+kNN	PPP-S+kNN	PPP-L+kNN	PCA+LDA	PPP-S+LDA	PPP-L+LDA
5	4.47 ± 2.92	40.04 ± 3.38	$\mathbf{40.16 \pm 0.69}$	17.34 ± 1.93	36.13 ± 3.17	$\mathbf{38.59 \pm 1.77}$
10	41.05 ± 2.93	64.21 ± 1.87	$\mathbf{69.46 \pm 1.64}$	62.53 ± 2.04	$\mathbf{74.83 \pm 3.28}$	72.71 ± 2.62
15	57.94 ± 1.32	75.50 ± 4.40	$\mathbf{79.08 \pm 2.79}$	82.21 ± 2.22	84.56 ± 1.15	$\mathbf{87.81 \pm 1.68}$
20	70.36 ± 2.94	78.97 ± 0.74	$\mathbf{81.77 \pm 1.25}$	87.92 ± 1.04	86.80 ± 2.21	$\mathbf{91.05 \pm 1.38}$
30	79.19 ± 2.91	79.42 ± 2.36	$\mathbf{82.89 \pm 1.86}$	90.72 ± 1.28	90.04 ± 2.06	$\mathbf{91.39 \pm 1.77}$
50	83.00 ± 1.97	82.44 ± 2.13	$\mathbf{83.11 \pm 0.95}$	91.50 ± 2.26	92.06 ± 2.76	$\mathbf{92.17 \pm 2.45}$

Table 2: Recognition accuracy(%) on the Extended Yale-B database based on Gabor features.

art performance for face recognition. Rather, the goal is to compare the perception preserving performance of PPP as the same role where PCA may be applied.

5 Related Work

There are many variants of PCA, such as the probablilistic PCA [20], the kernel PCA [18], the Laplacian PCA [32], the generalized PCA [24]. Recent years have witnessed increasing interest in low rank matrix approximation, such as the low-rank matrix completion [3], robust PCA (RPCA) [4]. In particular, these methods pursue a low-rank matrix approximation to the data matrix by isolating certain structured noise (e.g., gross error on certain dimension, outliers). RPCA is generally used for removing the corruptiong in the data and recovering the low rank data. To explicitly obtain the projection basis, a standard PCA operation is necessary following the RPCA. Our proposed method focuses on the projection learning instead of corruption removing and data recovering. Several methods have been proposed to to improve the perceptual(visual) quality of reconstructed images. Smart PCA [31] is based on the probabilistic interpretation of PCA, the inverse Wishart distribution is used as conjugate prior for the population covariance and domain knowledge can be transferred by the prior hyperparameters. However the domain knowledge in smart PCA is still constrained on intensity-domain feature distance functions(such as pixel locations). 2DPCA [29] is based on 2D image matrices rather than 1D vectors so the image matrix does not need to be transformed into a vector prior to feature extraction. Instead, an image covariance matrix is constructed directly using the original image matrices, and its eigenvectors are derived for image feature extraction. Though 2DPCA is useful in feature extraction, it is not able to reduce dimensions of the images. However, the main contribution of our work is not to gain good perceptual quality on the intensity domain. What we want to preserve specializes to a certain type of features that are provided by a specific perception system (or user), meanwhile retain the simplicity and unsupervised manner of classic PCA. It is also interesting to see how PPP can be regarded as an implementation of unsupervised joint embedding [26] of different domains(original and feature space in our case).

6 Conclusion

We propose the perception preserving projection (PPP) method, which is able to preserve the important information for specific perception system in the image projection process. In particular, we explicitly embed the feature preserving metric provided by a certain type

of perception systems into the loss function and propose a low-rank approximation optimization method to the problem. In two concrete application scenarios, we take Gabor and gradient features as illustration and experimentally evaluate that PPP can better preserve the discriminative and domain-specific feature information.

Acknowledgements

We would like to thank the anonymous reviewers. This work was partially supported by Singapore Ministry of Education under research Grant MOE2010-T2-1-087. Xie and Lu were also supported by National Natural Science Foundation of China (No.61272247).

References

[1] P.A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. 2009.

[2] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[3] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 2009.

[4] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] John G Daugman et al. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Optical Society of America, Journal, A: Optics and Image Science*, 2(7):1160–1169, 1985.

[7] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001.

[8] Hervé Jégou and Ondřej Chum. Negative evidences and co-occurences in image retrieval: the benefit of pca and whitening. In *Proceedings of the 12th European conference on Computer Vision - Volume Part II*, pages 774–787, 2012.

[9] Ian T Jolliffe. *Principal component analysis*. Springer verlag, 2002.

[10] R.M. Larsen. Lanczos bidiagonalization with partial reorthogonalization. *DAIMI PB*, 27(537), 1998.

[11] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[12] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *NIPS*, 2011.

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

[14] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184, 2013.

[15] J.R. Movellan. Tutorial on gabor filters. 2002.

[16] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005.

[17] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3): 471–501, 2010.

[18] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[19] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002.

[20] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (3):611–622, 1999.

[21] Duan Tran and David A Forsyth. Configuration estimates improve pedestrian finding. In *Advances in neural information processing systems*, pages 1529–1536, 2007.

[22] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 1991.

[23] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Subspace learning from image gradient orientations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34 (12):2454–2466, 2012.

[24] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12): 1945–1959, 2005.

[25] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, pages 1–38, 2013.

[26] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[27] Jacob Whitehill and Javier Movellan. Discriminately decreasing discriminability with learned image filters. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2488–2495. IEEE, 2012.

[28] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(1):40–51, 2007.

[29] Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(1):131–137, 2004.

[30] Meng Yang and Lei Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *ECCV 2010*, pages 448–461. Springer, 2010.

[31] Yi Zhang. Smart pca. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[32] Deli Zhao, Zhouchen Lin, and Xiaoou Tang. Laplacian pca and its applications. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.