Face Recognition using Discriminatively Trained Orthogonal Rank One Tensor Projections

Gang Hua, Paul A. Viola, Steven M. Drucker
Microsoft Live Labs Research
One Microsoft Way, Redmond, WA 98052
{ganghua, viloa, sdrucker}@microsoft.com

Abstract

We propose a method for face recognition based on a discriminative linear projection. In this formulation images are treated as tensors, rather than the more conventional vector of pixels. Projections are pursued sequentially and take the form of a rank one tensor, i.e., a tensor which is the outer product of a set of vectors. A novel and effective technique is proposed to ensure that the rank one tensor projections are orthogonal to one another. These constraints on the tensor projections provide a strong inductive bias and result in better generalization on small training sets. Our work is related to spectrum methods, which achieve orthogonal rank one projections by pursuing consecutive projections in the complement space of previous projections. Although this may be meaningful for applications such as reconstruction, it is less meaningful for pursuing discriminant *projections. Our new scheme iteratively solves an eigenvalue problem with orthogonality constraints on one dimension, and solves unconstrained eigenvalue problems on the other dimensions. Experiments demonstrate that on small and medium sized face recognition datasets, this approach outperforms previous embedding methods. On large face datasets this approach achieves results comparable with the best, often using fewer discriminant projections.*

1. Introduction

Appearance based face recognition is often formulated as a problem of comparing labeled example images with unlabeled probe images. Viewed in terms of conventional machine learning, the dimensionality of the data is very high, the number of examples is very small, and the data is corrupted with large confounding influences such as changes in lighting and pose. As a result, conventional techniques such as nearest neighbor classification are not terribly effective. The predominant proposed solution is to find a projective embedding of the original data into a lower dimensional space that preserves discriminant information and discards confounding information. Techniques such as EigenFaces (PCA) [12], FisherFaces (LDA) [1], local discriminant embedding (LDE) [3], and variants of locality preserving projections (LPP) [8, 2], have proven to be effective to varying degrees.

All these techniques must address three challenges: *high dimensionality*, *learning capacity*, and *generalization ability*. Learning capacity, sometimes called inductive bias or discriminant ability, is the capacity of an algorithm to represent arbitrary class boundaries. It can be measured, for example, using Fisher's criterion or the Vapnik-Chervonenkis dimension [13]. Generalization ability is a measure of the expected errors on data outside of the training set. It is most famously measured by classification margin [13]. While tradeoffs of these factors apply in any practical machine learning approach, face recognition presents extreme challenges.

In general, complex models with more parameters (e.g., neural networks) have higher learning capacity but are prone to over-fit and thus have low generalization ability. When available, a large quantity of diversified training data can be used to better constrain the parameters. Simpler models with fewer parameters, tend to yield better generalization, but have limited learning capacity. How to tradeoff these issues, especially with high dimensional visual data, remains an open issue. In this paper, we address these challenges by pursuing a series of *orthogonal* rank one tensor projections designed to maximize discriminative information.

Many discriminant learning methods treat image data as vectors (such as the variants of LDA [1], LPP [8, 2], and LDE [3, 4]). These approaches have difficulty with high dimensionality, a matter made worse when there is only small set of training data. All the methods mentioned above involve solving an eigenvalue problem in the high dimensional input vector space (i.e., 1024 dimensions for 32×32 images). Solving the Eigen decomposition in high dimensions is not only computationally intensive, but also prone

1-4244-1180-7/07/$25.00 ©2007 IEEE

to numerical issues. For example, when the within class scattering matrix of LDA is singular, principal component analysis (PCA) [1] is usually performed beforehand. In this case it is clearly possible that the most discriminative projections may have been discarded. Vector based representations also ignore the spatial structure of image data which may be very useful for visual recognition.

An alternative is to regard image data as a tensor [7, 3, 14, 15] (i.e., multiple dimensional arrays). With the tensor representation, discriminant multi-linear projections (e.g., bi-linear projections for 2 dimensional tensor) are pursued to construct the discriminant embedding. In many cases, discriminant multi-linear projections can be obtained by solving the eigenvalue problems iteratively on the n different dimensions of the tensor space.

Tensor representations of images do not suffer from the same curse-of-dimensionality as vector space representations. Tensor projections are represented as the outer product of n lower dimensional vectors. Rather than expending 1024 parameters for each projection, two dimensional tensors can operate with as few as 64 parameters per projection. As discussed below, the GLOCAL tensor representation has the added benefit of respecting the geometric structure in images [4].

Most previous tensor based learning methods for discriminant embedding [3, 7, 15] constrain the spanning set of multi-linear projections to be formed by the combination of outer products of a small number of column vectors. This may have over-constrained the learning capacity of the projection vectors.

To address the conflicting goals of capacity and generalization, we propose to learn a projection which is a combination of orthogonal rank one tensors. Note that two rank one tensors are orthogonal if and only if they are orthogonal on at least one dimension of the tensor space. Using this insight we propose a novel scheme to achieve orthogonality. Our new scheme iteratively solves an eigenvalue problem with orthogonality constraints on one dimension, and solves unconstrained eigenvalue problems on the other dimensions of the tensor space.

Our approach is different from the rank one projections with adaptive margins (RPAM) proposed in [14]. Firstly, the rank one projections pursued in our approach are orthogonal, while those learned from RPAM are not. Previous research [2, 5] has shown that orthogonality increases the discriminative power of the projections. Note, we do not use adaptive margin in our formulation although that could be easily incorporated into our framework.

The remainder of the paper is organized as follows. Sec. 2 presents some notation and definition of tensors. Sec. 3 presents our new algorithm for pursuing discriminant ortho-normal rank one tensor projections, with a discussion of a limitation and an effective means to conquer it. Sec. 4 presents extensive experimental results and discussions. Finally we conclude in Sec. 5.

2. Rank one projection and orthogonality

In linear algebra, an order n real-valued tensor is a multiple dimensional array $\mathbf{X} \in \mathbb{R}^{m_0 \times m_1 \ldots \times m_{n-1}}$, and $x_{i_0 i_1 \ldots i_{n-1}}$ is the element at position $(i_0, i_1, \ldots, i_{n-1})$. We then define the rank one projection.

Definition 2.1 *Given an order n tensor $\mathbf{X}$, a rank one projection is a $\mathbf{X} \in \mathbb{R}^{m_0 \times m_1 \ldots \times m_n} \rightarrow y \in \mathbb{R}$ mapping defined by $\tilde{P} = \{\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_{n-1}\}$ where each $\mathbf{p}_i$ is a column vector of dimension m_i with the k^{th} element p_{ik}, such that*

$$y = \sum_{i_{n-1}} \ldots (\sum_{i_1} (\sum_{i_0} x_{i_0 i_1 \ldots i_{n-1}} p_{0 i_0}) p_{1 i_1} \ldots) p_{n-1 i_{n-1}} \quad (1)$$

The notation can be simplified using the k-mode product [9, 14], i.e.,

Definition 2.2 *The k-mode product of tensor $\mathbf{X} \in \mathbb{R}^{m_0 \times \ldots m_k \ldots \times m_{n-1}}$ and a matrix (i.e., an order 2 tensor) $\mathbf{B} \in \mathbb{R}^{m_k \times m'_k}$ is a $\mathbf{X} \in \mathbb{R}^{m_0 \times \ldots m_k \times \ldots \times m_{n-1}} \rightarrow \mathbf{Y} \in \mathbb{R}^{m_0 \times \ldots m'_k \times \ldots \times m_{n-1}}$ mapping, i.e., $\mathbf{Y} = \mathbf{X} \times_k \mathbf{B}$, where*

$$y_{i_0 \ldots i_{k-1} i'_k i_{k+1} \ldots i_{n-1}} = \sum_{j=0}^{m_k - 1} x_{i_0 \ldots i_{k-1} j i_{k+1} \ldots, i_{n-1}} b_{j i'_k} \quad (2)$$

Eq. 1 can then be written as $y = \mathbf{X} \times_0 \mathbf{p}_0 \ldots \times_{n-1} \mathbf{p}_{n-1}$, or in short $y = \mathbf{X} \odot \tilde{P}$. Let $\tilde{\mathbf{P}}^d = \{\tilde{P}^{(0)}, \ldots, \tilde{P}^{(d-1)}\}$ be a set of d rank one projections, we denote the mapping from $\mathbf{X}$ to $\mathbf{y} = [y_0, y_1, \ldots, y_{d-1}]^T \in \mathbb{R}^d$ as,

$$\mathbf{y} = [\mathbf{X} \odot \tilde{P}^{(0)}, \ldots, \mathbf{X} \odot \tilde{P}^{(d-1)}]^T \triangleq \mathbf{X} \odot \tilde{\mathbf{P}} \quad (3)$$

A rank one projection is also the sum of element-wise product of $\mathbf{X}$ and the reconstruction tensor of $\tilde{P}$.

Definition 2.3 *The reconstruction tensor of $\tilde{P}$ is $\acute{\mathbf{P}} \in \mathbb{R}^{m_0 \times m_2 \ldots \times m_{n-1}}$ such that*

$$\acute{\mathbf{P}} = [\acute{p}_{i_0 i_1 \ldots i_{n-1}}] = [\prod_{k=0}^{n-1} p_{k i_0}] \quad (4)$$

Then $y = \mathbf{X} \odot \tilde{P} = \sum_{i_0 i_1 \ldots i_{n-1}} x_{i_0 i_1 \ldots i_{n-1}} \acute{p}_{i_0 i_1 \ldots i_{n-1}}$. An order n rank one projection is indeed a constrained vector space linear projection $\mathbf{x} \in \mathbb{R}^{\prod_i m_i} \rightarrow y \in \mathbb{R}$ such that $y = \hat{\mathbf{p}}^T \mathbf{x}$, where $\mathbf{x}$ is the vector scanned dimension by dimension from $\mathbf{X}$, and $\hat{\mathbf{p}}$ is defined as

$$\hat{\mathbf{p}} = \mathbf{p}_{n-1} \otimes \mathbf{p}_{n-2} \otimes \mathbf{p}_0 \quad (5)$$

where $\otimes$ is the Kronecker product of matrices. We then define the orthogonality of two rank one projections, i.e.,

Definition 2.4 *Two rank one projection $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ are orthogonal, if and only if the corresponding vectors $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ calculated from Eq. 5 are orthogonal.*

Note that our definition of the orthogonality of rank one projections is equivalent to what is defined in [9]. Similarly, we call $\tilde{P}$ a *normal* rank one projection if and only if $\hat{\mathbf{p}}$ is a normal vector. It is obvious that if all $\mathbf{p}_i$ of $\tilde{P}$ are normal vectors, then $\tilde{P}$ is a normal rank one projection.

3. Ortho-rank-one Discriminant Analysis

3.1. Problem formulation

Given a training set $\{\mathbf{X}_i \in \mathbb{R}^{m_0 \times m_1 \ldots \times m_{n-1}}\}_{i=0}^{N-1}$, and set of pairwise labels $\mathcal{L} = \{l(i,j) : i < j; i, j \in \{0, \ldots, N-1\}\}$, where $l(i,j) = 1$ if $\mathbf{X}_i$ and $\mathbf{X}_j$ are in the same category, and $l(i,j) = 0$ otherwise. Let $\mathcal{N}_k(i)$ be the set of k-nearest neighbors of $\mathbf{X}_i$,

$$\mathcal{D} = \{(i,j)|i < j, l(i,j) = 0, \mathbf{X}_i \in \mathcal{N}_k(j)||\mathbf{X}_j \in \mathcal{N}_k(i)\}$$
$$\mathcal{S} = \{(i,j)|i < j, l(i,j) = 1, \mathbf{X}_i \in \mathcal{N}_k(j)||\mathbf{X}_j \in \mathcal{N}_k(i)\}$$

be the indices set of all example pairs which are k-nearest neighbors of one another, and are from different and same categories, respectively. Our objective is to learn a set of K ortho-normal rank one projections $\tilde{\mathbf{P}}^K = (\tilde{P}^{(0)}, \tilde{P}^{(1)}, \ldots, \tilde{P}^{(K-1)})$, such that in the projective embedding space, the distances of the example pairs in $\mathcal{S}$ are minimized, while the distances of those in $\mathcal{D}$ are maximized.

To achieve this, we propose to maximize a series of local weighted discriminant cost functions [3]. Suppose we have obtained k discriminant rank one projections indexed from 0 to $k-1$, to pursue the $(k+1)^{th}$ rank one projection, we want to solve the following constrained optimization problem,

$$\max_{\tilde{P}^{(k)}} \quad \frac{\sum_{\mathcal{D}} \omega_{ij} \|\mathbf{X}_i \odot \tilde{P}^{(k)} - \mathbf{X}_j \odot \tilde{P}^{(k)}\|^2}{\sum_{\mathcal{S}} \omega_{ij} \|\mathbf{X}_i \odot \tilde{P}^{(k)} - \mathbf{X}_j \odot \tilde{P}^{(k)}\|^2} \quad (6)$$

$$s.t. \quad \tilde{P}^{(k)} \perp \tilde{P}^{(k-1)}, \ldots, \tilde{P}^{(k)} \perp \tilde{P}^{(0)} \quad (7)$$

where $\|\cdot\|$ is the Euclidean distance, and ω_{ij} is a weight assigned according to the importance of the example pair $(\mathbf{X}_i, \mathbf{X}_j)$. We use the heat kernel weight [3], i.e., $\omega_{ij} = \exp\left\{-\frac{\|\mathbf{X}_i - \mathbf{X}_j\|_F^2}{t}\right\}$ where $\|\cdot\|_F$ denotes the Frobenius norm, and t is a constant parameter. It introduces heavy penalties to the cost function for example pairs which are close to one another. Notice that for $k = 0$, we only need to solve an unconstrained optimization problem of Eq. 6.

There are two difficulties in the constrained maximization of Eq. 6: firstly, it is in general difficult to keep both the rank one and orthogonality properties; secondly, there is no closed-form solution to the unconstrained optimization problem of Eq. 6. It is well known that the second problem can be addressed numerically by using a sequential iterative optimization scheme [3]. We present our solution to the first problem in the next section.

3.2. Learning algorithm

Our solution starts from the following proposition:

Proposition 3.1 *Two rank one projections $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$ are* orthogonal *to each other, if and only if for at least one i, $\mathbf{p}_i^{(1)} \in \tilde{P}^{(1)}$ is orthogonal to $\mathbf{p}_i^{(2)} \in \tilde{P}^{(2)}$, i.e., $\mathbf{p}_i^{(1)} \perp \mathbf{p}_i^{(2)}$.*

The proof is presented in Appendix A. From this Proposition, an equivalent set of constraints of Eq. 7 is,

$$\begin{aligned} \exists \quad & \{j_l : l \in \{0, \ldots, k-1\}; j_l \in \{0, \ldots, n-1\}\} \\ : \quad & \mathbf{p}_{j_{k-1}}^{(k)} \perp \mathbf{p}_{j_{k-1}}^{(k-1)}, \ldots, \mathbf{p}_{j_0}^{(k)} \perp \mathbf{p}_{j_0}^{(0)}. \end{aligned} \quad (8)$$

To make the optimization more tractable, we replace the constraints on Eq. 7 with the following stronger constraints.

$$\exists j \in \{0, \ldots, n-1\} \ : \ \mathbf{p}_j^{(k)} \perp \mathbf{p}_j^{(k-1)}, \ldots, \mathbf{p}_j^{(k)} \perp \mathbf{p}_j^{(0)}. \quad (9)$$

These constraints are stronger because it requires all j_l in Eq. 8 to be the same. It is obvious that the constraints in Eq. 9 are sufficient conditions for the constraints in Eq. 7.

It is well known that the unconstrained problem in Eq. 6 can be solved numerically in a sequential iterative fashion. That is, at each iteration, we fix $\tilde{P}_i^{(k)} = \{\mathbf{p}_0^{(k)}, \ldots, \mathbf{p}_{i-1}^{(k)}, \mathbf{p}_{i+1}^{(k)}, \ldots, \mathbf{p}_{n-1}^{(k)}\}$ for one $i \in \{0, \ldots, n-1\}$, and maximize Eq. 6 w.r.t. $\mathbf{p}_i^{(k)}$. For notation simplification, we denote

$$\begin{aligned} \mathbf{y}^{(k)} &= \mathbf{X} \times_0 \mathbf{p}_0^{(k)} \ldots \times_{i-1} \mathbf{p}_{i-1}^{(k)} \times_{i+1} \mathbf{p}_{i+1}^{(k)} \cdots \mathbf{p}_{n-1}^{(k)} \\ &\triangleq \mathbf{X} \odot \tilde{P}_i^{(k)}, \end{aligned} \quad (10)$$

which is a m_i dimensional vector. Then we need to solve

$$\max_{\mathbf{p}} \frac{\mathbf{p}^T \mathbf{A}_d^{(i)} \mathbf{p}}{\mathbf{p}^T \mathbf{A}_s^{(i)} \mathbf{p}} \quad (11)$$

where

$$\mathbf{A}_d^{(i)} = \sum_{\mathcal{D}} \omega_{op} (\mathbf{y}_o^{(k)} - \mathbf{y}_p^{(k)})(\mathbf{y}_o^{(k)} - \mathbf{y}_p^{(k)})^T \quad (12)$$

$$\mathbf{A}_s^{(i)} = \sum_{\mathcal{S}} \omega_{op} (\mathbf{y}_o^{(k)} - \mathbf{y}_p^{(k)})(\mathbf{y}_o^{(k)} - \mathbf{y}_p^{(k)})^T \quad (13)$$

$$\mathbf{y}_o^{(k)} = \mathbf{X}_o \odot \tilde{P}_i^{(k)}, o = 1, \ldots, N \quad (14)$$

It is also well known that the optimal solution of Eq. 11 can be obtained by solving the generalized eigenvalue problem

$$\mathbf{A}_d^{(i)} \mathbf{p} = \lambda \mathbf{A}_s^{(i)} \mathbf{p}, \quad (15)$$

and the optimal solution $\mathbf{p}_i^{(k)*}$ is the eigenvector associated with the largest eigenvalue. Eq. 15 is solved iteratively over $i = 1, 2, \ldots, n$ one by one until convergence. The final output $\tilde{P}^{(k)*} = \{\mathbf{p}_0^{(k)*}, \mathbf{p}_1^{(k)*}, \ldots, \mathbf{p}_n^{(k)*}\}$ is regarded as the optimal solution to the unconstrained Eq. 6. This iterative algorithm can only guarantee a local optimal solution.

To solve Eq. 6 with the constraints in Eq. 9, suppose j have been chosen, the iteration steps to optimize those $\mathbf{p}_i^{(k)}$ where $i \neq j$ should remain unchanged since the constraints do not apply to them. Now we need to address the problem of solving Eq. 11 for $i = j$, such that the constraints in Eq. 9 holds. It is equivalent to solving the following problem, i.e.,

$$\begin{aligned} \max_{\mathbf{p}_j^{(k)}} \quad & (\mathbf{p}_j^{(k)})^T \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} \\ s.t. \quad & (\mathbf{p}_j^{(k)})^T \mathbf{A}_s^{(j)} \mathbf{p}_j^{(k)} = 1 \\ & (\mathbf{p}_j^{(k)})^T \mathbf{p}_j^{(k-1)} = 0 \\ & \cdots \\ & (\mathbf{p}_j^{(k)})^T \mathbf{p}_j^{(0)} = 0. \end{aligned} \tag{16}$$

It can be shown (see Appendix B) that the solution can be obtained by solving the following eigenvalue problem, i.e.,

$$\tilde{\mathcal{M}} \mathbf{p}_j^{(k)} = \left(\mathcal{M} (\mathbf{A}_s^{(j)})^{-1} \mathbf{A}_d^{(j)} \right) \mathbf{p}_j^{(k)} = \lambda \mathbf{p}_j^{(k)} \tag{17}$$

where

$$\mathcal{M} = \mathbf{I} - (\mathbf{A}_s^{(j)})^{-1} \mathcal{A} \mathcal{B}^{-1} (\mathcal{A})^T \tag{18}$$

$$\mathcal{A} = \left[\mathbf{p}_j^{(0)}, \mathbf{p}_j^{(1)}, \ldots, \mathbf{p}_j^{(k-1)} \right] \tag{19}$$

$$\mathcal{B} = [b_{uv}] = \mathcal{A}^T (\mathbf{A}_s^{(j)})^{-1} \mathcal{A} \tag{20}$$

The optimal $\mathbf{p}_j^{(k)*}$ is the eigenvector corresponding the largest eigenvalue of $\tilde{\mathcal{M}}$. Note the derivation of the solution in Appendix B is motivated by [5, 2]. We summarize the new sequential iterative scheme, namely orthogonal rank one tensor discriminant analysis, in Fig.1. It can only guarantee a local optimal solution, too.

3.3. Remarks

There are several points to be clarified. First, there is no theoretic guidance on how to choose j in Eq. 9. Our intuition is not to put too many constraints on one specific dimension. Therefore we randomly choose one dimension j when pursuing each one of the K rank one projection.

Second, we always perform the constrained optimization on $\mathbf{p}_j^{(k)}$ first. This ensures that the constraints in Eq. 7 hold in all iterations.

Third, if we have obtained k rank one projections and $k \geq m_i$, then we can no longer pose orthogonality constraints on the i^{th} dimension. The reason is that $\{\mathbf{p}_i^{(l)} | l = 0, \ldots, k-1\}$ already span $\mathbb{R}^{m_i}$; we can only pursue $m = \max\{m_i\}_{i=0}^{n-1}$ orthogonal rank one projections.

We address this issue by transforming the tensor space from $\mathbb{R}^{m_0 \times m_1 \ldots \times m_{n-1}}$ to $\mathbb{R}^{m'_0 \times m'_1 \ldots \times m'_{n-1}}$, where $m' = \max\{m'_i\}_{i=0}^{n-1} > m = \max\{m_i\}_{i=1}^{n-1}$. In this new transformed space, our approach can now find a maximum of m' rank one projections. In this paper we explore second order tensors, in particular we use the GLOCAL transform motivated by [4].

The GLOCAL transform partitions a tensor of size $m_0 \times m_1$ into $m'_1 = \frac{m_0 \times m_1}{l_0 \times l_1}$ non-overlapping blocks of size $l_0 \times l_1$. The blocks are ordered by a raster scan. Each block i is then itself raster scanned to be a vector of dimension $m'_0 = l_0 \times l_1$, and put into the i^{th} column of the target tensor of size $m'_0 \times m'_1$ (see Fig. 2 for an example). The GLOCAL transform can be interpreted in the following way: the column space expresses local features in pixel level, and the row space expresses global features in appearance level(see [4] for details).

INPUT: $\{\mathbf{X}_i\}_{i=1}^{N-1}$, $\mathcal{S}$ and $\mathcal{D}$

OUTPUT: $\tilde{\mathbf{P}}^K = \{\tilde{P}^{(0)}, \tilde{P}^{(1)}, \ldots, \tilde{P}^{(K-1)}\}$

1. $k = 0$, iteratively solving Eq. 15 over $i = 0, 1, \ldots, n-1$ to obtain $\tilde{P}^{(0)}$. $k = k + 1$
2. Randomly initialize each $\mathbf{p}_i^{(k)}$ as a normal vector, randomly generate number $j \in \{l | l = 0, \ldots, n-1 \ \& \ m_l > k\}$
   (a) For each $i = [j, 0, 1, \ldots, j-1, j+1, \ldots, n-1]$, fixing all other $\mathbf{p}_m^{(k)}, m \neq i$. If $i = j$, update $\mathbf{p}_i^{(k)}$ by solving Eq. 17. Otherwise update $\mathbf{p}_i^{(k)}$ by solving Eq. 15. Then normalize $\mathbf{p}_i^{(k)}$.
   (b) Repeat Step 2a until the optimization of Eq. 6 is converged to obtain $\tilde{P}^{(k)}$
3. $k = k + 1$. If $k < K$, repeat Step 2, else output $\tilde{\mathbf{P}}^K = \{\tilde{P}^{(0)}, \tilde{P}^{(1)}, \ldots, \tilde{P}^{(K-1)}\}$

Figure 1. Orthogonal Rank One Tensor Discriminant Analysis.

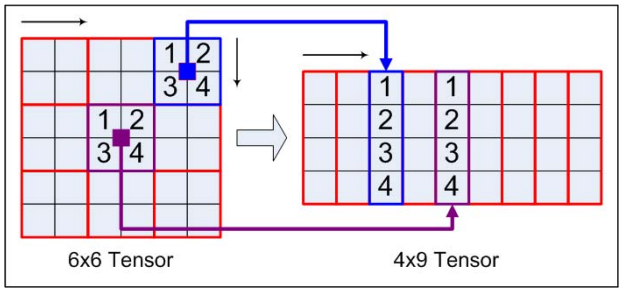


Figure 2. GLOCAL transform with 2×2 local blocks.

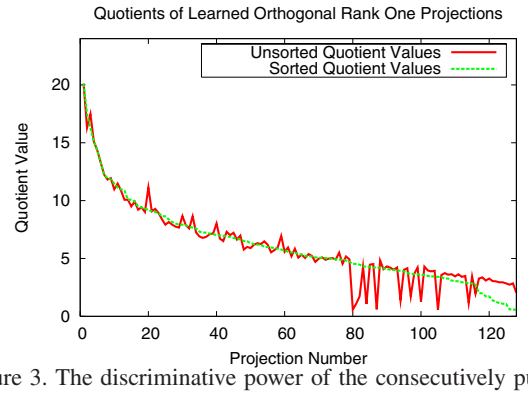


Figure 3. The discriminative power of the consecutively pursued orthogonal rank one tensor projections.

A final remark, the discriminant power (evaluated by the quotient Eq. 6) is not strictly decreasing over the sequentially pursued projections. In order to explore the effectiveness of projections with varying dimensions, we first sort them according to discriminant power. Fig. 3 displays the discriminant power of the orthogonal rank one projections obtained on a training set from CMU PIE dataset [11]. The red curve shows the unsorted quotients, and the green curve displays the sorted quotients. We perform GLOCAL transform with 4×2 blocks to form a tensor of size 8×128, allowing a total of 128 orthogonal projections.

4. Experimental results

The proposed approach is extensively tested for face recognition on some widely used benchmark data sets such as the CMU PIE database [11], the Yale face database [1], the Extended Yale Face Database B [6] and the Olivetti Research Laboratory (ORL) database [10]. We refer them to be PIE, Yale, YaleB and ORL respectively. On all datasets, the gray-scale face images are cropped and aligned by fixing the eye locations, and then resized to 32×32[1]. No other pre-processing is performed. For each data set, we randomly split it into training and testing sets. Recognition is performed using a nearest neighbor (NN) classifier based on the Euclidean distance in the learned embedding space.

We have tested our approach under three different settings: training and testing on the original images, on GLOCAL images with 4×4 blocks, and on GLOCAL images with 4×2 blocks. We call them ORO, $\text{ORO}_{4\times4}$ and $\text{ORO}_{4\times2}$, respectively. We compare the results from our approach with the state-of-the-art linear embedding methods such as PCA, [12], LDA [1], LPP [8], Tensor LPP [7], Orthogonal LPP (OLPP) [2], two dimensional local discriminant embedding with GLOCAL transform of 4×2 blocks ($\text{2DLDE}_{4\times2}$) [4], and the rank one projections with adaptive margin (RPAM) [14][2]. As a baseline, we also present

[1]The cropped Yale faces are from the authors of [2]. The other cropped data sets are from *http://ews.uiuc.edu/∼dengcai2/Data/data.html*

[2]The results of the variants of LPP are from published or public results of the authors. Our methods and our implementations of other methods are tested on the same datasets. We set $k = 5$ as the parameters of k-nearest neighbors for our methods, as well as other methods, if required.

	Error rate(%)$_{Dimension}$			
Methods	Yale	ORL	YaleB	PIE
Baseline	45.6	11.9	34.6	37.9
PCA	45.2_{71}	11.9_{189}	34.6_{780}	37.9_{1023}
LDA	22.5_{14}	6.1_{39}	18.7_{37}	10.9_{67}
LPP	21.7_{14}	6.3_{39}	13.6_{76}	$\mathbf{10.8}_{86}$
Tensor LPP	23.6_{35}	$\mathbf{4.2}_{71}$	$\mathbf{7.6}_{311}$	$\mathbf{9.7}_{68}$
OLPP	$\mathbf{17.9}_{14}$	$\mathbf{3.4}_{41}$	$\mathbf{5.7}_{241}$	$\mathbf{6.4}_{381}$
RPAM	$\mathbf{20.9}_{242}$	8.0_{219}	$\mathbf{7.6}_{389}$	$\mathbf{10.2}_{399}$
$\text{2DLDE}_{4\times2}$	$\mathbf{19.3}_{113}$	$\mathbf{4.5}_{87}$	$\mathbf{9.8}_{88}$	12.0_{104}
ORO	29.8_{32}	7.2_{30}	11.9_{32}	11.9_{31}
$\text{ORO}_{4\times4}$	$\mathbf{19.2}_{53}$	$\mathbf{4.8}_{58}$	10.9_{53}	$\mathbf{8.5}_{49}$
$\text{ORO}_{4\times2}$	$\mathbf{13.2}_{94}$ ($\mathbf{17.6}_{14}$)	$\mathbf{3.0}_{105}$ ($\mathbf{5.0}_{41}$)	$\mathbf{9.0}_{108}$ –	$\mathbf{6.4}_{73}$ –

Table 1. Face recognition on Yale, ORL, YaleB and PIE data.

the recognition results using the Euclidean distance in the original image space. The top 5 recognition results on each dataset are shown in boldface numbers. The results are discussed according to the size of the dataset, followed by a summarization of some general observations.

4.1. Face recognition on Yale database

To demonstrate the advantages of our approach for small training sets, we present our experimental results on the Yale data. It contains 165 faces of 15 individuals, each individual has 11 faces with different facial expressions and/or configurations (see the first column of Table 1). Note the subscripts of the error rates indicate the dimension of the embedding space where the best error rates are achieved (except the last row, the subscripts there indicate the dimension where the error rates are achieved). All experimental results in this column are the average of 20 random splitting of the dataset, with 5 faces from each person for training and the rest for testing. In each split there are 55 faces for training and 110 for testing.

$\text{ORO}_{4\times2}$ achieves the lowest error rate of 13.2% with 94 dimensions. Its performance is significantly better than all other methods. The second best result is from OLPP. It achieves an error rate of 17.9% with 14 dimensions. We plot the error rate versus dimension for the different methods in Fig. 4. It is clear that $\text{ORO}_{4\times2}$ (red curve) outperforms the other methods on all dimensions. After dimension 14, the error rate of $\text{ORO}_{4\times2}$ continues to decrease, while the error rate of both LPP and OLPP rise rapidly.

Note that ORO is not as good as Tensor LPP and RPAM. Our understanding is that for small training samples, the orthogonal constraints on the rank one projections of the 32×32 tensor are too strong. For this problem each rank one projection has only 64 parameters, which is already a

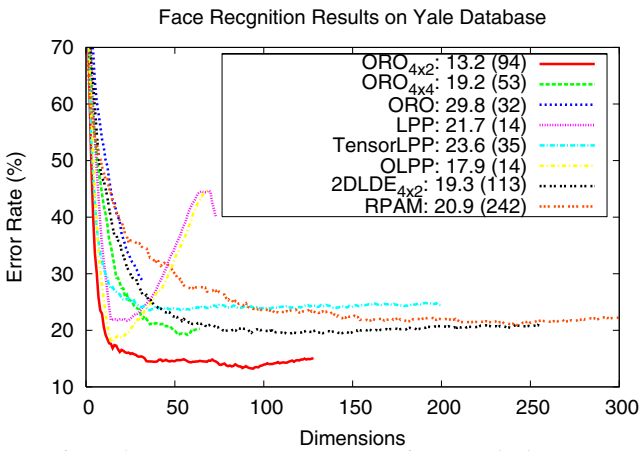


Figure 4. *Error Rate v.s. Dimensionality* on Yale data set.

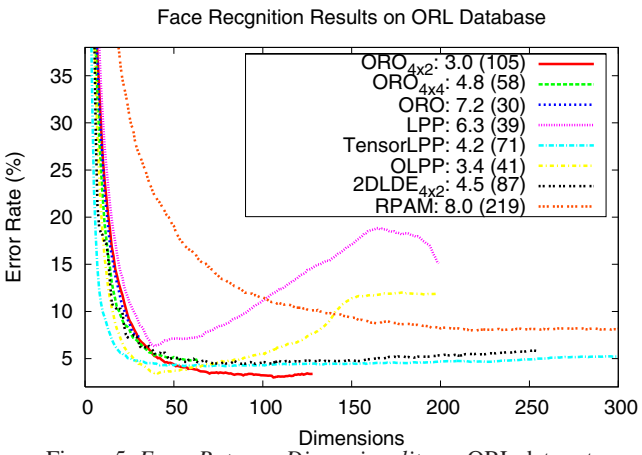


Figure 5. *Error Rate v.s. Dimensionality* on ORL data set.

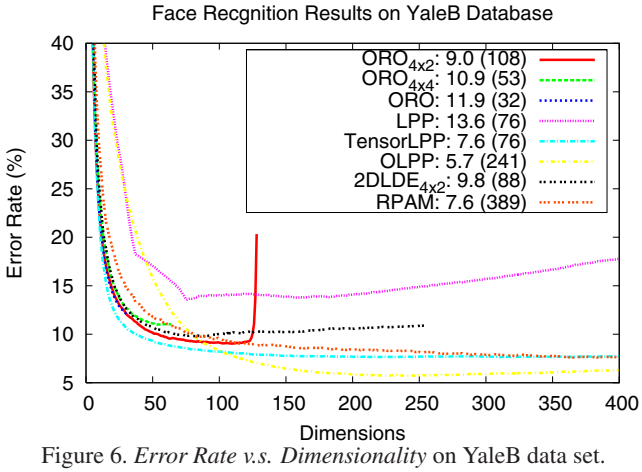


Figure 6. *Error Rate v.s. Dimensionality* on YaleB data set.

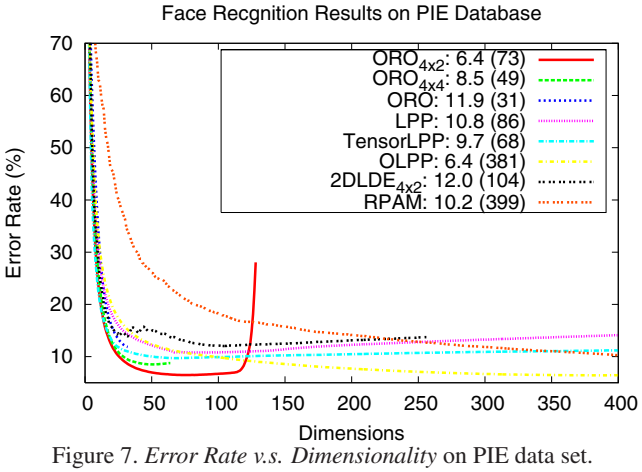


Figure 7. *Error Rate v.s. Dimensionality* on PIE data set.

fairly strong constraint. Tensor LPP and RPAM do not pose orthogonal constraints and leverage the additional capacity to achieve lower error rates. In this case the adaptive margin of RPAM may have played an important role.

$\text{ORO}_{4\times2}$, with 136 parameters for each rank one projection after the GLOCAL transform, has higher capacity. The effectiveness of orthogonality can be understood by comparing the result of $\text{ORO}_{4\times2}$ with that of $\text{2DLDE}_{4\times2}$. The discriminant cost functions Eq. 6 are similar to the 2DLDE formulation in [3, 4].

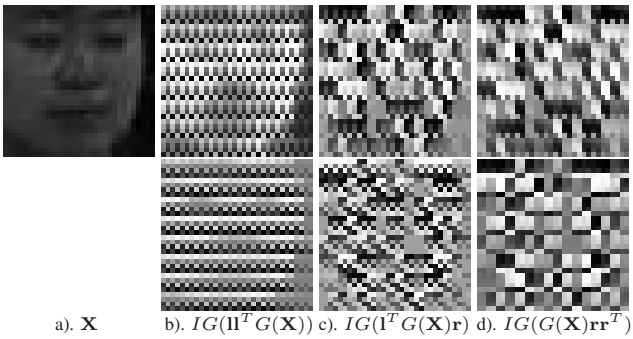
4.2. Face recognition on ORL database

The ORL dataset contains 400 face images of 40 persons, with 10 per person, which were taken at different time, under different lighting conditions, and with different facial expressions. We randomly select 5 images per person for training and the rest for testing (200 images for training and 200 for testing). The average recognition error rates of the different methods over 50 random splits are reported in the second column of Table 1.

$\text{ORO}_{4\times2}$ obtains the lowest error rate of 3.0% with 105 dimensions, followed by OLPP (3.4% with 41 dimensions), Tensor LPP (4.2% with 71 dimensions), and $\text{2DLDE}_{4\times2}$ (4.5% with 87 dimensions). $\text{ORO}_{4\times2}$ with 41 dimensions obtains an error rate of 5.0%, which is inferior to OLPP. But it is still better than PCA, LDA, LPP and RPAM. Another observation is that with the increased size of training set, the error rate of RPAM with 218 dimensions can not beat that of ORO with only 30 dimensions. Assuming that adaptive margin has positive effect, this shows that with the increased number of training examples, posing the orthogonality constraints increases generalization significantly. Again, we plot the error rate versus dimensionality in Fig. 5.

4.3. Face recognition on YaleB database

The YaleB dataset contains 21888 face images of 38 persons under 9 poses and 64 illumination conditions. We used the subset of all 2432 nearly frontal face images. We randomly choose 20 images per subject for training and the rest for testing, i.e., 760 training images and 1672 testing images. This training set is of medium size compared with the total dimension 1024. Results averaged over 50 random splits are summarized in the third column of Table 1. The error rate of $\text{ORO}_{4\times2}$ is 9.0% with 108 dimensions, better than LDA, LPP and $\text{2DLDE}_{4\times2}$, inferior yet comparable to



a). $\mathbf{X}$ b). $IG(\mathbf{l}\mathbf{l}^T G(\mathbf{X}))$ c). $IG(\mathbf{l}^T G(\mathbf{X})\mathbf{r})$ d). $IG(G(\mathbf{X})\mathbf{r}\mathbf{r}^T)$

Figure 8. For 2D tensor, a rank one projection is defined with left and right projection $\mathbf{l}$ and $\mathbf{r}$. a). source image; b). reconstructed left projection; c). full bilinear projection; d). reconstructed right projection. Each row visualize one rank one projection. $G(\cdot)$ and $IG(\cdot)$ denote GLOCAL and its inverse transform respectively.

RPAM, Tensor LPP, and OLPP. RPAM may have benefited from the adaptive margin step. With more training data, the negative effect of high dimensionality is less and thus OLPP may achieve better results. We plot error rates versus dimensionality of the different methods in Fig. 6.

4.4. Face recognition on PIE database

The PIE dataset contains 41368 images of 68 people (13 poses, 43 illumination conditions, and 4 expressions). We used the images of the 5 nearly frontal poses (C05, C07, C09, C27, C29) under all illumination conditions and expressions. This comes out to be a subset of 11560 face images with 170 images per person. We randomly select 30 images per person for training, and the rest for testing. The training set contains 2040 images, which is quite large. The average error rates over 50 random splits are summarized in the fourth column of Table 1.

Both $\text{ORO}_{4\times2}$ and OLPP achieve the lowest error rate of 6.4%. But $\text{ORO}_{4\times2}$ achieves that with only 73 dimensions while OLPP achieves that with 381 dimensions. The red curve in Fig. 7 shows how $\text{ORO}_{4\times2}$ can greedily pursue the smallest but most discriminant set of projections to achieve the lowest error rate. To better understand $\text{ORO}_{4\times2}$, we visualize the first two rank one projections applied to the first face image of this dataset in Fig. 8.

4.5. Discussions

Some general remarks are highlighted as follows:

• As shown in Fig. 6 and Fig. 7, on the YaleB and PIE datasets, adding in the last several rank one projections obtained by $\text{ORO}_{4\times2}$ dramatically degrades the recognition performance. In this case the orthogonality constraint forces these last projections to preserve non-discriminative information.

• The performance of ORO is limited by the number of orthogonal rank one projections our method can pursue. However, on YaleB, it achieves the error rate of 11.9% with 32 projections, which is much better than LDA (18.7% with 37 dimensions) and LPP (13.6% with 76 dimensions).

• Posing orthogonal constraints on the discriminant rank one projections in general helps to improve the learning capacity. This conclusion comes from comparing the results of $\text{2DLDE}_{4\times2}$ and $\text{ORO}_{4\times2}$. $\text{ORO}_{4\times2}$ performs consistently better than $\text{2DLDE}_{4\times2}$ over all datasets.

• Overall, the two orthogonal constrained algorithms, $\text{ORO}_{4\times2}$ and OLPP are the best. $\text{ORO}_{4\times2}$ outperforms OLPP on Yale and ORL, and achieves equivalent results to that of OLPP on PIE. It is only inferior to OLPP on YaleB.

• The locality preserving criterion (LPC) [7, 2], and the local discriminant criterion(LDC) [3, 4] used in this paper are two criteria for selecting discriminant projections. It has been shown that LDC is superior to unsupervised LPC [3]. Our current experiments compared LDC with supervised LPC, but it is still not clear which one is better. More investigation is necessary and we defer to our future work.

• RPAM [14] tends to require more projections to achieve a good performance due to the adaptive margin step. Adaptive margin is effective in our experiments. Incorporating it to our approach is straightforward and is part of planned future work.

5. Conclusion and future work

A novel embedding method for visual recognition is proposed, which sequentially pursues a set of discriminant orthogonal rank one projections. It was applied to the task of face recognition. Extensive experiments demonstrate that it outperforms most state-of-the-art linear embedding methods such as LDA, LPP, Tensor LPP, 2DLDE and OLPP.

Future work includes testing the proposed approach on higher order (≥ 3) tensor data, exploiting adaptive margins [14], and exploring nonlinear projections using kernel tricks [15]. We also plan to investigate both theoretically and empirically to better understand the pros and cons of the two discriminative criteria, supervised LPC and LDC, for the task of visual recognition.

Acknowledgment

We thank Dr. Shuicheng Yan (UIUC) for discussions. We also thank Deng Cai (UIUC) for providing his MATLAB code and experimental results of the variants of LPP.

A. Proof of Proposition 3.1

Proof Denote $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b}$ to be the dot product of two vectors. For two rank one projection $\tilde{P}^{(1)}$ and $\tilde{P}^{(2)}$, from the properties of Kronecker product, it is easy to show that

$$\hat{\mathbf{p}}^{(1)} \cdot \hat{\mathbf{p}}^{(2)} = \prod_{i=0}^{n-1} \mathbf{p}_i^{(1)} \cdot \mathbf{p}_i^{(2)} \tag{21}$$

"$\Longrightarrow$": if $\tilde{P}^{(1)}$ is orthogonal to $\tilde{P}^{(2)}$, by Definition 2.4 we have that $\hat{\mathbf{p}}^{(1)}$ is orthogonal to $\hat{\mathbf{p}}^{(2)}$. Therefore, Eq. 21 varnishes to zero. It is easy to see that it can not be zero if

$\mathbf{p}_i^{(1)} \cdot \mathbf{p}_i^{(2)} \neq 0$ for all $i = 0, \ldots, n-1$. Therefore, there exists at least one $i \in \{0, \ldots, n-1\}$ such that $\mathbf{p}_i^{(1)} \cdot \mathbf{p}_i^{(2)} = 0$. "$\Longleftarrow$": if for at least one $i \in \{0, \ldots, n-1\}$, $\mathbf{p}_i^{(1)} \cdot \mathbf{p}_i^{(2)} = 0$, from Eq. 21, we have $\hat{\mathbf{p}}^{(1)} \cdot \hat{\mathbf{p}}^{(2)} = 0$ thus $\hat{\mathbf{p}}^{(1)}$ is orthogonal to $\hat{\mathbf{p}}^{(2)}$, therefore by Definition 2.4, $\tilde{P}^{(1)}$ is orthogonal to $\tilde{P}^{(2)}$. Proposition 3.1 is proven. ∎

B. Derivation of the solution in Eq. 17

Proof We first formulate the Lagrangian multipliers, i.e.,

$$\begin{aligned} & L(\mathbf{p}_j^{(k)}, \lambda, \mu_0, \ldots, \mu_{k-1}) \\ = \; & (\mathbf{p}_j^{(k)})^T \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} - \lambda((\mathbf{p}_j^{(k)})^T \mathbf{A}_s^{(j)} \mathbf{p}_j^{(k)} - 1) \\ - \; & \mu_0 (\mathbf{p}_j^{(k)})^T \mathbf{p}_j^{(0)} - \ldots - \mu_{k-1} (\mathbf{p}_j^{(k)})^T \mathbf{p}_j^{(k-1)} \end{aligned} \quad (22)$$

Set the derivative of $L(\cdot)$ w.r.t. $\mathbf{p}_j^{(k)}$ to zero, we have

$$2\mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} - 2\lambda \mathbf{A}_s^{(j)} \mathbf{p}_j^{(k)} - \mu_0 \mathbf{p}_j^{(0)} \ldots - \mu_{k-1} \mathbf{p}_j^{(k-1)} = 0 \quad (23)$$

Multiplying Eq. 23 by $(\mathbf{p}_j^{(k)})^T$, we have

$$2(\mathbf{p}_j^{(k)})^T \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} - 2\lambda (\mathbf{p}_j^{(k)})^T \mathbf{A}_s^{(j)} \mathbf{p}_j^{(k)} = 0 \quad (24)$$

We then have $\lambda = \frac{(\mathbf{p}_j^{(k)})^T \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)}}{(\mathbf{p}_j^{(k)})^T \mathbf{A}_s^{(j)} \mathbf{p}_j^{(k)}}$, which is exactly the quantity we want to maximize.

Multiplying Eq. 23 by $(\mathbf{p}_j^{(l)})^T (\mathbf{A}_s^{(j)})^{-1}$ for each $l = 0, \ldots, k-1$, we obtain a set of k equations

$$\sum_{i=0}^{k-1} \mu_i (\mathbf{p}_j^{(l)})^T (\mathbf{A}_s^{(j)})^{-1} \mathbf{p}_j^{(i)} = 2(\mathbf{p}_j^{(l)})^T (\mathbf{A}_s^{(j)})^{-1} \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} \quad (25)$$

Denote $\mathbf{u} = [\mu_0, \mu_1, \ldots, \mu_{k-1}]^T$ and also use the notation in Eq 19 and Eq. 20, we can write the equation set in Eq. 25 more concisely in matrix format as

$$\mathcal{B}\mathbf{u} = 2\mathcal{A}^T (\mathbf{A}_s^{(j)})^{-1} \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} \quad (26)$$

Therefore

$$\mathbf{u} = 2\mathcal{B}^{-1} \mathcal{A}^T (\mathbf{A}_s^{(j)})^{-1} \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} \quad (27)$$

Multiplying Eq. 23 by $(\mathbf{A}_s^{(j)})^{-1}$, rearrange it to matrix form, we have

$$2(\mathbf{A}_s^{(j)})^{-1} \mathbf{A}_d^{(j)} \mathbf{p}_j^{(k)} - 2\lambda \mathbf{p}_j^{(k)} - (\mathbf{A}_s^{(j)})^{-1} \mathcal{A}\mathbf{u} = 0 \quad (28)$$

Embedding Eq. 27 into Eq. 28, we obtain Eq. 17. Since λ is exactly the quantity we want to maximize, we have the conclusion that the optimal solution of $\mathbf{p}_j^{(k)}$ is the Eigenvector corresponding the largest eigenvalue of matrix $\tilde{\mathcal{M}}$ defined in Eq. 18. ∎

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 19(7):711–720, 1997. Special Issue on Face Recognition.

[2] D. Cai, X. He, J. Han, and H.-J. Zhang. Orthogonal laplacianfaces for face recognition. *IEEE Transaction on Image Processing*, 15(11):3608–3614, November 2006.

[3] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *Proc. of IEEE Conf. on Computer Vision and Patter Recognition*, volume 2, pages 846–853, San Diego, CA, June 2005.

[4] H.-T. Chen, T.-L. Liu, and C.-S. Fuh. Learning effective image metrics from few pairwise examples. In *Proc. of IEEE International Conf. on Computer Vision*, pages 1371–1378, Beijing, China, Octobor 2005.

[5] J. Duchene and S. Leclercq. An optimal transformation for discriminant and principal component analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 10(6):978–983, November 1988.

[6] 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 Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[7] X. He, D. Cai, and P. Niyogi. Tensor subspace analysis. In *Advances in Neural Information Processing Systems*, volume 18, Vancouver, Canada, December 2005.

[8] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(3):328–340, March 2005.

[9] T. G. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–257, 2001.

[10] F. Samaria and A. Harter. Parameterization of a stochastic model for human face identification. In *Proc. of IEEE Workshop on Applications of Computer Vision*, pages 138–142, Sarasota, FL, USA, December 1994.

[11] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, December 2003.

[12] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proc. of IEEE Conf. on Computer Vision and Patter Recognition*, pages 586–591, June 1991.

[13] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, 1995.

[14] D. Xu, S. Lin, S. Yan, and X. Tang. Rank-one projections with adaptive margins for face recognition. In *Proc. of IEEE Conf. on Computer Vision and Patter Recognition*, volume 1, pages 175–181, New York City, NY, June 2006.

[15] S. Yan, D. Xu, L. Zhang, B. Zhang, and H. Zhang. Coupled kernel-based subspace learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 645–650, San Diego, CA, June 2005.