5

Multilinear Supervised Neighborhood Preserving Embedding Analysis of Local Descriptor Tensor

Xian-Hua Han and Yen-Wei Chen
Ritsumeikan University
Japan

1. Introduction

Subspace learning based pattern recognition methods have attracted considerable interests in recent years, including Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA), and some extensions for 2D analysis. However, a disadvantage of all these approaches is that they perform subspace analysis directly on the reshaped vector or matrix of pixel-level intensity, which is usually unstable under appearance variance. In this chapter, we propose to represent an image as a local descriptor tensor, which is a combination of the descriptor of local regions (K*K-pixel patch) in the image, and is more efficient than the popular Bag-Of-Feature (BOF) model for local descriptor combination. As we know that the idea of BOF is to quantize local invariant descriptors, e.g., obtained using some interest-point detector techniques by Harris & Stephens (1998), and a description with SIFT by Lowe (2004) into a set of visual words by Lazebnik et al. (2006). The frequency vector of the visual words then represents the image, and an inverted file system is used for efficient comparison of such BOFs. However. the BOF model approximately represents each local descriptor feature as a predefined visual word, and vectorizes the local descriptors of an image into a orderless histogram, which may lose some important (discriminant) information of local features and spatial information hold in the local regions of the image. Therefore, this paper proposes to combine the local features of an image as a descriptor tensor. Because the local descriptor tensor retains all information of local features, it will be more efficient for image representation than the BOF model and then can use a moderate amount of local regions to extract the descriptor for image representation, which will be more effective in computational time than the BOF model. For feature representation of image regions, SIFT proposed by Lowe (2004) is improved to be a powerful local descriptor by Lazebnik et al. (2006) for object or scene recognition, which is somewhat invariant to small illumination change. However, in some benchmark database such as YALE and PIE face data sets by Belhumeur et al. (1997), the illumination variance is very large. Then, in order to extract robust features invariant to large illumination, we explore an improved gradient (intensity-normalized gradient) of the image and use histogram of orientation weighed with the improved gradient for local region representation.

With the local descriptor tensor of image representation, we propose to use a tensor subspace analysis algorithm, which is called as multilinear Supervised Neighborhood Preserving Embedding (MSNPE), for discriminant feature extraction, and then use it for object or scene recognition. As we know, subspace learning approaches, such as PCA and LDA by Belhumeur et al. (1997), have widely used in computer vision research filed for feature extraction or selection and have been proven to be efficient for modeling or classification.

Recently there are considerable interests in geometrically motivated approaches to visual analysis. Therein, the most popular ones include locality preserving projection by He et al. (2005), neighborhood preserving embedding, and so on, which cannot only preserve the local structure between samples but also obtain acceptable recognition rates for face recognition. In real applications, all these subspace learning methods need to firstly reshape the multilinear data into a 1D vector for analysis, which usually suffers an overfitting problem. Therefore, some researchers proposed to solve the curse-of-dimension problem with 2D subspace learning such as 2-D PCA and 2-D LDA by ming Wang et al. (2009) for analyzing directly on a 2D image matrix, which was proven to be suitable in some extend. However, all of the conventional methods usually perform subspace analysis directly on the reshaped vector or matrix of pixel-level intensity, which would be unstable under illumination and background variance. In this paper, we propose MSNPE for discriminant feature extraction on the local descriptor tensor. Unlike tensor discriminant analysis by Wang (2006), which equally deals with the samples in the same category, the proposed MSNPE uses neighbor similarity in the same category as a weight of minimizing the cost function for N^{th} order tensor analysis, which is able to estimate geometrical and topological properties of the sub-manifold tensor from random points ("scattered data") lying on this unknown sub-manifold. In addition, compared with TensorFaces by Casilescu & D.Terzopoulos (2002) method, which also directly analyzes multi-dimensional data, the proposed multilinear supervised neighborhood preserving embedding uses supervised strategy and thus can extract more discriminant features for distinguishing different objects and, at the same time, can preserve samples' relationship of inner object instead of only dimension reduction in TensorFaces. We validate our proposed algorithm on different benchmark databases such as view-based object data sets (Coil-100 and Eth-70) and Facial image data sets (YALE and CMU PIE) by Belhumeur et al. (1997) and Sim et al. (2001).

2. Related work

In this section, we firstly briefly introduce the tensor algebra and then review subspace-based feature extraction approaches such as PCA, LPP.

Tensors are arrays of numbers which transform in certain ways under coordinate transformations. The order of a tensor $\mathcal{X} \in R^{N_1 \times N_2 \times \cdots \times N_M}$, represented by a multi-dimensional array of real numbers, is M. An element of $\mathcal{X}$ is denoted as $\mathcal{X}_{i_1,i_2,\cdots,i_M}$, where $1 \leq i_j \leq N_j$ and $1 \leq j \leq M$. In the tensor terminology, the mode-j vectors of the nth-order tensor $\mathcal{X}$ are the vectors in R^{N_j} obtained from $\mathcal{X}$ by varying the index i_j while keeping the other indices fixed. For example, the column vectors in a matrix are the mode-1 vectors and the row vectors in a matrix are the mode-2 vectors.

Definition. (**Modeproduct**). The tensor product $\mathcal{X}_{\times d}\mathbf{U}$ of tensor $\mathcal{X} \in R^{N_1 \times N_2 \times \cdots \times N_M}$ and a matrix $\mathbf{U} \in R^{N_d \times N'}$ is the $N_1 \times N_2 \times \cdots \times N_{d-1} \times N' \times N_{d+1} \times \cdots \times N_M$ tensor:

$$(\mathcal{X}_{\times d}\mathbf{U})_{i_1,i_2,\cdots,i_{d-1},j,i_{d+1},\cdots,i_M} = \sum_{i_d}(\mathcal{X}_{i_1,i_2,\cdots,i_{d-1},i_d,i_{d+1},\cdots,i_M}\mathbf{U}_{i_d,j}) \tag{1}$$

for all index values. $\mathcal{X}_{\times d}\mathbf{U}$ means the mode d's product of the tensor $\mathcal{X}$ with the matrix $\mathbf{U}$. The mode product is a special case of a contraction, which is defined for any two tensors not just for a tensor and a matrix. In this paper, we follow the definitions in Lathauwer (1997) and avoid the use of the term "contraction".

In tensor analysis, Principal Component Analysis (PCA) is used to extract the basis for each mode. The proposed MSNPE approach is based on the basis idea of Locality Preserving Projection (LPP). Therefore, we simply introduce PCA, LPP and a 2D extension of LPP as the following.

(1) Principal component analysis extracts the principal eigen-space associated with a set (matrix) $\mathbf{X} = [\mathbf{x}_i|_{i=1}^N]$ of training samples ($\mathbf{x}_i \in R^n$ with $1 \leq i \leq N$; N: sample number; n: dimension of the samples). Let $\mathbf{m}$ be the mean of the N training samples, and $\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$ be the covariance matrix of the $\mathbf{x}_i$. One solves the eigenvalue equation $\lambda \mathbf{u}_i = \mathbf{C}\mathbf{u}_i$ for eigenvalues $\lambda_i \geq 0$. The principal eigenspace $\mathbf{U}$ is spanned by the first K eigenvectors with the largest eigenvalues, $\mathbf{U} = [\mathbf{u}_i|_{i=1}^K]$. If $\mathbf{x}_t$ is a new feature vector, then it is projected to eigenspace $\mathbf{U}$: $\mathbf{y}_t = \mathbf{U}^T(\mathbf{x}_t - \mathbf{m})$. The vector $\mathbf{y}_t$ is used in place of $\mathbf{x}_t$ for representation and classification.

(2)Locality Preserving Projection: LPP seeks a linear transformation $\mathbf{P}$ to project high-dimensional data into a low-dimensional sub-manifold that preserves the local Structure of the data. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ denotes the set representing features of N training image samples, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N] = [\mathbf{P}^T\mathbf{x}_1, \mathbf{P}^T\mathbf{x}_2, \cdots, \mathbf{P}^T\mathbf{x}_N]$ denotes the samples feature in transformed subspace. Then, the linear transformation $\mathbf{P}$ can be obtained by solving the following minimization problem with some constraints, which will be given later:

$$\min_{\mathbf{P}} \sum_{ij} ||\mathbf{y}_i - \mathbf{y}_j||^2 W_{ij} = \min_{\mathbf{P}} \sum_{ij} ||\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j||^2 W_{ij} \tag{2}$$

where W_{ij} evaluate the local structure of the image space. It can be simply defined as follows:

$$W_{ij} = \begin{cases} 1 \text{ if } \mathbf{x}_i \text{ is among the } k \text{ nearest neighbors of } \mathbf{x}_j \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

By simple algebra formulation, the objective function can be reduced to:

$$\begin{aligned} \frac{1}{2}\sum_{ij}(\mathbf{P}^T\mathbf{x}_i - \mathbf{P}^T\mathbf{x}_j)^2 W_{ij} &= \sum_i \mathbf{P}^T\mathbf{x}_i D_{ii}\mathbf{x}_i^T\mathbf{P} - \sum_{ij}\mathbf{P}^T\mathbf{x}_i W_{ij}\mathbf{x}_i^T\mathbf{P} \\ &= \mathbf{P}^T\mathbf{X}(\mathbf{D} - \mathbf{W})\mathbf{X}^T\mathbf{P} = \mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P} \end{aligned} \tag{4}$$

where each column $\mathbf{P}_i$ of the LPP linear transformation matrix $\mathbf{P}$ can not be zero vector, and a constraint is imposed as follows:

$$\mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{I} \Rightarrow \mathbf{P}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P} = \mathbf{I} \tag{5}$$

where $\mathbf{I}$ in constraint term $\mathbf{P}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P} = \mathbf{I}$ or $\mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{I}$ is an identity matrix. $\mathbf{D}$ is a diagonal matrix; its entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, $D_{ii} = \sum_j W_{ij}$; $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix [5]. Matrix D provides a natural measure on the data samples. The bigger the value D_{ii} (corresponding to $\mathbf{y}_i$) is, the more importance is $\mathbf{y}_i$. The constraint for the sample $\mathbf{y}_i$ in $\mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{I}$ is $D_{ii} * \mathbf{y}_i^T\mathbf{y}_i = 1$, which means that the more importance (D_{ii} is larger) the sample $\mathbf{y}_i$ is, the smaller the value of $\mathbf{y}_i^T\mathbf{y}_i$ is. Therefor, the constraint $\mathbf{Y}^T\mathbf{D}\mathbf{Y} = \mathbf{I}$ will try to make the important point (has density distribution around the important point) near the origin of the projected subspace. Then, the density region near the origin of the

projected subspace includes most of the samples, which can make the objecrive function in Eq. (2) as small as possible, and at same time, can avoid the trivial solution $||\mathbf{P}_i||^2 = 0$ for the transformation matrix $\mathbf{P}$.

Then, The linear transformation $\mathbf{P}$ can be obtained by minimizing the objective function under constraint $\mathbf{P}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P} = \mathbf{I}$:

$$\underset{\mathbf{P}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P}=\mathbf{I}}{\mathbf{argmin}} \ \mathbf{P}^T\mathbf{X}(\mathbf{D}-\mathbf{W})\mathbf{X}^T\mathbf{P} \tag{6}$$

Finally, the minimization problem can be converted to solve a generalized eigenvalue problem as follows:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P} \tag{7}$$

In Face recognition application, He et al [8] extended LPP method into 2D dimension analysis, named as Tensor Subspace Analysis (TSA). TSA can directly deal with 2D gray images, and achieved better recognition results than the conventional 1D subspace learning methods such as PCA, LDA and LPP. However, for object recognition, color information also plays an important role for distinguishing different objects. Then, in this paper, we extend LPP to ND tensor analysis, which can directly deal with not only 3D Data but also ND data structure. At the same time, in order to obtain stable transformation tensor basis, we regularize a term in the proposed MSNPE objective function for abject recognition, which is introduced in Sec. 3 in detail.

3. Local descriptor tensor for image representation

In computer vision, local descriptors (i.e., features computed over limited spatial support) have been proven to be well-adapted for matching and recognition tasks as they are robust to partial visibility and clutter. The current popular one for a local descriptor is the SIFT feature, which is proposed by Lowe (2004). With the local SIFT descriptor, usually there are two types of algorithms for object recognition. One is to match the local points with SIFT features in two images, and the other one is to use the popular BOF model, which forms a frequency histogram of a predefined visual-words for all sampled region features by Belhumeur et al. (1997). For a matching algorithm, it is usually not enough to recognize the unknown image even if there are several points that are well matched. The popular BOF model usually can achieve good recognition performance in most applications such as scene and object recognition. However, in BOF model, in order to achieve an acceptable recognition rate, it is necessary to sample a lot of points for extracting SIFT features (usually more than 1000 in an image) and to compare the extracted local SIFT feature with the predefined visual words (usually more than 1000) to obtain the visual-word occurrence histogram. Therefore, BOF model needs a lot of computing time to extract visual-words occurrence histogram. In addition, BOF model just approximately represents each local region feature as a predefined visual-word; then, it may lose a lot of information and will be not efficient for image representation. Therefore, in this paper, we propose to represent a color or gray image as a combined local descriptor tensor, which can use different features (such as SIFT or other descriptors) for local region representation.

In order to extract the local descriptor tensor for image representation, we firstly grid-segment an image into K regions with some overlapping, and in each region, we extract some descriptors (can be consider tensor) for local region representation. For a gray image, a M-dimensional feature vector, which can be considered as a 1D tensor, is extracted from

the local gray region. For a color image, a M-dimensional feature vector can be extracted from each color channel such as R, G and B color channels. With the feature vectors of the three color channels, a combined 2D $M \times 3$ tensor can represent the local color region. Furthermore we combine the K 1D or 2D local tensor (M-dimensional vector or $M \times 3$ 2D tensor) into a 2D or 3D tensor with of size $M \times K \times L$ (L: 1 or 3). The tensor feature extraction procedure of a color image is shown in Fig. 1(a). For feature representation of the local regions such as the red, orange and green rectangles in Fig. 1 (a), the popular SIFT proposed by Lowe (2004) is proved to be a powerful one for object recognition, which is somewhat invariant to small illumination change. However, in some benchmark database such as YALE and CMU PIE face datasets, the illumination variance is very large. Then, in order to extract robust feature invariant to large illumination, we explore an normalized gradient (intensity-normalized gradient) of the image, and use Histogram of Orientation weighed with Normalized Gradient (NHOG) for local region representation. Therefore, for the benchmark databases without large illumination variance such as COIL-100 dataset or where the illumination information is also useful for recognition such as scene dataset, we use the popular SIFT for local region representation. However, for the benchmark database with large illumination variation, which will be harmful for subject recognition such as YALE and CMU PIE facial datasets, we use Histogram of Orientation weighed with Normalized Gradient (NHOG) for local region representation.

(1) SIFT: The SIFT descriptor computes a gradient orientation histogram within the support region. For each of 8 orientation planes, the gradient image is sampled over a 4 by 4 grid of locations, thus resulting in a 128-dimensional feature vector for each region. A Gaussian window function is used to assign a weight to the magnitude of each sample point. This makes the descriptor less sensitive to small changes in the position of the support region and puts more emphasis on the gradients that are near the center of the region. To obtain robustness to illumination changes, the descriptors are made invariant to illumination transformations of the form $aI(x) + b$ by scaling the norm of each descriptor to unity [8]. For representing the local region of a color image, we extract SIFT feature in each color component (R, G and B color components), and then can achieve a $128 * 3$ 2D tensor for each local region.

(2) Histogram of Orientation weighed with the Normalized Gradient (NHOG): Given an image $\mathbf{I}$, we calculate the improved gradient (Intensity-normalized gradient) using the following Eq.:

$$\begin{aligned} \mathbf{I}_x(i,j) &= \frac{\mathbf{I}(i+1,j) - \mathbf{I}(i-1,j)}{\mathbf{I}(i+1,j) + \mathbf{I}(i-1,j)} \\ \mathbf{I}_y(i,j) &= \frac{\mathbf{I}(i,j+1) - \mathbf{I}(i,j-1)}{\mathbf{I}(i,j+1) + \mathbf{I}(i,j-1)} \\ \mathbf{I}_{xy}(i,j) &= \sqrt{\mathbf{I}_x(i,j)^2 + \mathbf{I}_y(i,j)^2} \end{aligned} \tag{8}$$

where $\mathbf{I}_x(i,j)$ and $\mathbf{I}_y(i,j)$ mean the horizontal and vertical gradient in pixel position i,j, respectively, $\mathbf{I}_{xy}(i,j)$ means the global gradient in pixel position i,j. The idea of the normalized gradient is from χ^2 distance: a normalized Euclidean distance. For x-direction, the gradient is normalized by summation of the upper one and the bottom one pixel centered by the focused pixel; for y-direction, the gradient is normalized by that of the right and left one. With the intensity-normalized gradient, we can extract robust and invariant features to illumination changing in a local region of an image. Some examples with the intensity-normalized and conventional gradients are shown in Fig. 2

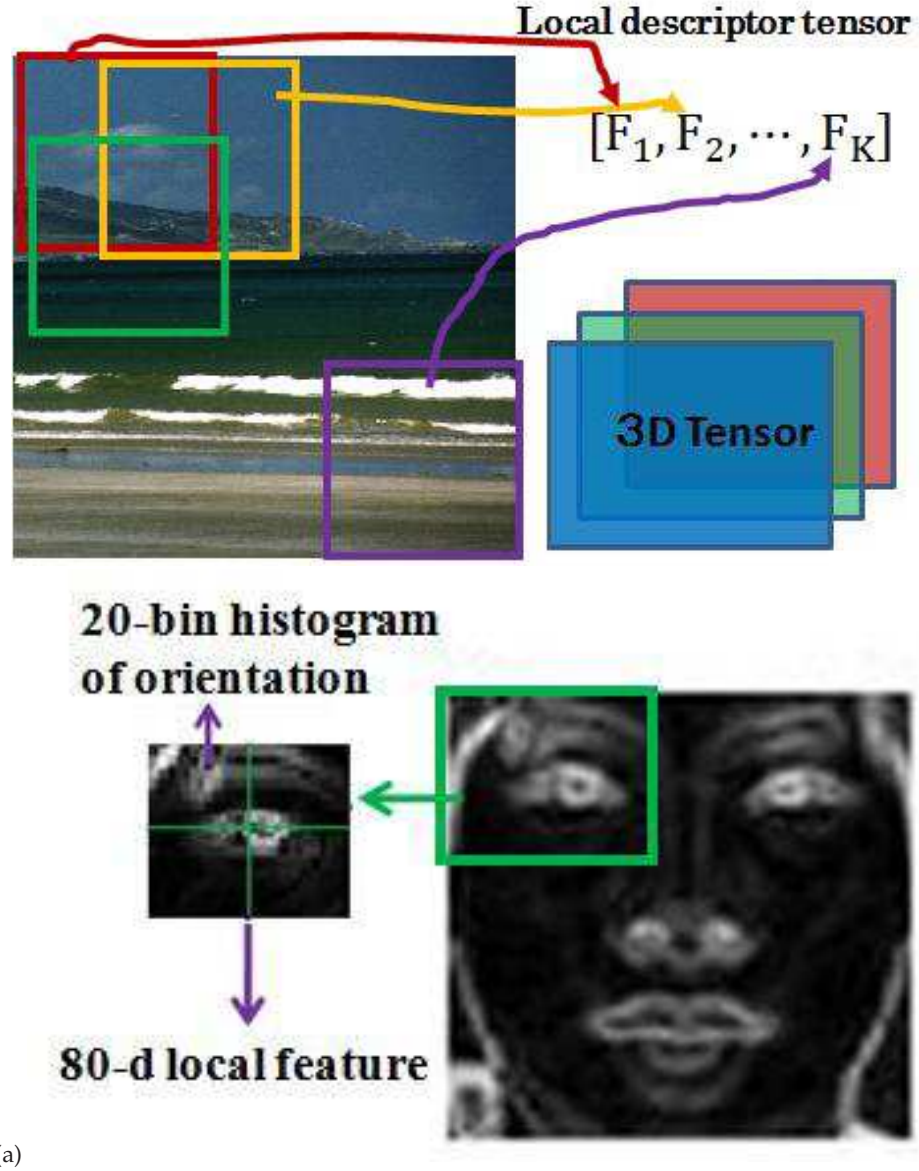


Fig. 1. (a) Extraction of local descriptor tensor for color image representation; (b)NHOG feature extraction from a gray region.

(a)Samples of YALE facial database

(b)Samples of PIE facial database

Fig. 2. Gradient image samples. Top row: Original face images; Middle row: the intensity-normalized gradient images; Bottom row: the conventional gradient images.

For feature extraction of a local region I^R in the normalized gradient image shown in Fig. 1(b), we firstly segment the region into 4 (2×2) patches,and then in each patch extract a 20-bin histogram of orientation weighted by global gradient $\mathbf{I}_{xy}^R$ calculated using the intensity-normalized gradients $\mathbf{I}_x^R$, $\mathbf{I}_y^R$. Therefore, each region in a gray image can be represented by 80-bin (20×4) histogram as shown in Fig. 1(b).

4. Multilinear supervised neighborhood preserving embedding

In order to model *N*-Dimensional data without rasterization, tensor representation is proposed and analyzed for feature extraction or modeling. In this section, we propose a multilinear supervised neighborhood preserving embedding by Han et al. (2011) Han et al.

(2011) to not only extract discriminant feature but also preserve the local geometrical and topological properties in same category for recognition. The proposed approach decompose each mode of tensor with objective function, which consider neighborhood relation and class label of training samples.

Suppose we have ND tensor objects $\mathcal{X}$ from C classes. The c^{th} class has n^c tensor objects and the total number of tensor objects is n. Let $\mathcal{X}_{i_c} \in R^{N_1 \times N_2 \times \cdots \times N_L} (i_c = 1, 2, \cdots, n^c)$ be the i^{th} object in the c^{th} class. For color object image tensor, L is 3, N_1 is the row number, N_2 is the column number, and N_3 is the color space components (N_3=3). We can build a nearest neighbor graph $\mathcal{G}$ to model the local geometrical structure and label information of $\mathcal{X}$. Let $\mathbf{W}$ be the weight matrix of $\mathcal{G}$. A possible definition of $\mathbf{W}$ is as follows:

$$W_{ij} = \begin{cases} exp^{-\frac{\|\mathcal{X}_i - \mathcal{X}_j\|^2}{t}} & \text{if sample } i \text{ and } j \text{ is in same class} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $\|\mathcal{X}_i - \mathcal{X}_j\|^2$ means Euclidean distance of two tensor, which is the summation square root of all corresponding elements between $\mathcal{X}_i$ and $\mathcal{X}_j$, and $\| \bullet \|$ means l_2 norm in our paper.

Let $\mathbf{U}_d$ be the d-mode transformation matrices (Dimension: $N_d \times N'_d$). A reasonable transformation respecting the graph structure can be obtained by solving the following objective functions:

$$\min_{\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_L} \frac{1}{2} \sum_{ij} \|\mathcal{X}_{i \times 1} \mathbf{U}_{1 \times 2} \mathbf{U}_2 \cdots {}_{\times L} \mathbf{U}_L - \mathcal{X}_{j \times 1} \mathbf{U}_{1 \times 2} \mathbf{U}_2 \cdots {}_{\times L} \mathbf{U}_L \|_2 W_{ij} \tag{10}$$

Algorithm 1: ND tensor supervised neighborhood embedding

Input: Tensor objects $\mathcal{X}_i^c$ from C classes, $\mathcal{X}_i^c$ denots the i^{th} tensor object in the c_{th} class
Graph-based weights: Building nearest neighbor graph in same class and calculate the graph weight **W** according to Eq. 9 and **D** from **W**
Initialize: Randomly initialize $\mathbf{U}_r^d \in R^{N_d}$ for d =1,2,$\cdots$, L
for t=1:T (Iteration steps) or until converge **do**
  for d=1:L (Iteration steps) **do**
    • Calculate $\mathbf{D}_d$ and $\mathbf{S}_d$ assuming $\mathbf{U}_i (i = 1, 2, \cdots, d-1, d+1, \cdots, L)$ fixed.
    • Solve the minimizing problem: $\min_{\mathbf{U}_d} tr(\mathbf{U}_d^T (\mathbf{D}_d - \mathbf{S}_d) U_d)$ with eigenspace analysis
  end for
end for
output: the MSNPE tensor $\mathcal{T}_j = \mathbf{U}_1 \times \mathbf{U}_2 \times \cdots \times \mathbf{U}_L, j = 1, 2, \cdots, (N'_1 \times N'_2 \times \cdots \times N'_L)$.

Table 1. The flowchart of multilinear supervised neighborhood preserving embedding (MSNPE).

where $\mathcal{X}_i$ is the tensor representation of the i^{th} sample; $\mathcal{X}_{i\times 1}\mathbf{U}_1$ means the mode 1's product of the tensor $\mathcal{X}_i$ with the matrix $\mathbf{U}_1$, and $\mathcal{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2$ means the mode 2's product of the tensor $\mathcal{X}_{i\times 1}\mathbf{U}_1$ with the matrix $\mathbf{U}_2$, and so on. The above objective function incurs a heavy penalty if neighboring points of same class $\mathcal{X}_i$ and $\mathcal{X}_j$ are mapped far apart. Therefore, minimizing it is an attempt to ensure that if $\mathcal{X}_i$ and $\mathcal{X}_j$ are ″close″, then $\mathcal{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2\cdots_{\times L}\mathbf{U}_L$ and $\mathcal{X}_{j\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2\cdots_{\times L}\mathbf{U}_L$ are ″close″ as well. Let $\mathcal{Y}_i = \mathcal{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2\cdots_{\times L}\mathbf{U}_L$ with dimension $N_1\times N_2\times\cdots\times N_L$, and $(\mathbf{Y}_i)^d = (\mathcal{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d$ with dimension:$N_d\times(N_1\times N_2\times\cdots\times N_{d-1}\times N_{d+1}\times\cdots\times N_L)$ is the d-mode extension of tensor $\mathcal{Y}_i$, which is a 2D matrix. Let D be a diagonal matrix,$D_{ii}=\sum_j W_{ij}$. Since $\|\mathbf{A}\|^2 = tr(\mathbf{A}\mathbf{A}^T)$, we see that

$$
\begin{aligned}
&\frac{1}{2}\sum_{ij}\|\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times L}\mathbf{U}_L-\mathcal{X}_{j\times 1}\mathbf{U}_1\cdots_{\times L}\mathbf{U}_L\|^2 W_{ij}\\
=&\frac{1}{2}\sum_{ij}tr(((\mathbf{Y}_i)^d-(\mathbf{Y}_j)^d)((\mathbf{Y}_i)^d-(\mathbf{Y}_j)^d)^T)W_{ij}\\
=&tr(\sum_i D_{ii}(\mathbf{Y}_i)^d((\mathbf{Y}_i)^d)^T-\sum_{ij}W_{ij}(\mathbf{Y}_i)^d((\mathbf{Y}_j)^d)^T)\\
=&tr(\sum_i D_{ii}(\mathbf{U}_d^T(\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d\\
&\qquad((\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d)^T\mathbf{U}_d\\
&-\sum_{ij}W_{ij}(\mathbf{U}_d^T(\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d\\
&\qquad((\mathcal{X}_{j\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d)^T\mathbf{U}_d)\\
=&tr(\mathbf{U}_d^T(\sum_i D_{ii}((\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d\\
&\qquad((\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d)^T\\
&-\sum_{ij}W_{ij}((\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d\\
&\qquad((\mathcal{X}_{j\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d)^T)\mathbf{U}_d)\\
=&tr(\mathbf{U}_d^T(\mathbf{D}_d-\mathbf{S}_d)\mathbf{U}_d)
\end{aligned}
\tag{11}
$$

where $\mathbf{D}_d = \sum_i D_{ii}(\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\ \mathbf{U}_{d+1}\cdots_{\times L}\ \mathbf{U}_L)^d\ ((\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\ \mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d)^T$ and $\mathbf{S}_d = \sum_{ij}W_{ij}(\mathcal{X}_{i\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\ \mathbf{U}_L)^d\ ((\mathcal{X}_{j\times 1}\mathbf{U}_1\cdots_{\times d-1}\mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1}\cdots_{\times L}\mathbf{U}_L)^d)^T$. In optimization procedure of each mode, we also impose a constraint to achieve the transformation matrix (such as $\mathbf{U}_d$ in mode d) as the following:

$$\mathbf{U}_d^T\mathbf{Y}^d\mathbf{D}(\mathbf{Y}^d)^T\mathbf{U}_d=\mathbf{I}\Rightarrow\mathbf{U}^T\mathbf{D}_d\mathbf{U}=\mathbf{I} \tag{12}$$

For the optimization problem of all modes, we adopt an alternative least square (ALS) approach. In ALS, we can obtain the optimal base vectors on one mode by fixing the base vectors on the other modes and cycle for the remaining variables. The d-mode transformation matrix $\mathbf{U}_d$ can be achieved by minimizing the following cost function:

$$\underset{\mathbf{U}_d^T\mathbf{D}_d\mathbf{U}_d=\mathbf{I}}{\mathbf{argmin}}\ \mathbf{U}_d^T(\mathbf{D}_d-\mathbf{S}_d)\mathbf{U}_d \tag{13}$$

In order to achieve the stable solution, we firstly regularize the symmetric matrix $\mathbf{D}_d$ as $\mathbf{D}_d = \mathbf{D}_d + \alpha\mathbf{I}$ (α is a small value, $\mathbf{I}$ is an identity matrix of same size with the matrix $\mathbf{D}_d$). Then, the minimization problem for obtaining d-mode matrix can be converted to solve a generalized eigenvalue problem as follows:

$$(\mathbf{D}_d - \mathbf{S}_d)\mathbf{U}_d = \lambda\mathbf{D}_d\mathbf{U}_d \tag{14}$$

We can select the corresponding generalized eigenvectors with the first N'_d smaller eigenvalues in Eq.(14), which can minimize the objective function in Eq.(13). However, the eigenvectors with the smallest eigenvalues are usually unstable. Therefore, we convert Eq. (14) into:

$$\mathbf{S}_d\mathbf{U}_d = (1-\lambda)\mathbf{D}_d\mathbf{U}_d \Rightarrow \mathbf{S}_d\mathbf{U}_d = \beta\mathbf{D}_d\mathbf{U}_d \tag{15}$$

The corresponding generalized eigenvectors with the first N'_d smaller eigenvalues λ in Eq. (14) means those with the first N'_d larger eigenvalues $\beta(1-\lambda)$ in Eq. (15). Therefore, the corresponding generalized eigenvectors with the first N'_d larger eigenvalues can be selected for minimizing the objective function in Eq.(13). The details algorithm of MSNPE are listed in Algorithm 1. In MSNPE algorithm, we need to decide the retained number of the generalized eigenvectors (mode dimension) for each mode. Usually, the dimension numbers in most discriminant tensor analysis methods are decided empirically or according to applications. In our experiments, we retain different dimension numbers for different modes, and do recognition for objects or scene categories. The recognition accuracy with varied dimensions in different modes are also given in the experiment part. The dimension numbers is decided empirically in the compared results with the state-of-art algorithms.

After obtaining the MSNPE basis of each mode, we can project each tensor object into these MSNPE tensors. For classification, the projection coefficients can represent the extracted feature vectors and can be inputted into any other classification algorithm. In our work, beside Euclidean distance as KNN (k=1) classifier, we also use Random Forest (RF) for recognition.

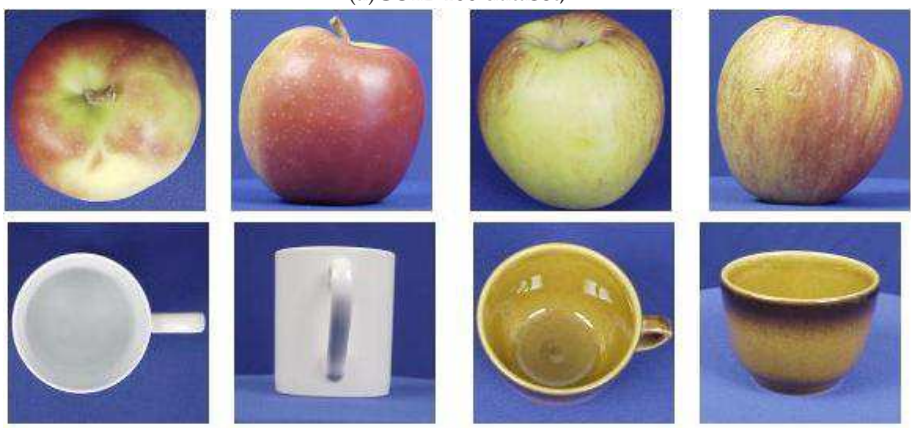
5. Experiments

5.1 Database

We evaluated our proposed framework on two different types of datasets.

(i) View-based object datasets, which includes two datasets: The first one is the Columbia COIL-100 image library by Nene et al. (1996). It consists of color images of 72 different views of 100 objects. The images were obtained by placing the objects on a turntable and taking a view every 5°. The objects have a wide variety of complex geometric and reflectance characteristics. Fig. 3(a) shows some sample images from COIL-100. The second one is the ETH Zurich CogVis ETH-80 dataset by Leibe & Schiele (2003a). This dataset was setup by Leibe and Schiele to explore the capabilities of different features for object class recognition. In this dataset, eight object categories including apple, pear, tomato, cow, dog, horse, cup and car have been collected. There are 10 different objects spanned large intra-class variance in each category. Each object has 41 images from viewpoints spaced equally over the upper viewing hemisphere. On the whole we have 3280 images, 41 images for each object and 10 object for each category. Fig.3(b) shows some sample images from ETH-80.

(a)COIL-100 dataset;



(b) ETH80 dataset;

Fig. 3. Sample images from view-based object data sets.

(ii) Facial dataset: We use two facial datasets for evaluating the tensor representation with the proposed NHOG for image representation. One is Yale databae which includes 15 people and 11 facial images of each individual with different illuminations and expressions. Some sample facial images are shown in the top row of Fig. 2(a). The other one is CMU PIE, which includes 68 people and about 170 facial images for each individual with 13 different poses, 43 different illumination conditions, and with 4 different expressions. Some sample facial images are shown in the top row of Fig. 2(b).

5.2 Methodology

The recognition task is to assign each test image to one of a number of categories or objects. The performance is measured using recognition rates.

For view-based object databases, we take different experimental setup in COIL-100 and ETH80 datasets. For COIL-100, the objective is to discriminate between the 100 individual

objects. In most previous experiments on object recognition using COIL-100, the number of views used as training set for each object varied from 36 to 4. When 36 views are used for training, the recognition rate using SVM was reported approaching 100% by Pontil & Verri (1998). In practice, however, only very few views of an object are available. In our experiment, in order to compare experimental results with those by Wang (2006), we follows the experiment setup, which used only 4 views of each object for training and the rest 68 views for testing. In total it is equivalent to 400 images for training and 6800 images for testing. The error rate is the overall error rate over 100 objects. The 4 training viewpoints are sampled evenly from the 72 viewpoints, which can capture enough variance on the change of viewpoints for tensor learning. For ETH-80, it aims to discriminate between the 8 object categories. Most previous experiments using ETH-80 dataset all adopted leave-one-object-out cross-validation. The training set consists of all views from 9 objects from each category. The testing set consists of all views from the remaining object from each category. In this setting, objects in the testing set have not appeared in the training set, but those belonging to the same category have. Classification of a test image is a process of labeling the image by one of the categories. Reported results are based on average error rate over all 80 possible test objects by Leibe & Schiele (2003b). Similar to the above, instead of taking all possible views of each object in the training set, we take only 5 views of each object as training data. By doing so we have decreased the number of the training data to 1/8 of that used by Leibe & Schiele (2003b), Marrr et al. (2005). The testing set consists of all the views of an object. The recognition rate with the proposed scheme is compared to those of different conventional approaches by Wang (2006) and those with MSNPE analysis directly on pixel-level intensity tensor.

For facial dataset, which has large illumination variance in images, we validate that the tensor representation with the proposed NHOG for image representation will be much more efficient for face recognition than that with the popular SIFT descriptor, which only is somewhat robust to small illumination variance. In experiments Yale dataset, we randomly select 2, 3, 4 and 5 facial images from each individual for training, and the remainders for test. For CMU PIE dataset, we randomly select 5 and 10 facial images from each individual for training, and the remainder for test. We do 20 runs for different training number and average recognition rate in all experiments. The recognitions with our proposed approach are compared to those by the state-of-art algorithm by Cai et al. (2007a), Cai et al. (2007b).

6. Experimental results

(1) View-based object data sets

We investigate the performance of the proposed MSNPE tensor learning compared with conventional tensor analysis such as tensor LDA by Wang (2006), which is also used in view-base object recognition, and the efficiency of the proposed tensor representation compared to the pixel-level intensity tensor, which directly consider a whole image as a tensor, on COIL-100 and ETH80 datasets. In these experiments, all samples are also color images, and SIFT descriptor for local region representation is used. Therefore, the pixel-level intensity tensor is 3rd tensor with dimension $R1 \times C1 \times 3$, where $R1$ and $C1$ is row and column number of the image, and the local descriptor tensor is with $128 \times K \times 3$, where K is the segmented region number of an image (here K=128). In order to compare with the state-of-art works by Wang (2006), simple KNN method (k=1 in our experiments) is also used for recognition. Experimental setup was given in Sec. 5, and we did 18 runs so that all samples can be as test. Figure 6(a) shows the compared results of MSNPE using pixel-level tensor and local descriptor tensor (denoted MSNPE-PL and MSNPE with KNN classifier, respectively, MSNPE-RF-PL and MSNPE-RF with random forest) and traditional methods by Wang (2006) on COIL-100

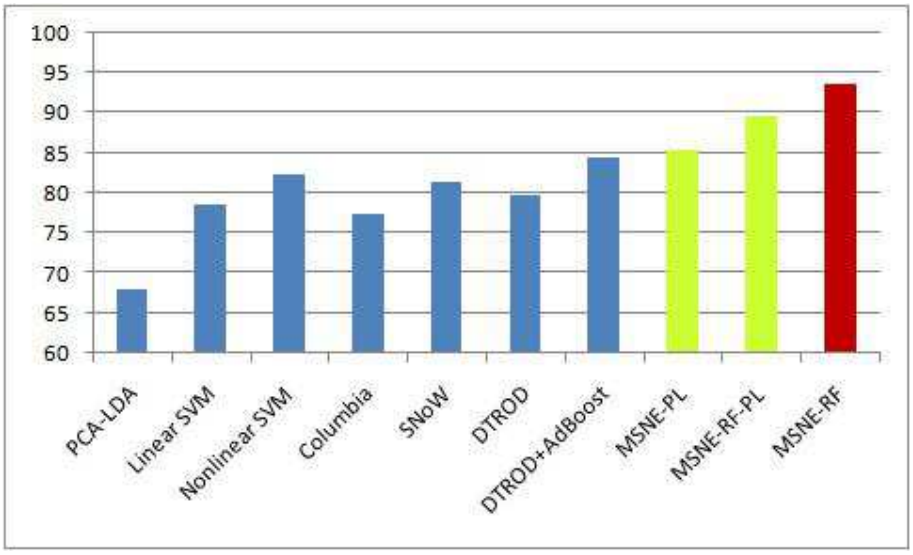
Methods	DTROD	DTROD+AdaB	RSW	LS
Rate(%)	70.0	76.0	75.0	65.0
Methods	MSNPE-PL	RF-PL	MSNPE	RF
Rate(%)	76.83	77.74	83.54	85.98

Table 2. The compared recognition rates on ETH-80. RSW denotes random subwindow method Marrr et al. (2005) and LS denotes the results from Leibe and Schiele Leibe & Schiele (2003b) with 2925 samples for training and 328 for testing. The others are with 360 samples for training. MSNPE-PL and RF-PL mean MSNPE analysis on pixel-level intensity tensor using simple Euclidean distance and random forest classifier, respectively; MSNPE and RF mean the proposed MSNPE analysis on local SIFT tensor using simple Euclidean distance random forest classifier, respectively.

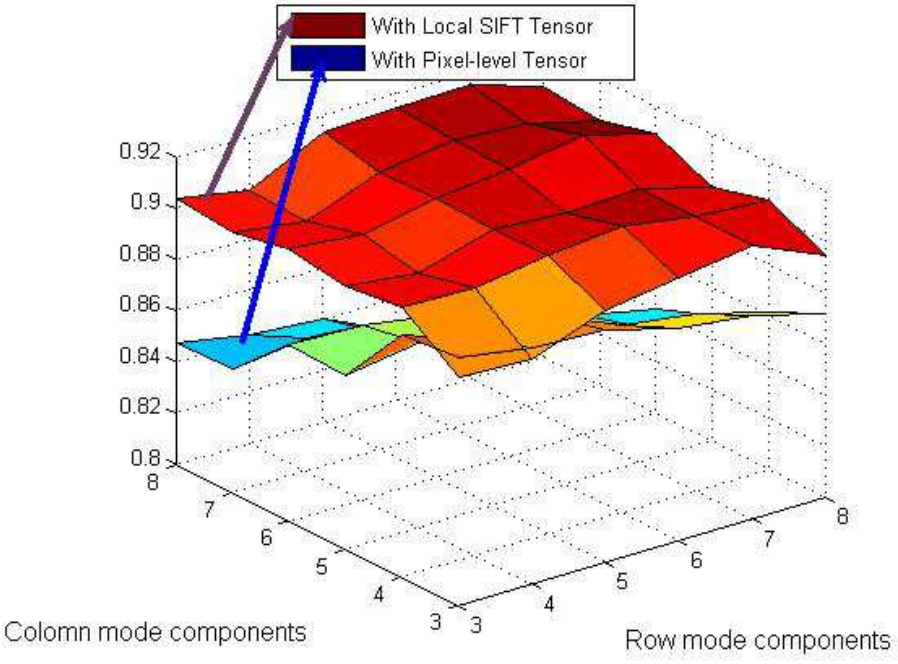
dataset . The best result with the same experiment setup (400 training samples and 6800 test samples) on COIL-100 is reported by Wang (2006), in which the average recognition rate using tensor LDA and AdBoost classifier (DTROD+AdBoost) is 84.5%, and the recognition rate of the tensor LDA and simple Euclidean distance(DTROD) by Wang (2006) (same as KNN method with k=1) is 79.7. However, The MSNPE approach with pixel-intensity tensor can achieve about 85.28% with same classifier (KNN), and 90% average recognition rate with random forest classifier. Furthermore, the MSNPE approach with local SIFT tensor achieved 93.68% average recognition rate. The compared recognition rate results with the state-of-art approaches are shown in Fig. 4 (a). Figure 4(b) shows the compared recognition rates of one run on different mode dimension of MSNPE between using pixel-level intensity and local SIFT tensor with random forest classifier. It is obvious that the recognition rates by using pixel-level tensor have very large variance with differen mode dimension changing. Therefore, we must select a optimized row and column mode dimension to achieve better recognition rate. However, it is usually difficult to decide the optimized dimension number of different modes automatically. If we just shift the mode dimension number a little from the optimized mode dimension, the recognition rate can be decreased significantly shown in Fig. 4(b) when using pixel-level tensor. For the local SIFT tensor representing an object image, the average recognition rates in lager mode dimension changing (Both row and column mode dimension numbers are from 3 to 8; color mode dimension is 3) are very stable.

For ETH-80 dataset, we also do similar experiments to COIL-100 using the proposed MSNPE analysis with pixel-level and local SIFT tensor, respectively. The compared results with the state of the art approach are shown in Table 2. From Table 2, it can be seen that our proposed approach can greatly improve the overall recognition rate compared with the state of the art method (from 60-80% to about 86%).

(2) Facial Datasets: With the two used facial datasets, we investigate the efficiency of the proposed local NHOG feature on large illumination variance dataset compared with local SIFT descriptor. We do 20 runs for different training number and average recognition rate. For comparison, we also do experiments using the proposed MSNPE analysis directly on the gray face image (pixel-level intensity, denoted MSNPE-PL), local feature tensor with SIFT descriptor (denoted MSNPE-SIFT) and our proposed intensity-normalized histogram of orientation (denoted MSNPE-NHOG). Table 3 gives the compared results using MSNPE analysis with different tensors using KNN classifier (k=1) and other subspace learning methods by Cai et al. (2007a), Cai et al. (2007b), Cai (2009) and Cai (n.d.) on YALE dataset, and the compared results on CMU PIE dataset are shown in Table 4 with our proposed framework and the conventional ones by Cai et al. (2007a) Cai et al. (2007b) Cai (2009) Cai



(a)



(b)

Fig. 4. (a) The compared recognition rates on COIL-100 between the proposed framework and the state-of-art approaches Wang (2006). (b) Average recognition rate with different mode dimension using random forest classifier.

Method	2 Train	3 Train	4 Train	5 Train
PCA	56.5	51.1	57.8	45.6
LDA	54.3	35.5	27.3	22.5
Laplacianface	43.5	31.5	25.4	21.7
O-Laplacianface	44.3	29.9	22.7	17.9
TensorLPP	54.5	42.8	37	32.7
R-LDA	42.1	28.6	21.6	17.4
S-LDA	37.5	25.6	19.7	14.9
MSNPE	41.89	31.67	24.86	23.06
MSNPE-SIFT	35.22	26.33	22.19	20.83
MSNPE-NHOG	**29.74**	**22.87**	**18.52**	**17.44**

Table 3. Average recognition error rates (%) on YALE dataset with different training number.

Method	5 Train	10 Train
PCA	75.33	65.5
LDA	42.8	29.7
LPP	38	29.6
MSNPE	37.66	23.57
MSNPE-NHOG	**33.85**	**22.06**

Table 4. Average recognition error rates (%) on PIE dataset with different training number.

(n.d.). From Table 3 and 4, it is obvious that our proposed algorithm can achieve the best recognition performances for all most cases, and the recognition rate improvements become greater when the training sample number is small compared to those by the conventional subspace learning methods by Cai et al. (2007a), Cai et al. (2007b), Cai (2009) and Cai (n.d.). In addition, as we have shown in the previous section, our proposed strategy can be applied not only for recognition of face with small variance (such as mainly frontal face database), but also for recognition of generic object with large variance. With generic object dataset with large variance, the recognition rates are also improved greatly compared with using pixel-level tensor.

7. Conclusion

In this paper, we proposed to represent an image as a local descriptor tensor, which is a combination of the descriptor of local regions ($K * K$-pixel patch) in the image, and more efficient than the popular Bag-Of-Feature (BOF) model for local descriptor combination, and at the same time, we explored a local descriptor for region representation for databases with large illumination variance, Which is improved to be more efficient than the popular SIFT descriptor. Furthermore, we proposed to use Multilinear Supervised Neighborhood Preserving Embedding (MSNPE) for discriminant feature extraction from the local descriptor tensor of different images, which can preserve local sample structure in feature space. We validate our proposed algorithm on different Benchmark databases such as view-based and facial datasets, and experimental results show recognition rate with our method can be greatly improved compared conventional subspace analysis methods.

8. References

Belhumeur, P. N., Hepanha, J. P. & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (7): 711–720.

Cai, D. (2009). Spectral regression: A regression framework for efficient regularized subspace learning.

Cai, D. (n.d.). *http://www.zjucadcg.cn/dengcai/Data/Yale/results_new.html* .

Cai, D., He, X., Hu, Y., Han, J. & Huang, T. (2007a). Learning a spatially smooth subspace for face recognition, *CVPR* .

Cai, D., He, X., Hu, Y., Han, J. & Huang, T. (2007b). Spectral regression for efficient regularized subspace learning, *ICCV* .

Casilescu, M. & D.Terzopoulos (2002). Multilinear analysis of image ensembles: Tensorfaces, *ECCV* .

Han, X.-H., Qiao, X. & wei Chen, Y. (2011). Multilinear supervised neighborhood embedding with local descriptor tensor for face recognition, *IEICE Trans. Inf. & Syst.* .

Han, X.-H., wei Chen, Y. & Ruan, X. (2011). Multilinear supervised neighborhood embedding of local descriptor tensor for scene/object recognition, *IEEE Transaction on Image Processing* .

Harris, C. & Stephens, M. (1998). A combined corner and edge detector, *In Proc. Alvey Vision Conference* .

He, X., Yan, S., Hu, Y., Niyogi, P. & Zhang, H.-J. (2005). Face recognition using laplacianfaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (3): 328–340.

Lazebnik, S., Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *CVPR* pp. 2169–2178.

Leibe, B. & Schiele, B. (2003a). Analyzing appearance and contour based methods for object categorization, *CVPR* .

Leibe, B. & Schiele, B. (2003b). Analyzing appearance and contour based methods for object categorization, *CVPR* .

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* (2): 91–110.

Marrr, R., Geurts, P., Piater, J. & Wehenkel, L. (2005). Random subwindows for robust image classification, *CVPR* pp. 34–40.

ming Wang, X., Huang, C., ying Fang, X. & gao Liu, J. (2009). 2dpca vs. 2dlda: Face recognition using two-dimensional method, *International Conference on Artificial Intelligence and Computational Intelligence* pp. 357–360.

Nene, S. A., Nayar, S. K. & Murase, H. (1996). Columbia object image library (coil-100), *Technical Report CUCS-006-96* .

Pontil, M. & Verri, A. (1998). Support vector machines for 3d object recognition, *PAMI* pp. 637–646.

Sim, T., Baker, S. & Bsat, M. (2001). The cmu pose, illumination, and expression (pie) database of human faces, *Robotics Institute, CMU-RI-TR-01-02, Pittsburgh, PA,* .

Wang, Y. & Gong, S. (2006). Tensor discriminant analysis for view-based object recognition, *ICPR* pp. 439–454.

L.D. Lathauwer. Signal processing based on multilinear algebra, *Ph.D. Thesis*, Katholike Universiteit Leu- ven.

Principal Component Analysis
Edited by Dr. Parinya Sanguansat

ISBN 978-953-51-0195-6
Hard cover, 300 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

This book is aimed at raising awareness of researchers, scientists and engineers on the benefits of Principal Component Analysis (PCA) in data analysis. In this book, the reader will find the applications of PCA in fields such as image processing, biometric, face recognition and speech processing. It also includes the core concepts and the state-of-the-art methods in data analysis and feature extraction.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Xian-Hua Han and Yen-Wei Chen (2012). Multilinear Supervised Neighborhood Preserving Embedding Analysis of Local Descriptor Tensor, Principal Component Analysis, Dr. Parinya Sanguansat (Ed.), ISBN: 978-953-51-0195-6, InTech, Available from: http://www.intechopen.com/books/principal-component-analysis/multilinear-supervised-neighborhood-preserving-embedding-analysis-of-local-descriptor-tensor

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821