Face Recognition Using Multilinear Manifold Analysis of Local Descriptors

Xian-Hua Han and Yen-Wei Chen

Ritsumeikan University, 1-1-1, NojiHigashi, Kusatsu, Shiga, 525-8577, Japan

Abstract. In this paper, we propose to represent a face image as a local descriptor tensor and use a Multilinear Manifold Analysis (MMA) method for discriminant feature extraction, which is used for face recognition. The local descriptor tensor, which is a combination of the descriptor of local regions (K*K-pixel patch) in the image, can represent image more efficient than pixel-level intensity representation, and also than the popular Bag-Of-Feature (BOF) model, which approximately represents each local descriptor as a predefined visual word. Therefore it should be more effective in computational time than the BOF model. For extracting discriminant and compact features from the local descriptor tensor, we propose to use the proposed TMultilinear Manifold Analysis (MMA) algorithm, which has several benefits compared with conventional subspace learning methods such as PCA, ICA, LDA and so on: (1) a natural way of representing data without losing structure information, i.e., the information about the relative positions of pixels or regions; (2) a reduction in the small sample size problem which occurs in conventional supervised learning because the number of training samples is much less than the dimensionality of the feature space; (3) a neighborhood structure preserving in tensor feature space for face recognition and a good convergence property in training procedure. We validate our proposed algorithm on Benchmark database Yale and PIE, and experimental results show recognition rate with the proposed method can be greatly improved compared with conventional subspace analysis methods especially for small training sample number. ...

1 Introduction

Many face recognition techniques have been developed over the past few decades. One of the most successful and well-studied face recognition techniques is the appearance-based method [1,2]. When using appearance-based methods, an image of size $n_1 \times n_2$ pixels is usually represented by a vector in an $n_1 \times n_2$-dimensional space. In practice, however, these $n_1 \times n_2$-dimensional spaces are too large to allow robust and fast face recognition. Previous works have demonstrated that the face recognition performance can be improved significantly in lower dimensional linear subspaces [2-3]. Two of the most popular appearance-based face recognition methods include Eigenface [2] and Fisherface. Eigenface is based on Principal Component Analysis (PCA). PCA projects the face images along the directions of maximal variances. It also aims to preserve the Euclidean

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 734–742, 2012.

distances between face images. Fisherface is based on Linear Discriminant Analysis (LDA) [2]. Unlike PCA which is unsupervised, LDA is supervised. When the class information is available, LDA can be used to find a linear subspace which is optimal for discrimination. Recently there is considerable interest in geometrically motivated approaches to visual analysis. Therein, the most popular ones include Locality Preserving Projection (LPP) [3], Neighborhood Preserving Embedding (NPE) and so on, which can not only preserve the local structure between samples, and also obtain acceptable recognition rate for face recognition. In real application, all these subspace learning methods need to firstly reshape the 2D face image into 1D vector for analysis, which usually suffers "curse of dimension". Therefore, some researchers proposed to solve the "curse of dimension" problem with 2D subspace learning such as 2D-PCA, 2D-LDA [4] for analyzing directly on 2D image matrix, which was improved to be suitable to some extent. 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 or pose variance.

In this paper, we propose to represent a face 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 [5] for local descriptor combination. In order to extract discriminant feature from the local regions, we explore an improved gradient (intensity-normalized gradient) of the face image, which is robust to illumination variance, and use histogram of orientation weighed with the improved gradient for local region representation. Furthermore, we propose to use a multilinear subspace learning algorithm for discriminant feature extraction from the local descriptor tensor of face images, which can preserve local sample structure in feature space. Compared with tensorfaces [6] method which also directly analyze multi-dimensional data, the proposed MMA uses supervised strategy, and thus can extract more discriminant features for distinguishing different objects (here facial images of different persons) and at the same time, can preserve samples' relationship of inner-person instead of only dimension reduction in tensorfaces. We validate our proposed algorithm on benchmark database Yale[2] and CMU PIE[7], and experimental results show recognition rate with our method can be greatly improved compared conventional subspace analysis methods especially for small training sample number.

The remaining parts of this paper are organized as follows. We introduce the local descriptor tensor for face images in section 2. Section 3 propose a Multilinear Manifold Analysis (MMA) for extracting discriminant feature for face representation. Finally, we report experiment setup and results in section 4, and give conclusion remarks in section 5.

2 Local Descriptor Tensor for Face Image Representation

In computer vision, local descriptors (i.e. features computed over limited spatial support) have proved well-adapted for matching and recognition tasks, as

Fig. 1. Gradient image samples. Top row: Original face images; Middle row: the intensity-normalized gradient images; Bottom row: the conventional gradient images.

they are robust to partial visibility and clutter. The current popular one for local descriptor is SIFT feature, which is proposed by in [11] and is robust to small illumination variance. However with large illumination variance usually appeared in face recognition, it is still difficult to recognize correctly, and achieve acceptable recognition rate. Therefore, we proposed a histogram of orientation weighted with the improved gradient for local image representation. With the local descriptor, usually there are two types of algorithms for object recognition. One is to match the local point with SIFT feature in two images, and the other one is to use the popular Bag-Of-Feature model (BOF), which forms a frequency histogram of a predefined visual-words for all sampled region features [5]. For matching algorithm, it is usually not enough to recognize the unknown image even if there are several points well matched. How to combine more features is not unsolved still. 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 acceptable recognition rate it is necessary to sample a lot of points for extracting SIFT features (usually more than 1000 in an image), and compare the extracted local feature with the predefined visual-words (Usually more than 1000) to obtain the visual-word occurrence histogram. Therefore, BOF model need a lot of computing time to extract visual-words occurrence histogram. In addition, BOF model just approximately represent each local region feature as the predefined visual-words, and then, it maybe lose a lot of information and will be not efficient for image representation. Therefor, in this paper, we propose to represent a face image as a combined local descriptor tensor.

In our work, we combine two types of local features as a tensor for face image representation: SIFT feature and a intensity-Normalized Histogram of Orientation Gradient–NHOG.

(1) 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 a$\mathbf{I}$(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 $\times$3 2D tensor for each local region.

(2) In order to extract robust feature to illumination variance, we need to obtain the improved gradient. Given an image $\mathbf{I}$, we calculate the improved gradient (Intensity-normalized gradient) using the following Eq.:

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

where $\mathbf{I}_x(i,j)$ and $\mathbf{I}_y(i,j)$ means 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. 1. The local NHOG feature can be extracted as shown in Fig. 2. given a local region I^R in an face image, we firstly segment the region into 4 (2×2) patches,and in each patch, we extract a 20-bin histogram of orientation weighted by global gradient using the intensity-normalized gradients $\mathbf{I}_x^R$, $\mathbf{I}_y^R$ and $\mathbf{I}_{xy}^R$. Therefore, each region in a face image can be represent by 80-bin (20×4) histogram as shown in the left part of Fig. 2.

In order to efficiently represent a face image, we combine the extracted local SIFT or NHOG descriptors for face image representation. Firestly, we grid-segment an image, and can obtain $M2$ overlapping regions as shown in the right part of Fig. 2, and then in each region, we extract a L-dimension (128 for SIFT, 80 for NHOG) local feature (1D tensor). Furthermore we combine the $M2$ vectors (local descriptors) into a 2D tensor with of size $L \times M2$ in the space $R_{128or80} \bigotimes R_{M2}$ for representing a face image. The tensor NHPG feature extraction procedure of a face image is shown in Fig. 2.

3 Multilinear Manifold Analysis

In order to model N-D data without rasterization (2D is a special case), tensor representation is proposed and analyzed for feature extraction or modeling. In

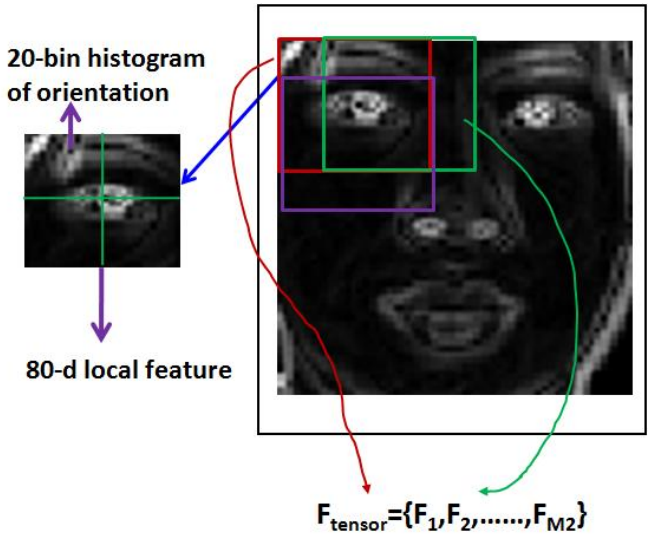

Fig. 2. Extraction procedure of local descriptor tensor from a face image. The red rectangle in the right part of this figure is the first extracted region for calculating local descriptor (a 80-bin edge histogram); The green rectangle is the next extracted region after moving several pixels from the red one (predefined interval) along row, and continue this step until the end of row pixels. The purple rectangle is the first extracted region after moving several pixels for the red one along column, and then obtain next regions through moving pixel in row. The total number of extracted regions is $M2$.

this section, we propose a tensor supervised neighborhood embedding 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 model 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 $\mathcal{X}_{i_c} \in R^{N_1} \bigotimes R^{N_2} \bigotimes \cdots \bigotimes R^{N_L} (i_c = 1, 2, \cdots, n^c)$ be the i^{th} object in the c^{th} class. For a gray face image, we can directly represent it as pixel-level intensity tensor, where L is 2, N_1 is the row number, N_2 is the column number. We also can represent the face image as a feature-based tensor such as local descriptor feature tensor introduced in Sec. 2, where L is also 2, N_1 is the local feature dimension, N_2 is the sampled region number in an image. Then, we can build a nearest neighbor graph $\mathcal{G}$ to model the local geometrical structure and label information of $\mathcal{X}$. Let W be the weight matrix of $\mathcal{G}$. A possible definition of W is as follows:

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

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

$$\min_{\mathbf{U}_1,\mathbf{U}_2,\cdots,\mathbf{U}_L} \sum_{ij} \|\mathcal{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2 \cdots_{\times L} \mathbf{U}_L - \mathcal{X}_{j\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2 \cdots_{\times L} \mathbf{U}_L\| W_{ij} \tag{3}$$

The objective function incurs a heavy penalty if neighboring points $\mathcal{X}_i$ 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 $\mathcal{Y}_i = \mathcal{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2 \cdots_{\times L} \mathbf{U}_L$ (Dimension: $D_1 \times D_2 \times \cdots \times D_L$) , and $(\mathbf{Y}_i)^d = (\mathbf{X}_{i\times 1}\mathbf{U}_{1\times 2}\mathbf{U}_2 \cdots_{\times L} \mathbf{U}_L)^d$ (2D matrix, Dimension:$D_d \times (D_1 \times D_2 \times \cdots \times D_{d-1} \times D_{d+1} \times \cdots \times D_L)$) is the d-mode extension of tensor $\mathcal{Y}_i$. 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

$$\begin{aligned}
&\frac{1}{2}\sum_{ij} \|\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times L} \mathbf{U}_L - \mathcal{X}_{j\times 1}\mathbf{U}_1 \cdots_{\times L} \mathbf{U}_L\| W_{ij} \\
=&\frac{1}{2}\sum_{ij} tr(((\mathbf{Y}_i)^d - (\mathbf{Y}_j)^d)((\mathbf{Y}_i)^d - (\mathbf{Y}_j)^d)^T) W_{ij} \\
=&tr(\sum_i D_{ii}(\mathbf{Y}_i)^d((\mathbf{Y}_i)^d)^T - \sum_{ij} W_{ij}(\mathbf{Y}_i)^d((\mathbf{Y}_j)^d)^T) \\
=&tr(\mathbf{U}_d^T(\sum_i D_{ii}((\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L) \\
&\qquad (\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L)^T \\
&\quad - \sum_{ij} W_{ij}((\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L) \\
&\qquad (\mathcal{X}_{j\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L)^T)U_d) \\
=&tr(\mathbf{U}_d^T(\mathbf{D}_d - \mathbf{S}_d)U_d)
\end{aligned} \tag{4}$$

where $\mathbf{D}_d = \sum_i D_{ii}((\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L)\ (\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L)^T$ and $\mathbf{S}_d = \sum_{ij} W_{ij}((\mathcal{X}_{i\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L)\ (\mathcal{X}_{j\times 1}\mathbf{U}_1 \cdots_{\times d-1} \mathbf{U}_{d-1\times d+1}\mathbf{U}_{d+1} \cdots_{\times L} \mathbf{U}_L)^T$.Therefore the linear transformation $\mathbf{U}_d$ can be obtained by minimizing the objective function under constraint:

$$\mathbf{U}_d = \underset{\mathbf{U}_d^T\mathbf{D}_d\mathbf{U}_d=1}{\mathbf{argmin}}\ (\mathbf{U}_d^T(\mathbf{D}_d - \mathbf{S}_d)U_d) \tag{5}$$

In order to achieve the stable solution, we firstly regularize the symmetric matrix $\mathbf{D}$ as $D_{ii} = D_{ii} + \alpha$ (α is a small value). Finally, the minimization problem can be converted to solving a generalized eigenvalue problem as follows:

$$\mathbf{D}_d U_d = \lambda \mathbf{S}_d U_d \tag{6}$$

After obtaining the MMA basis of each mode, we can project each tensor object into these MMA basis for each mode. For face recognition, the projection coefficients can represent the extracted feature vectors and can be used for classification using Euclidean distance or other similar measurement.

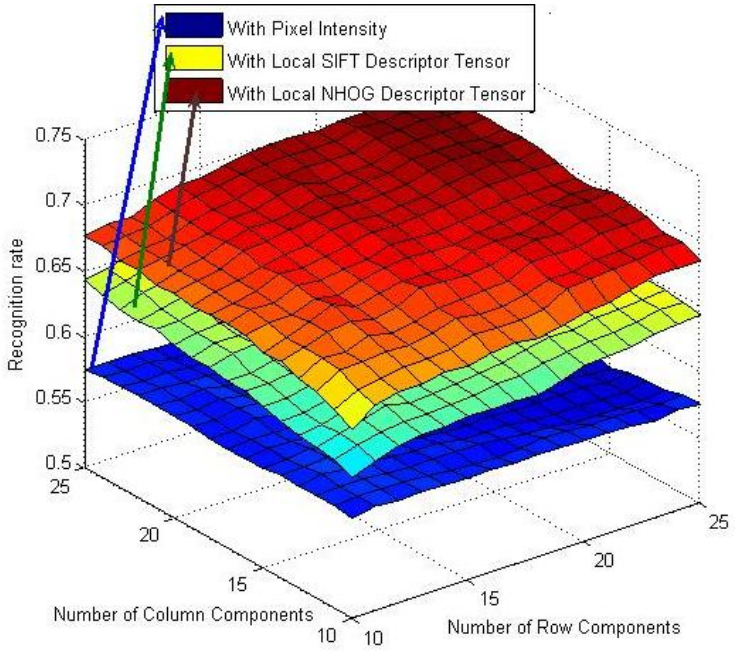

Fig. 3. Compared recognition rates using MMA for feature extraction with pixel-intensity tensor, local SIFT tensor and local NHOG tensor, respectively. X-axis denotes the number of retained row-mode components of the used tensor; Y-axis denotes the number of retained column-mode components of the used tensor.

4 Experimental Results

In this paper, we use the benchmark face dataset YALE, which includes 15 people and 11 facial images of each individual with different illuminations and expressions, and 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. For YALE dataset, we randomly select 2, 3, 4 and 5 facial images from each individual for training, and the remainders for test. We do 20 runs for different training number and average recognition rate. For comparison, we also do experiments using the proposed MMA analysis directly on the gray face image (pixel-level intensity, denoted MMA), and our proposed local descriptor tensor with SIFT descriptor (denoted MMA-SIFT) and intensity-Normalized Histogram of Orientation Gradient (denoted MMA-NHOG). Figure 3 gives the compared recognition rates after discriminant and compact feature extraction by the proposed MMA with the three types tensor (Pixel-intensity tensor, Local SIFT tensor and NHOG tensor), respectively. It is obvious from Fig. 3 that the proposed two local descriptor tensor representations for face image can achieve much higher recognition rates than those directly with pixel intensity tensor on any extracted feature number, and then the recognition

Table 1. Average recognition error rates (%) on YALE dataset with different training number

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
MMA	41.89	31.67	24.86	23.06
MMA-SIFT	35.22	26.33	22.19	20.83
MMA-NHOG	**29.74**	**22.87**	**18.52**	**17.44**

Table 2. Average recognition error rates (%) on YALE dataset with different training number

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

rates with the proposed NHOG feature, which is robust to large illumination variance, are better than those with SIFT feature, which just can deal with small illumination variation. In order to validate our proposed MMA algorithm with conventional subspace learning methods, we also give the compared results shown in Table 1 using MMA analysis with different tensors and the state-of-art subspace learning methods by He [3,9,10]. From Table 1, it is obvious that our proposed algorithm can obtain the best recognition performance especially using small training samples. For CMU PIE dataset, we randomly select 5 and 10 facial images from each individual for training, and the remainder for test. We also do 20 runs for achieving average recognition error rate. The compared recognition error rates between our proposed algorithms and the conventional subspace learning methods by He [3,9,10] are shown in Table 2.

5 Conclusions

In this paper, we proposed to represent a face 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. Furthermore, we proposed to use Multilinear Manifold Analysis (MMA) for discriminant feature extraction from the local descriptor tensor of face images, which can preserve local sample structure in feature space. We validate our proposed algorithm on Benchmark database Yale and PIE, and experimental results show recognition rate with our method can be

greatly improved compared conventional subspace analysis methods especially for small training sample numbers.

Acknowledgments. This work was supported in part by the Grant-in Aid for Scientific Research from the Japanese MEXT under the Grant No. 2430076, 24700179 and in part by the Research Matching Fund for Private Universities from MEXT (Ministry of Education, Culture, Sports, Science, and Technology).

References

1. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-d objects from appearance. International Journal of Computer Vision 14(1), 5–24 (1995)
2. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
3. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.-J.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)
4. Wang, X.-M., Huang, C., Fang, X.-Y., Liu, J.-G.: 2DPCA vs. 2DLDA: Face Recognition Using Two-Dimensional Method. In: International Conference on Artificial Intelligence and Computational Intelligence, vol. 2, pp. 357–360 (2009)
5. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categoraization with bags of keypoints. In: Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–16.
6. Vasilescu, M.A.O., Terzopoulos, D.: Multilinear Analysis of Image Ensembles: TensorFaces. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 447–460. Springer, Heidelberg (2002)
7. Sim, T., Baker, S., Bsat, M.: The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces. Robotics Institute, CMU-RI-TR-01-02, Pittsburgh, PA (2001)
8. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Cai, D., He, X., Hu, Y., Han, J., Huang, T.: Learning a Spatially Smooth Subspace for Face Recognition. In: CVPR 2007 (2007)
10. Cai, D., He, X., Han, J.: Spectral Regression for Efficient Regularized Subspace Learning. In: ICCV 2007 (2007)