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

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) [2] in particular is one of the most popular projection learning methods for dimensionality reduction and feature extraction. Representations learned with PCA have proven useful for tasks such as face and object recognition, tracking, detection, and background modeling.

Compared with the descriptive features learned from the raw intensity domain, many advanced features work better either for machine perception (pattern recognition accuracy), or human perception (physical interpretability), or both. In those cases, the standard PCA is not tuned to any domain-specific features, thus the discriminative features for certain perception systems may be lost in the projection process.

In this work, we propose a perception preserving projection (PPP) method that is able to preserve the desired feature characteristic for projected images. PPP pursues a set of suitable projection basis which minimize the reconstruction loss for both the original images and the extracted features simultaneously. 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 objective function of PPP can then be formulated as:

$$\min_{U^TU=I_r} \mathcal{L}(U) = \|\mathcal{P}'(X) - \mathcal{P}'(UU^TX)\|_F^2, \tag{1}$$

There are various feature extractors can be seen as linear operators $\mathcal{P}$ over the data vector $\mathbf{x} \in \mathbb{R}^d$. For example, the convolution of data with linear filter, pixel-wise "masking", and the sum of filters. Here $P' = [(1-\alpha)P, \alpha I_d]$, and α is a trade-off parameter between the original data space and feature space. 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 $\mathcal{P}(UU^TX)$ will not deviate from the features extracted from the original data $\mathcal{P}(X)$ too much.

The most straightforward method to solve the problem (1) is performing gradient descent on the Stiefel manifold defined by $U^TU = I$. Though state-of-the-art algorithms can be directly exploited, the computational cost is quite high.

Inspired by the Robust PCA work [1], 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 (1) can be relaxed as follows:

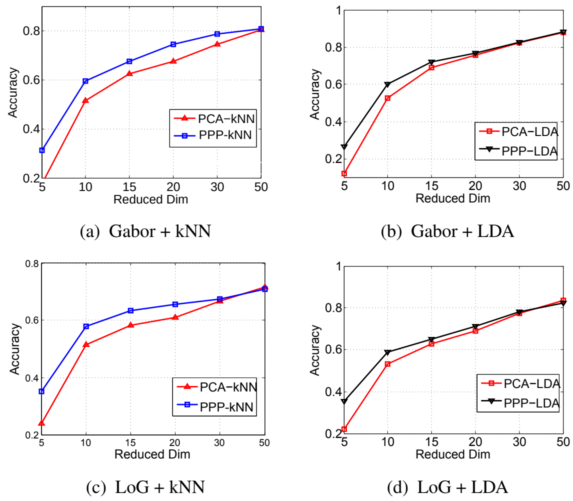
$$\min_W \|\mathcal{P}(X) - \mathcal{P}(WX)\|_F^2, \text{ s.t. } \mathrm{rank}(W) \le r, \tag{2}$$

The above objective function can be further relaxed as:

$$\min_W \|W\|_* + \lambda\|E\|_F^2, \text{ s.t. } \mathcal{P}(X) - \mathcal{P}(WX) = E, \tag{3}$$

where E explicitly accommodates the reconstruction errors.

The optimization problem in (3) can be solved by the Alternating Direction Method (ADM) [3] efficiently. 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, it is difficult to solve the sub-problem for optimizing the function w.r.t. W. Directly minimizing the subproblem above will lead to solving a discrete-time Sylvester equation in each iteration, which is infeasible for the ADM algorithm and makes the low-rank relaxation rewardless.



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

Figure 1: The performance comparison for reconstructed face recognition from (a) PCA and (b) PPP on FRGC dataset based on Gabor/LoG feature.

To alleviate such difficulty, we adopt the recently developed Linearized Alternating Direction Method (LADM) [4] and linearize the quadratic term in the above Lagrangian function at the point W^k:

$$\mathcal{L}(W,W_k) = \|W\|_* + \langle Y, -\mathcal{P}(W_kX)\rangle + \mu\left\langle \mathcal{P}^*\left(\mathcal{P}(X) - \mathcal{P}(W^kX) - E\right)X^T, W - W_k\right\rangle + \frac{\mu\eta}{2}\|W - W_k\|_F^2. \tag{4}$$

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:

$$\mathcal{L}(W,Y,W_k) = \|W\|_* + \frac{\mu\eta}{2}\|W - M_k\|_F^2, \tag{5}$$

where $M_k = W_k - \mathcal{P}^*(\mathcal{P}(X) - \mathcal{P}(W_kX) - E)X^T/\eta + \mathcal{P}^*YX^T/\mu\eta$. $\mathcal{P}^*$ 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:

$$W_{k+1} = U\mathcal{S}_{\frac{1}{\mu\eta}}(\Sigma)V^T, \tag{6}$$

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 [4]:

$$\mu_{k+1} = \begin{cases} \rho_0\mu_k, & \text{if } \mu_k \max(\sqrt{\eta}\varepsilon_W, \varepsilon_E)/\|\mathcal{P}(X)\| < \varepsilon_2, \\ \mu_k, & \text{otherwise.} \end{cases} \tag{7}$$

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

$$\|\mathcal{P}(X) - \mathcal{P}(W^kX) - E\|/\|\mathcal{P}(X)\| < \varepsilon_1. \tag{8}$$

In our implementation, we also adopt partial SVD and rank prediction techniques using PROPACK and represent W as its skinny SVD to avoid full matrix multiplications and thus yielding a complexity of $\mathcal{O}(nd^2)$.

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

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

[3] 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.

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