Volume 45, Issue 12
December 2012
ISSN 0031-3203
ELSEVIER
PATTERN RECOGNITION
Available online at www.sciencedirect.com
SciVerse ScienceDirect

Pattern Recognition 45 (2012) 4438–4450

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

The small sample size problem of ICA: A comparative study and analysis

Weihong Deng [a],*, Yebin Liu [b], Jiani Hu [a], Jun Guo [a]

[a] School of Information and Telecommunications Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China
[b] Department of Automation, Tsinghua University, Beijing 100084, People's Republic of China

ARTICLE INFO

Article history:
Received 21 October 2011
Received in revised form
4 June 2012
Accepted 16 June 2012
Available online 26 June 2012

Keywords:
Projection pursuit
Independent component analysis
Whitened principal component analysis
Locality pursuit
Face recognition

ABSTRACT

On the small sample size problems such as appearance-based recognition, empirical studies have shown that ICA projections have trivial effect on improving the recognition performance over whitened PCA. However, what causes the ineffectiveness of ICA is still an open question. In this study, we find out that this small sample size problem of ICA is caused by a special distributional phenomenon of the high-dimensional whitened data: all data points are similarly distant, and nearly perpendicular to each other. In this situation, ICA algorithms tend to extract the independent features simply by the projections that isolate single or very few samples apart and congregate all other samples around the origin, without any concern on the clustering structure. Our comparative study further shows that the ICA projections usually produce misleading features, whose generalization ability is generally worse than those derived by random projections. Thus, further selection of the ICA features is possibly meaningless. To address the difficulty in pursuing low-dimensional features, we introduce a locality pursuit approach which applies the locality preserving projections in the high-dimensional whitened space. Experimental results show that the locality pursuit performs better than ICA and other conventional approaches, such as Eigenfaces, Laplacianfaces, and Fisherfaces.

1. Introduction

In pattern classification, feature extraction is defined as a mapping from a typically high-dimensional data space to space of reduced dimension while preserving the class separability [1]. PCA and ICA are the two most widely used unsupervised feature extraction techniques. PCA minimizes second-order dependency of the input data to find the basis along which the data (when projected onto them) have maximal variance. ICA minimizes both second-order and higher-order dependencies to find the basis along which the data are statistically independent. PCA is optimal for gaussian signals only, because it neglects the extra information contained in the higher-order statistics. In contrast, ICA uses this higher-order statistical information and is good at describing nonGaussian data.

In the area of appearance-based face recognition, Bartlett et al. claimed that a lot of important information might be contained in the high-order relationships among features (pixels) [2], and thus ICA was commonly considered as a more powerful tool than PCA. Several studies have been conducted for face recognition using ICA algorithm, namely independent Gabor feature method [3], for enhanced ICA by selecting PCA dimension [4]. ICA were also combined with LDA for face recognition [5] and gender classification [6]. Although most empirical studies [2–4] have claimed ICA is better than PCA for feature extraction in the high-dimensional classification system, some studies [7,8] reported contradictory results.

In high-dimensional applications, the ICA pipeline actually contains PCA process (for dimension reduction), whitening process (for scale normalization), and pure ICA process.[1] Yang et al. used the "PCA+whitening" (whitened PCA) as the baseline to revaluate the ICA-based face recognition systems, and the experimental results showed that the performance of ICA is nearly identical to that of whitened PCA [13]. In other words, pure ICA projection has trivial effect on the recognition performance. Based on similar experimental results, Vicente et al. [14] further pointed out that, if all the ICA projections are used, the feature vector derived by ICA is just a rotation of the whitened data, which is meaningless for classification. Therefore, the contradictory results between PCA and ICA can be explained by the effect of whitening process on different data sets. On many data sets, whitening process is effective to improve PCA-based recognition performance [15–18]. The studies used these data sets would report ICA is better than PCA. In some other cases, however, whitening process would lead to overfitting, hence it is not surprising that ICA is inferior.

* Corresponding author. Tel.: +86 10 62283059; fax: +86 10 62285019.
E-mail addresses: whdeng@bupt.edu.cn, whdeng@it.usyd.edu.au (W. Deng).

[1] Throughout this paper, we use FastICA as a representative of various ICA algorithms. Previous studies have shown that the performance difference between FastICA and other ICA implementations, such as Informax [9] and Common's algorithm [10], is not significant [11–13].

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.06.010

The equivalence between ICA and whitened PCA is based on the special condition that all the extracted ICA projections are used for classification. In general, ICA is commonly considered a variant of projection pursuit, and a subset of ICA projections can be selected for classification. The usefulness of ICA projections for pattern recognition is often illustrated by some toy samples like Fig. 1, where the projection direction with maximum non Gaussianity clearly highlight the clustered structure of the data. The projection on the first principal component, on the other hand, fails to show this structure. Hence, it is widely believed that selecting a subset of the ICA projections for feature extraction can significantly improve the classification performance [14]. However, the low-dimensional examples like Fig. 1 are not sufficient to verify the effectiveness of ICA on the high-dimensional applications such as the appearance-based recognition, because the high-dimensional data have fundamentally different distribution property from the low dimensional ones.

In this paper, we reveal a *small sample size problem of ICA*: For the high-dimensional data sets, ICA algorithms tend to extract the independent features simply by the projections that isolate single or very few samples apart and congregate all other samples around the origin, without any concern on the clustering structure. To address the difficulty in pursuing low-dimensional features, we introduce two alternative approaches: random pursuit and locality pursuit (LP). Further, we perform a comparative study on ICA, random pursuit, locality pursuit, as well as other state-of-the art dimension reduction methods. Specifically, the contributions of this paper are as follows:

1. We justify that under small sample size condition, *the pairwise distances of the high-dimensional whitened data points are identical*. In other words, the whitening procedure can strictly uniform the pairwise distance between samples, regardless of the intrinsic distribution of the data. As illustrated in Fig. 2, there are three images in the high-dimensional space, and the whitening process maps them onto the vertexes of a equilateral triangle. This finding of the *equaldistant whitened space* unveils a special property of the whitening procedure in the small sample size situation, which brings a new understanding of whitening process beyond the data scaling.
2. We show that the failure of ICA roots from the similarly distant data distribution in the high-dimensional whitened space, where the non-Gaussianity measures of ICA tends to derive the projection directions that isolate a very small number of (even one) data point apart and collapse the others near the origin. To convictively evaluate the applicability of ICA projections, we apply a random projections based algorithm as the baseline, and empirically find that *the ICA projections are less discriminative than the random projections in the high-dimensional whitened space*, which indicates that the ICA model is somehow misleading (worse than random) for high-dimensional classification problems.

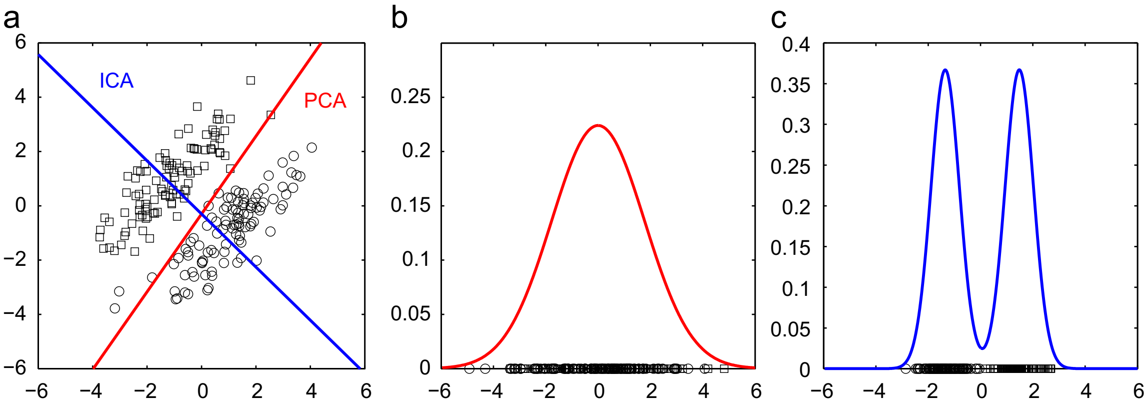

Fig. 1. An illustration of projection pursuit and the "interestingness" of non-Gaussian projections. The data in this figure is clearly divided into two clusters. However, the principal component, i.e. the direction of maximum variance, provides no separation between the clusters. In contrast, the strongly non-Gaussian projection pursuit direction provides optimal separation of the clusters. (a) Projection directions of PCA and ICA. (b) PCA projections. (c) ICA projections.

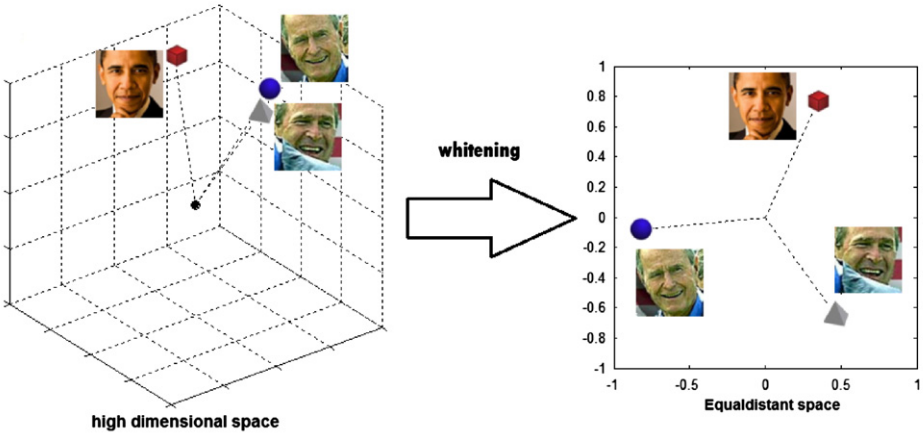

Fig. 2. Illustration of the high-dimensional whitening. The whitening has an unique effects on the high-dimensional data: uniforming the pairwise distance.

3. We introduce a locality pursuit method which applies the locality preserving criterion [19] to seek the classification-oriented projections in the high-dimensional whitened space. To address the distance concentration problem, we further propose a generalized heat kernel, which extends the L_2 distance metric of the heat kernel to the L_f distance metrics (f can be any positive number), to weight the pairwise distance in the locality measure. Experimental results show that the generalized heat kernel weights can improve the recognition performance.

We conduct a comparative study on appearance-based recognition using the FERET (face), AR (face), COIL-20 (object) and USPS (handwriting digit) data sets, which show the locality criterion is much better than the non-Gaussianity criterion in terms of both the computational efficiency and the recognition accuracy. In particular, on all the data sets, LP outperforms ICA by a margin of 10%–60% when using low-dimensional features. On the face data sets, the superiority of the LP algorithm over the manifold-based methods such as Laplacianfaces and UDP is shown by both the better recognition accuracy with lower feature dimension, and the stability against varying neighborhood size. Moreover, its performance is even comparable to the supervised feature extraction methods such as Fisherfaces, Enhanced FLD Model, Direct-LDA, Null-space LDA and LDA/GSVD.

The remainder of the paper is organized as follows. Section 2 studies the distributional properties of the high-dimensional whitened data. Section 3 explains why the ICA model is not applicable to the classification problems. Section 4 details the LP algorithm and discusses its differences with the manifold-based methods. Section 5 compares ICA and the LP algorithm on the face, object and digit recognition tasks. Finally, Section 6 concludes this paper.

2. High-dimensional whitening: beyond data scaling

Whitening is a data preprocessing method that takes the form of a linear transformation considering both the data scale and the correlations amongst the variables. A zero-mean random vector is said to be white if its elements are uncorrelated and have unit variance. Hence, the whitening transformation is usually performed by first decorrelating the data using PCA and then scaling each principal direction to uniform the spread of the data. The more complex whitening, also called whitened PCA, involves reduction of the data dimensionality in the PCA stage, which may lead to better performance by reducing the noisy components.

2.1. SVD based high-dimensional whitening

Let $X=[x_1,x_2,\ldots,x_n] \in \mathbb{R}^{m\times n}$ be the data matrix, and the centered data matrix can be expressed as follows:

$$\widehat{X} = X(I_n-(1/n)1_n1_n^T) \tag{1}$$

where $I_n \in \mathbb{R}^{n\times n}$ is an identical matrix, and $1_n \in \mathbb{R}^{n\times 1}$ is a vector with all elements of 1. A basic assumption of our study is that the n data points in the data matrix $X \in \mathbb{R}$ are linear independent, and thus we have $\text{rank}(X)=n$ and $\text{rank}(\widehat{X})=n-1$. The assumption is commonly made on the small sample size problems, and Ye [20] has shown that this assumption is satisfied on the various high-dimensional data sets, such as images, texts, and microarrays, when the data dimensionality is much larger than the sample size, i.e. $m>n$.

For the high-dimensional data, whitening transformation can be performed efficiently by compact singular value decomposition (SVD) on the centered data matrix [14]. Suppose the compact SVD of $\widehat{X}$ is

$$\widehat{X} = X(I_n-(1/n)1_n1_n^T) = UDV^T \tag{2}$$

where $D \in \mathbb{R}^{(n-1)\times(n-1)}$ is the diagonal matrix with the non-zero singular values as diagonal elements, and the columns of $U \in \mathbb{R}^{m\times(n-1)}$ and $V \in \mathbb{R}^{n\times(n-1)}$ are the corresponding left and right singular vectors, respectively. The whitened data matrix $\tilde{X}$ can be derived directly from the compact SVD as follows:

$$\tilde{X} = V^T = D^{-1}U^T\widehat{X} \tag{3}$$

Hence, the columns of $V^T=[v_1,v_2,\ldots,v_n] \in \mathbb{R}^{(n-1)\times n}$ are the whitened sample vectors.

2.2. The equaldistant whitened space

Theorem 2.1 (*Equaldistant Whitened Space*). *Assuming the n sample vectors are linearly dependent, the $n-1$ dimensional whitened sample vectors are equally distant.*

Proof. According to the properties of SVD [21], the range (column) space of $\widehat{X}^T$ is the $n-1$ dimensional subspace spanned by the columns of V, i.e.

$$\text{ran}(\widehat{X}^T) = \text{span}(V) \tag{4}$$

Meanwhile, in light of the basic linear algebra [22], the range space of $\widehat{X}^T$ equals the orthogonal complement of the null space of the $\widehat{X}$, i.e.

$$\text{ran}(\widehat{X}^T) = \text{null}(\widehat{X})^{\perp} \tag{5}$$

Combining (4) and (5), we have

$$\text{span}(V) = \text{null}(\widehat{X})^{\perp} \tag{6}$$

On one hand, $\text{null}(\widehat{X}) = \text{span}\{1_n^T\}$ due to $\{x|\widehat{X}x=0\}=\{x=1_n^T\}$, and thus the orthogonal projection onto $\text{null}(\widehat{X})$ is $P_1 = x(x^Tx)^{-1}x=(1/n)1_n1_n^T$. On the other hand, in light of the property of SVD with $V^TV=I$, the matrix $P_2=V(V^TV)^{-1}V^T=VV^T$ is the orthogonal projection onto $\text{span}(V)$. Note that the orthogonal projection onto a subspace is unique. Since $\text{span}(V)=\text{null}(\tilde{X})^{\perp}$, we have $P_2=I-P_1$, i.e.

$$VV^T = I-(1/n)1_n1_n^T \tag{7}$$

The matrix VV^T characterizes the geometry of the whitened sample vectors. Specifically, the inner product of any two whitened vectors can be calculated as follows:

$$v_i^Tv_j = \begin{cases} 1-\dfrac{1}{n} & \text{if } i=j \text{ (diagonal element of } VV^T) \\ -\dfrac{1}{n} & \text{if } i\neq j \text{ (non-diagonal element of } VV^T) \end{cases} \tag{8}$$

In light of the above property, we can easily find that the length of any whitened vector is

$$\|v_i\|_2 = \sqrt{v_i^Tv_i} = \sqrt{1-\frac{1}{n}} \tag{9}$$

The angle between two whitened vectors is

$$\angle(v_i,v_j) = \arccos\left(\frac{v_i^Tv_j}{\|v_i\|\|v_j\|}\right) = \arccos\left(-\frac{1}{n-1}\right) \tag{10}$$

Since the number of samples n is usually large, the whitened vectors are nearly perpendicular to each other. The Euclidean

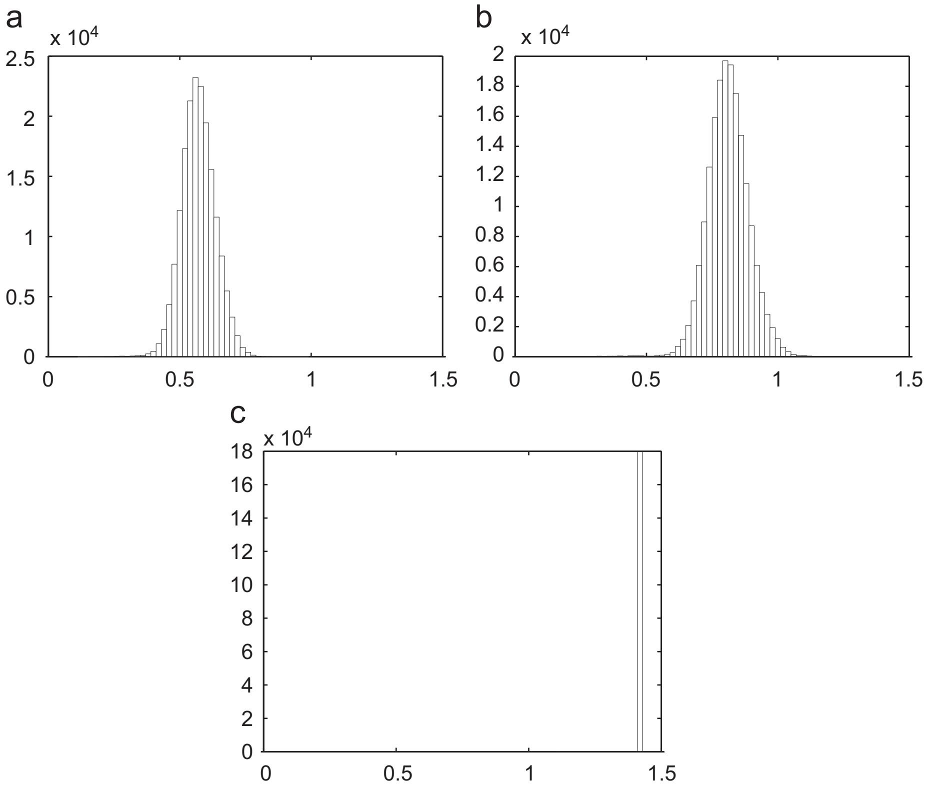

Fig. 3. The histogram of the pairwise distances between the 600 FERET facial images in varying dimensional whitened PCA space. (a) 100 dimensions. (b) 200 dimensions. (c) 599 dimensions.

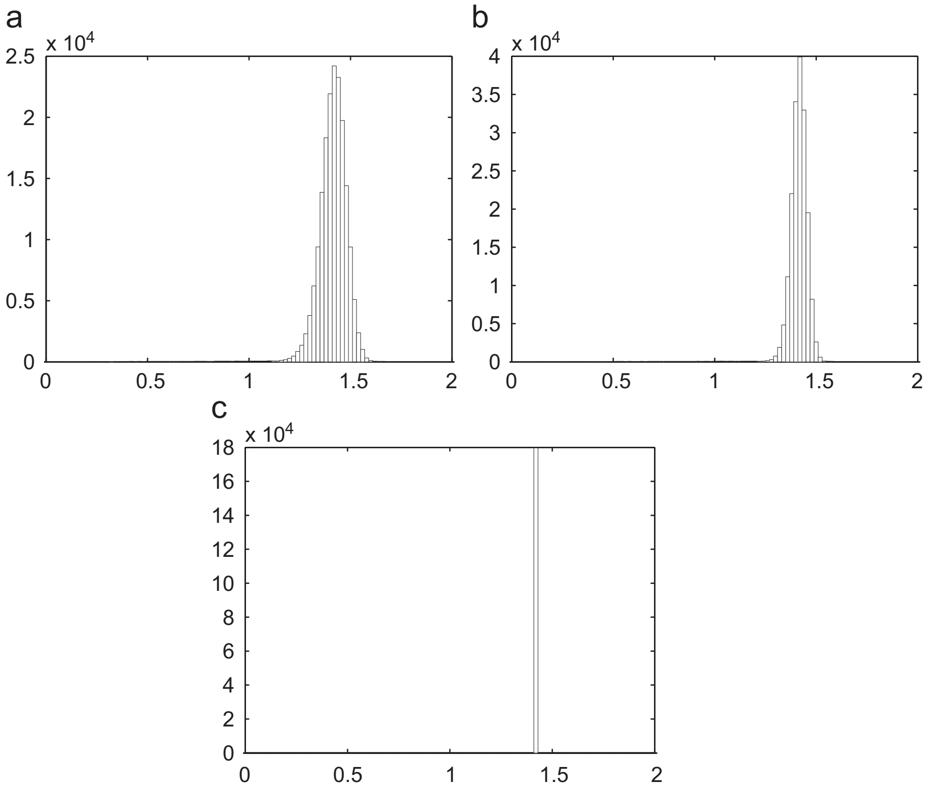

Fig. 4. The histogram of the pairwise distances between the 600 FERET facial images in varying dimensional whitened PCA space, when the whitened data vectors are normalized to *unit length*. (a) 100 dimensions. (b) 200 dimensions. (c) 599 dimensions.

distance between two points v_i and v_j can be computed as

$$\|v_i-v_j\|_2=\sqrt{(v_i-v_j)^T(v_i-v_j)}=\sqrt{2} \qquad \square \tag{11}$$

Surprisingly, our justification shows that the data distribution in the undersampled whitened space is determinated, regardless of the original data distribution in the observation space. In the $n-1$ dimensional whitened space, the pairwise distance between any two of the n samples is $\sqrt{2}$. Because all pairwise distances are the same, the whitened data lie at the vertices of an *regular simplex*. Fig. 2 illustrates an example where $n=3$.

2.3. Reduced-dimensional whitened PCA space

As the number of training samples increases, some high-dimensional data may become nearly dependent, and thus the trivial singular values often characterize the noisy components. Hence, it is a common practice for whitened PCA to retain a data dimension lower than $n-1$ by discarding the directions with too small variance [4]. In this situation, the pairwise distances between the whitened samples are not identical, but similar to each other. As evidence, we collect 600 FERET images of 200 persons (those used in [3,23,17]) to measure the characteristic of pairwise distances in the reduced-dimensional whitened space, and the results are showed in Fig. 3. We have reduce the dimensionality of a set 600 facial images (Gabor feature vectors) from 10,420 to 599, 200, 100, and 50 respectively, using SVD and measured the histograms of the pairwise distances of the whitened vectors. As shown in Fig. 3(a), when the dimension is 599, i.e. $n-1$, the histogram forms an impulsion at the distance $\sqrt{2}$. When the dimension is lower than 599, the distances still distribute in a narrow extent, suggesting the pairwise distances between the reduced-dimensional whitened samples are similar. To test the angles between whitened pattern vector, we normalize them to unit length and measure the histogram of pairwise distance again. Fig. 4 shows that their pairwise distances are all near $\sqrt{2}$, which suggests that all the whitened sample vectors are nearly perpendicular to each other. Our previous study have shown that the cosine similarity measure based NN classifier within the whitened space can achieve 100% accuracy on this standard data set [17], when the dimensionality of the whitened space is higher than 200.

Remark 2.2. In the high-dimensional whitened space, the sample vectors tend to be similarly distant, and nearly perpendicular to each other.

As the increase in the training data size, the dimensionality of the whitened PCA space must be increased in order to preserve enough information for accurate classification. High feature dimensionality would inevitably induce the efficiency and over-fitting problems in the classification stage. To address this limitation, a common way is to search for lower dimensional feature code with enhanced discriminatory power, i.e. feature extraction. Note that, although feature extraction itself has been intensively studied, feature extraction

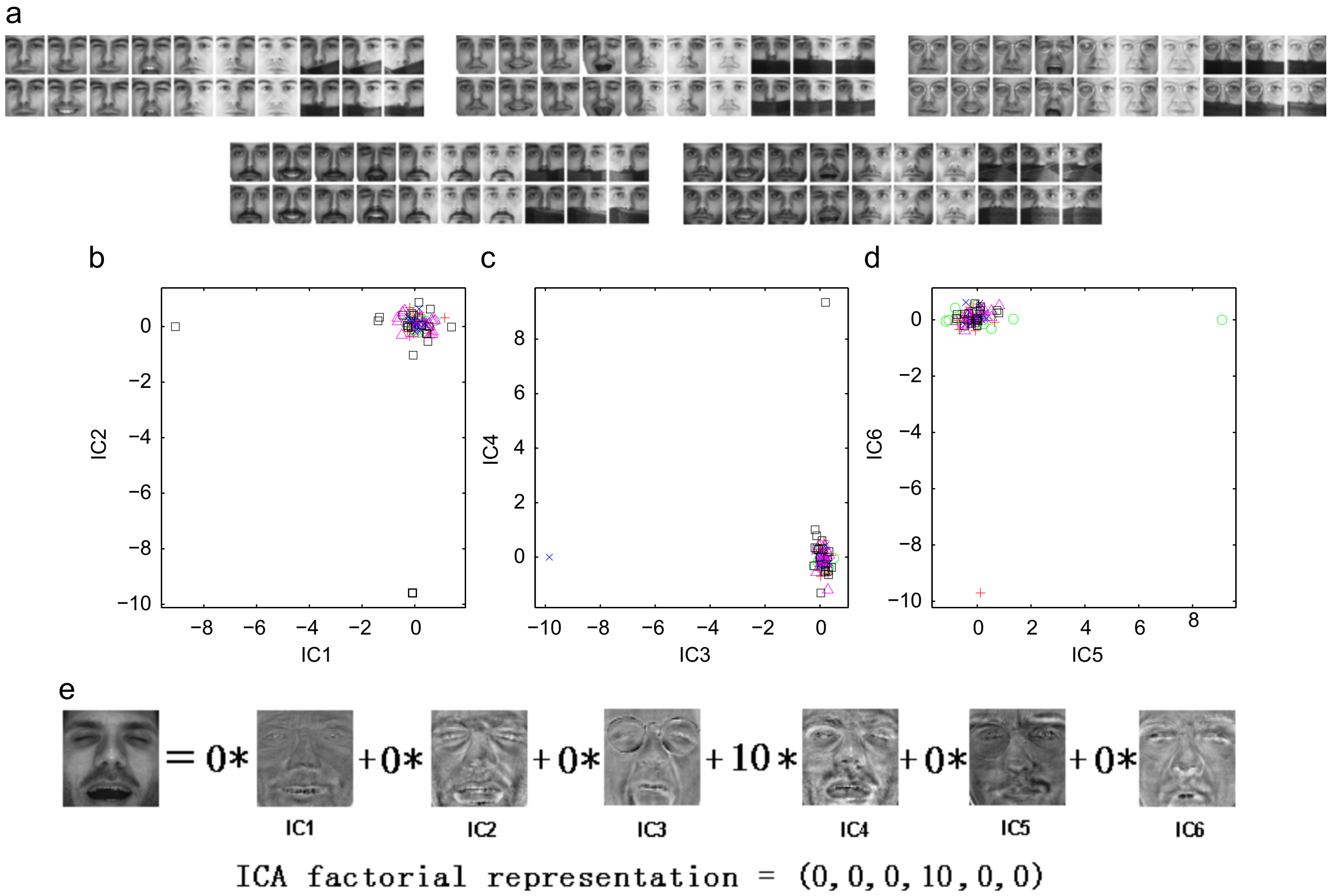

Fig. 5. 2D scatter plots of 100 face images of five individuals. (a) the original 100 images. (b,c,d) the scatter plots of the first six features sought by maximum non-Gaussianity criterion. (e) An image is represented by a linear combination of the six ICA projection bases shown in an image form. (a) 100 images of five persons. (b) 1,2. (c) 3,4. (d) 5,6. (e) projection basis.

within the high-dimensional whitened space, where the data points are similarly distant, is a rather new and interesting issue, which motivates us to conduct this study. In the rest of this paper, we will study such a special feature extraction problem by first pointing out the limitations of ICA, and then proposing a better method.

3. A critique on high-dimensional ICA

Why does ICA fail in the high-dimensional whitened space? In light of Remark 2.2, it is clear that the data distribution of the high-dimensional whitened space is totally different from that of the low-dimensional space. The ICA is effective for low-dimensional data classification because the non-Gaussianity of the low-dimensional data often relates to the multimodal structures. In contrast, in the high-dimensional whitened space, the relative separation of points vanishes as the pairwise distances concentrate. Since the data do not display any multimodal structure at all, the non-Gaussianity, as well as the ICA model, breaks down inevitably.

The measures of non-Gaussianity, such as the kurtosis and negentropy, depend primarily on the high-order moment of the data, and thus emphasize the tails of the distribution. When the feature space is very sparse, the tails of distribution are corresponding to the isolated samples that are projected far from the origin, because a very small number of isolated points induces very large high-order statistics. Therefore, non-Gaussianity criterion favors the directions on which a very small number, even single, data points is isolated far from the origin, and simultaneously the others are near the origin. The total data variance is constrained to 1 in the whitened space, and the data variance on this direction is mainly caused by the isolated point.

To provide a convincible evidence, we collect 100 images of five persons from the AR database with different facial expression, lighting conditions, and occlusions, as shown in Fig. 5(a). The PCA is first applied to reduce the dimension from 16,384 (128×128 pixels) to 50, followed by the whitening process. Then, we seek the maximum non-Gaussianity directions using the FastICA [24] with kurtosis based non-Gaussianity measure. As shown in 5(b,c,d), all the first six ICA projections collapse 99 images together, leaving one single image far apart. Numerically, on each projection direction, the distance between the isolated image and the origin is about 10, which suggests that the variance of the 100 images, i.e. $var(y) \approx 10^2/100 = 1$, is mainly caused by the isolated image. At the same time, the high-order statistics of the projected data is maximized by a single isolated image, e.g. $kurt(y) \approx 10^4/100 = 100$. It is clear that such a low-dimensional ICA representation is not appropriate for subsequent classification.

Interestingly, this limitation of ICA can also be visually perceived by its projection basis images. As shown in Fig. 5(e), the ICA bases tend to describe the characteristic of a specific facial image, rather than that of a specific face (class). Indeed, such ICA projection tends to generate the *sparse factorial representation*, where most extracted features (projected coordinates) are zeros. However, the features extracted by such bases are not appropriate for recognition, because they are only effective to discriminate a certain (training) image from the others. Actually, similar ICA basis images have already been demonstrated in many other studies such as [2,13], but they have not been aware of the ICA limitation discussed above.

It is well known that the higher-order statistics can be applied in a more robust way by using the nonlinear hyperbolic tangent function $\tanh(y)$, whose values always lie in the interval $(-1,1)$, or some other nonlinearity that grows slower than linearly with its argument value, such as $y\exp(-y^2/2)$ [25]. Unfortunately, we found that in high-dimensional space these contrast functions derive similarly overfitting results as the kurtosis in most cases. Some exceptions are shown in Fig. 6, which are still not appropriate for classification. The rationale behind this limitation is simple. *Roughly speaking, every point seems to be a outlier when the data points are similarly distant.* Although some robust estimators address the situation where a small proportion of data are outliers, they cannot handle the high-dimensional whitened space where every data point seems to be the outlier.

4. Alternative approaches: random pursuit and locality pursuit

Previous section has shown that ICA fails in the high-dimensional applications where the whitened data points are similarly distant. This section introduces two alterative approaches, which is fundamentally different from ICA, to pursue low-dimensional discriminative projections. The alternative approaches are based on the assumption as follows.

Assumption 1. In the high-dimensional whitened space, the relatively close data points tend to belong to the same class.

4.1. Random pursuit: random projections in whitened space

Random projection (RP) [26,27] refers to the technique of projecting a set of points from a high-dimensional space to a randomly chosen low-dimensional subspace. Random projections have also played a central role in providing feasible solutions to the well-known Johnson–Lindenstrauss (JL) lemma [26], which states that a point set in a high-dimensional Euclidean space can

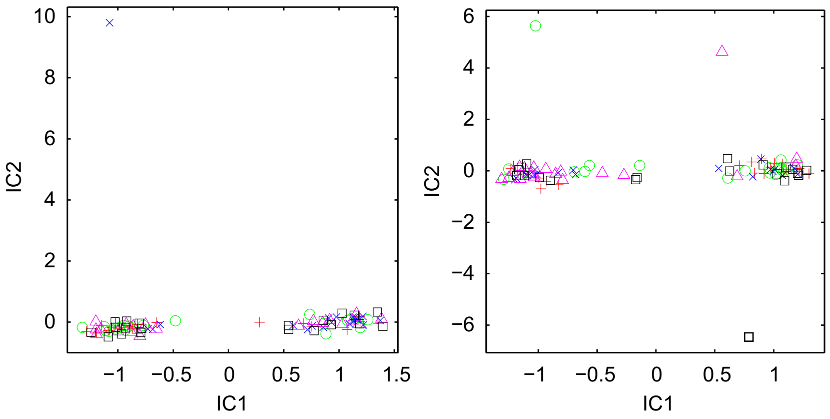

Fig. 6. Some results derived by the robust estimators of ICA. (a) $G(y) = \tanh(y)$ (b) $G(u) = y\exp(-y^2/2)$.

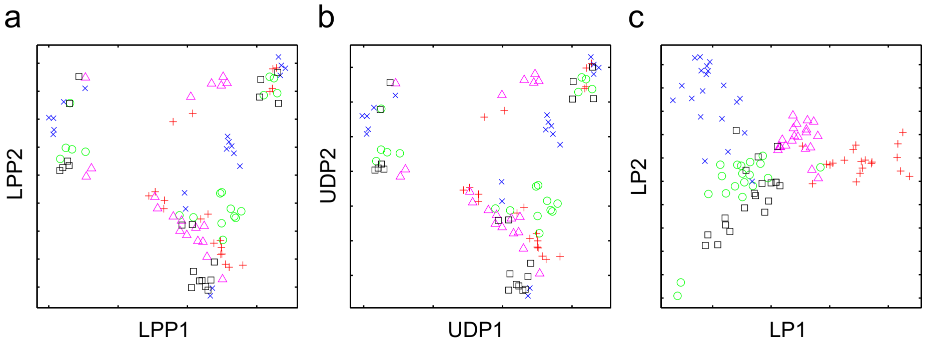

Fig. 7. The dramatical difference between manifold-based methods (LPP/UDP) and LP on face visualization: The scatter plots of 100 AR images of five persons (classes) derived by LPP, UDP, and LP, respectively, with different symbols representing the images of different persons. (a) LPP. (b) UDP. (c) LP.

be mapped down onto a space of dimension logarithmic in the number of points, such that the distances between the points are approximately preserved. If Assumption 1 holds in the whitened space, it is possible that the recognition performance in a randomly chosen low-dimensional subspace may be similar to that in the high-dimensional space.

In this study, we apply random projections as a standard baseline to evaluate the "feasibility" of ICA projections. Certainly, one could compare ICA with other state-of-the-art feature extraction algorithms. However, even if ICA performs worse with a certain number of features, one may still argue that ICA may be useful in the sense that it can be improved by some normalization or feature selection procedures [5,28,14]. To eliminate this possibility, a good baseline may be the random projections. Obviously, if ICA can outperform the random projections, it could be considered useful; otherwise, one can infer that ICA is ineffective, since it just generates some misleading (worse than random) projection directions for the classification problems.

In our study, the entries of the random projections matrix $R_p \in \mathbb{R}^{p\times p}$ are computed as follows[2]: (1) each entry of the matrix is an i.d.d. $N(0,1)$ value; (2) Orthogonalize the p columns of the matrix using Gram–Schmidt algorithm; (3) Normalize the columns of the matrix to unit length.

4.2. *Locality pursuit: Locality preserving projections in whitened space*

Based on Assumption 1, a desirable projection criterion might be formulated by pairwise distances among the neighboring points, rather than the global statistics of all points. By seeking the projection direction that concentrates the neighboring points together, the similarly distant points may become well clustered on the projected line. Let $N_k(\tilde{x}_i)$ be the set of the k nearest neighbors of $\tilde{x}_i$. The affinity matrix of the neighborhood graph is commonly defined as

$$A_{ij} = \begin{cases} k(\tilde{x}_i,\tilde{x}_j) & \text{if } \tilde{x}_i \in N_k(\tilde{x}_j) \text{ or } \tilde{x}_j \in N_k(\tilde{x}_i) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $k(\tilde{x}_i,\tilde{x}_j)$ is the function that measures the affinity between $\tilde{x}_i$ and $\tilde{x}_j$, such as the binary function and heat kernel (radial basis) function. Cosine similarity measure is applied to search for the nearest neighbors as suggested in [29,17]. The projection of mapping the neighboring points together is well defined by the *locality preserving criterion* [19] as follows:

$$w_{opt} = \arg\min_{w} w^T \tilde{X} L \tilde{X}^T w, \quad \text{s.t. } w^T w = 1 \tag{13}$$

where $\tilde{X}$ is the whitened data matrix and L is the Laplacian matrix of the neighborhood graph. Note that the norm constrain $w^T w = 1$ is equivalent to the global scatter constrain $w^T \tilde{X}^T \tilde{X} w = 1$ in the whitened space. The optimization problem can be solved by computing the eigenvectors of matrix $\tilde{X} L \tilde{X}^T$ corresponding to the smallest eigenvalues.

As the manifold structure of data is largely demolished during the whitening process, the low-dimensional embedding by locality pursuit is fundamentally different from those of conventional manifold analysis methods, such as LPP [19] and UDP [30]. As evidence, Fig. 7 demonstrates several two-dimensional manifold structures of the 100 facial images of five subjects. As shown in (a) and (b), LPP/UDP indeed discovers the local structures of the face space, but these structures are mostly corresponding to the distinct groups of expression, lighting, and wearing, rather than distinct face identity. In contrast, measured in the whitened PCA space, multi-modality structure in (c) recovered by the LP algorithm corresponds to the face identity directly. Clearly, the low-dimensional space derived by LP would be more appropriate for the classification purpose.

Moreover, locality preserving criterion provides a natural rank of the discrimination power of the extracted features [31]. Applying the locality preserving criterion (13) to the whitened feature vectors of the 100 AR images, we get the six optimal projections that are shown in Fig. 8. As expected, the features extracted by locality concentrating criterion is much more separable than those by ICA (shown in Fig. 5(b)–(d)). In Fig. 8, as the value of $w\tilde{X}L\tilde{X}^T w$ increasing from "LP1" to "LP6", and the five faces (classes) become less and less compact from (a) to (c). In light of this ranking, a small proportion of features could be selected for the efficient recognition purposes.

In the high-dimensional whitened space, the L_2 distance metric between any two points would become very similar. Hence, the commonly used "heat kernel" weights might become nearly identical for all neighboring points, which makes its performance similar to that of the binary weights. Aggarwal et al. have justified that L_f distance metric with $f \in (0,2)$ provide larger distance contrast than the L_2 distance metric [32]. Inspired by this property, we generalize the L_2 norm of the heat kernel to L_f norm where f is any positive number, i.e.

$$k(x_i,x_j) = \exp[-L_f(x_i,x_j)/\sigma] \tag{14}$$

where $L_f(x_i,x_j) = \sum_{d=1}^{p} |x_i^{(d)} - x_j^{(d)}|^f$, and σ is the scaling factor. Obviously, the commonly used heat kernel is a special case of the generalized heat kernel with $f=2$.

5. Experiments

In this section, we evaluate the feasibility of ICA (FastICA) algorithm and the effectiveness of the locality pursuit algorithm on the high-dimensional classification problems, using the

[2] In our implementation, the matrix R_p is generated simply by a single Matlab script Rp=orth(rand(p));

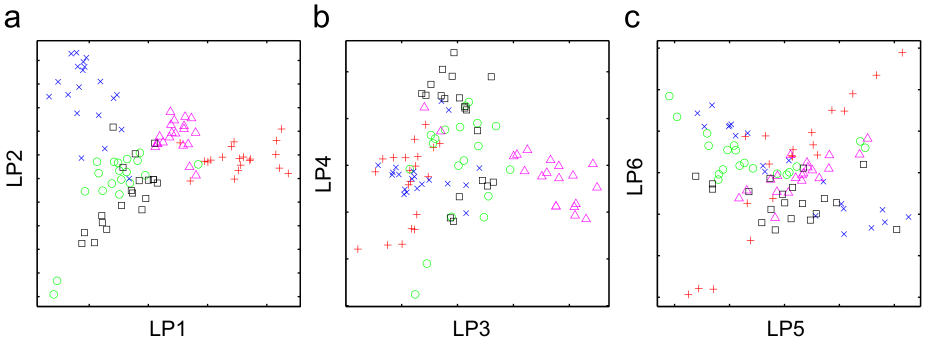

Fig. 8. 2D scatter plots of 100 face images of five individuals. (a,b,c) the scatter plots of the first six features sought by locality pursuit. In the scatter plots, different symbols representing the images of different persons.

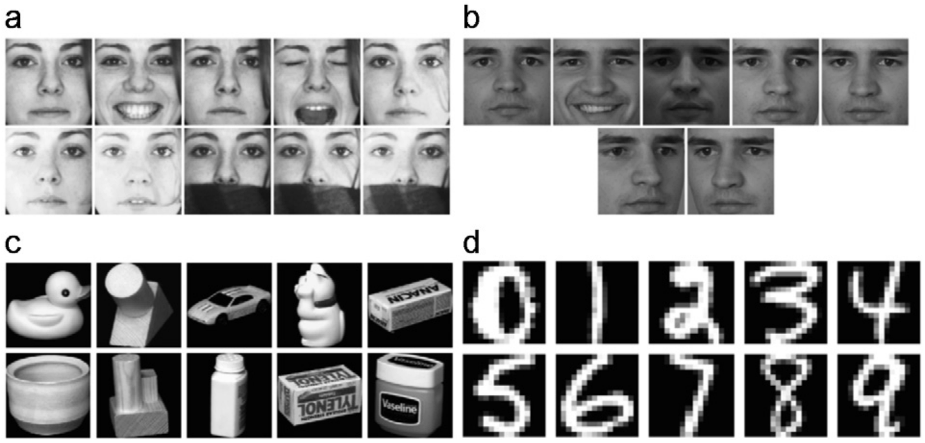

Fig. 9. Example images used in our experiments. (a) The ten images of one person marked "a"–"j" in the AR database. We discard the images with sun glasses since they are difficult to accurately align (normalize) even by manual labelling. (b) The seven images of a person marked by "ba," "bj,", "bk,", "be", "bf", "bd", "bg" in the FERET database. (c) COIL-20 data set. (d) USPS data set.

random pursuit as the baseline. We also perform a comparative study on state-of-the-art feature extraction algorithms for the face recognition task.

5.1. Comparison of ICA, random pursuit, and locality pursuit

We compare ICA with RP and LP on the face, object, and handwriting recognition tasks. For face recognition, we employ the AR and FERET databases. The AR data set contains 100 persons and each person has 10 frontal face images taken in different facial expressions, illuminations, and occlusion. The FERET data set includes 1400 images of 200 individuals with seven images per person with variations in facial expression, illumination and pose. The 10,240 dimensional Gabor feature is used for representation, using the extraction procedure in [18]. For object recognition, we use the Columbia Object Image Library (COIL-20) database, which contains 1440 gray-scale images of 20 objects. We randomly choose 20 images of each object for training, and remaining 52 images for testing. For the digit recognition, we use the USPS data set which contains handwritten digit images of a resolution 16 × 16. We randomly select 20 and 50 images for training and testing, respectively. The gray value of the pixel is directly used as features, generating a feature vector with the dimensionality of 16,384 and 256 for object and digit, respectively. Fig. 9 shows the example images of the four data sets, and Table 1 summarizes the experimental settings. For all data sets, we conduct the hand out procedure 10 times, and the performance is compared in terms of the average recognition rate.

Table 1
Best recognition accuracy using different unsupervised feature extraction methods and corresponding feature dimensions. The number in the bracket indicates the optimal feature dimension.

Dataset	#Class	#Train/Test	Feature	Dim.	PCA Dim (p)
AR	100	5/5	Gabor	10 240	200
FERET	200	4/3	Gabor	10 240	200
COIL-20	20	20/52	Pixel	16 384	100
USPS	10	20/50	Pixel	256	50

For all data sets, PCA is first applied to reduce the feature dimension to p. ICA and LP are then fairly compared in the same PCA subspace. ICA, implemented by the FastICA algorithm, has been evaluated using three contrast functions: $G_1(u)=u^3$, $G_2(u)=\tanh(u)$, and $G_3(u)=u\exp(-u^2/2)$. The results of different contrast functions are almost identical (typically smaller than 1%), and we reported the best result achieved. For the LP methods, we empirically set $k=1$ for the face data sets, and $k=3$ for the object and digit data sets. The "goodness" of different feature spaces are evaluated in terms of the classification accuracy of the nearest neighbor classifier, using Euclidean distance measure (Ed) and cosine similarity measure (Cos), respectively. Fig. 10 shows their comparative recognition accuracy at the dimension of 5,10,15,20, and p. The results reveal a number of interesting observations:

1. The accuracies of ICA, RP and LP are identical when the feature dimension is p. This is because ICA, RP, and LP pursue orthogonal basis in the p dimensional whitened PCA space.

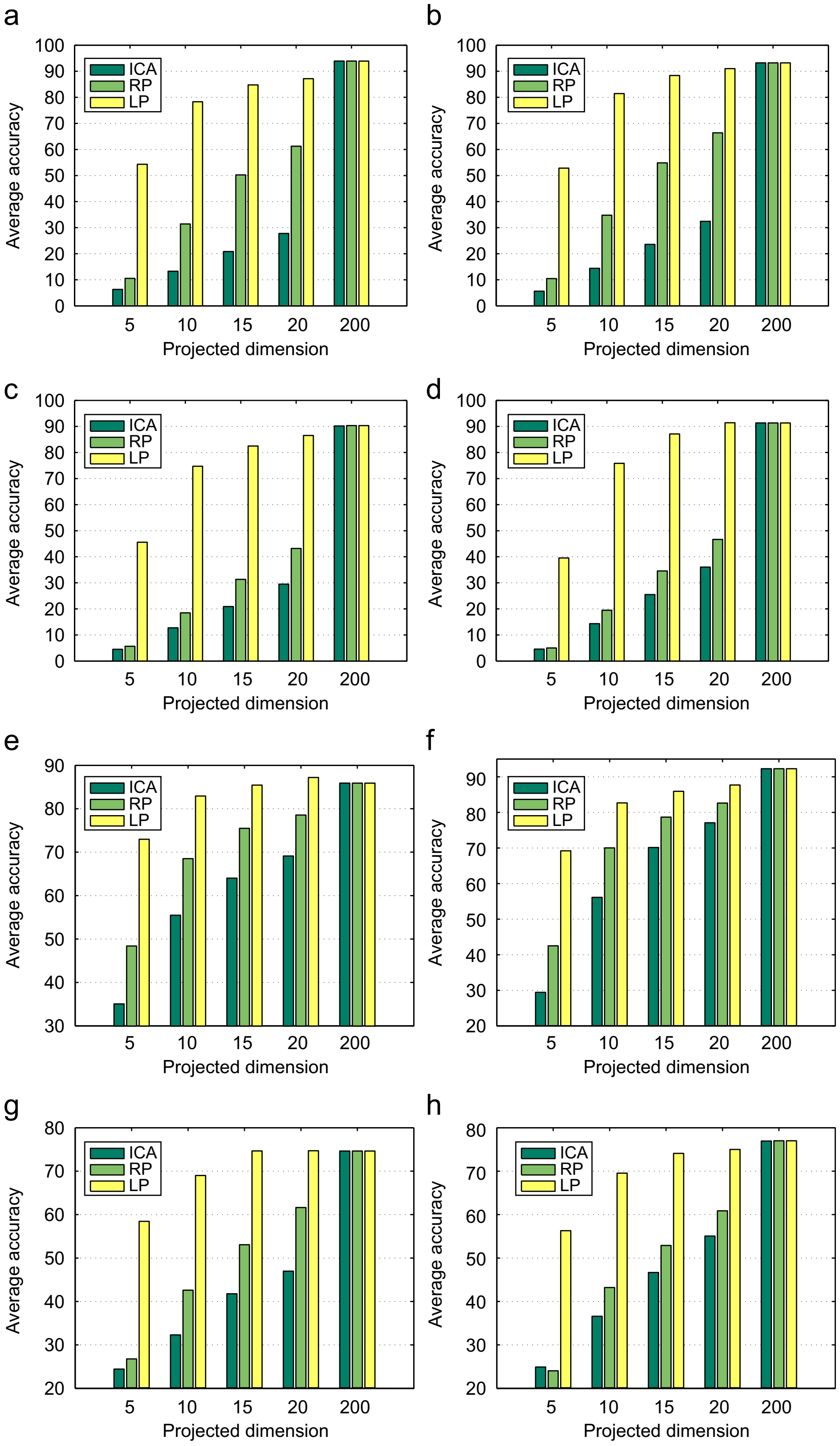

Fig. 10. Comparative performance of ICA, Random Projection (RP) and Locality Projections (LP) using different number of projection directions. The classification is performed by nearest-neighbor classifier using Euclidean (Ed) and cosine similarity measure (Cos) respectively. The average accuracy from 10 runs are plotted a function of the projection dimension. (a) AR(Ed). (b) AR(Cos). (c) FERET(Ed). (d) FERET(Cos). (e) COIL(Ed). (f) COIL(Cos). (g) USPS(Ed). (h) USPS(Cos).

When all p bases are used, all the three methods perform only a axes rotation of the whitened PCA space, and thus do not change the recognition performance at all [14].

2. If the reduced number of bases are used, the accuracies of the three methods differ to a large extent: LP performs much better than the RP baseline, followed by ICA. Based on the

Table 2
Best recognition accuracy of ICA and LP methods and corresponding feature dimensions. The number in the bracket indicates the optimal feature dimension.

Data	Euclidean distance measure		Cosine similarity measure	
	$\text{ICA}_{(opt)}$	$\text{LP}_{(opt)}$	$\text{ICA}_{(opt)}$	$\text{LP}_{(opt)}$
AR	$93.94 \pm 0.74_{(190)}$	$94.02 \pm 0.97_{(105)}$	$93.36 \pm 0.90_{(185)}$	$95.24 \pm 0.63_{(70)}$
FERET	$90.18 \pm 0.93_{(200)}$	$91.90 \pm 0.97_{(95)}$	$92.30 \pm 0.51_{(150)}$	$96.57 \pm 0.46_{(65)}$
COIL	$85.89 \pm 1.41_{(100)}$	$90.26 \pm 1.25_{(65)}$	$92.31 \pm 1.21_{(95)}$	$93.25 \pm 1.21_{(40)}$
USPS	$74.63 \pm 3.37_{(40)}$	$77.90 \pm 2.67_{(25)}$	$76.93 \pm 3.41_{(50)}$	$78.17 \pm 3.40_{(35)}$

result that ICA is significantly worse than the random projection method on all data sets, we can infer that the non-Gaussianity measure of ICA might be a misleading criterion on the low-dimensional representation for classification.

3. At the low feature dimensions such as 20, the LP method outperforms ICA methods by a margin of 10%–60%. The large performance margins clearly suggest the superiority of the locality measure over the non-Gaussianity measure for the classification purpose. In particular, the 20 dimensional LP features achieve similar performance to the full p dimensional whitened features. This indicates that the LP method can preserve the class separability while largely reducing the feature dimension.

Besides the recognition performance using low-dimensional features, we also compare ICA and LP in terms of the best accuracy achieved. Table 2 summarizes the best accuracy achieved by ICA and LP and the corresponding optimal feature dimension searched from 5 to p with an interval of 5. One can see from the Table that the LP method not only achieves better accuracy but also uses lower feature dimension than ICA on all the four data sets. Moreover, the training speed of LP algorithm is faster than ICA by over one magnitude in all experiments, even though we have used the FastICA implementation, which is regarded as the fastest ICA algorithm. Therefore, in terms of both recognition accuracy and computational efficiency, our experimental results clearly show the LP algorithm is a better choice for feature extraction.

5.2. Effectiveness of the generalized heat kernel

Previous experiments have verified the effectiveness of the LP algorithm using binary weights for the simplicity of parameter settings. This set of experiments further evaluates the LP algorithm with the proposed generalized heat kernel. For the comparison purpose, we use the heat kernel weights as the baseline, which is a special case of the generalized heat kernel with $f=2$. Since cosine similarity measure based NN classifier is better than Euclidean distance on all data sets, we therefore report the results based on the cosine similarity measure based NN classifier. Fig. 11 shows the recognition performance of LP method with heat kernel weights and generalized heat kernel weights. For fair comparison, both heat kernel weights and generalized heat kernel weights are with $\sigma = 1$. The generalized heat kernel is tested using $f=1$, $f=0.9$, and $f=0.8$. One can see from the figure that (1) Comparing to the results of binary weights in Fig. 10, the effect of the heat kernel weights is not significant for the LP algorithm. This may because the data points are similarly distant, and the heat kernel weights are almost equivalent to the binary weights. (2) The generalized heat kernel weights perform better than the heat kernel in all the four recognition tasks, which clearly suggests the L_f distance metric with $f \in (0,2)$ is a more effective metric to measure the proximity than the L_2 distance metric.

The advantage of generalized heat kernel is most significant in the object recognition (COIL-20) task, where the recognition accuracy is improved by 6%–15%. The highest accuracy of binary weight based LP is 93.25% using 40 features (See Table 2). In contrast, the accuracy reaches to 96% using only 15 features when the neighborhood graph is weighted by the generalized heat kernel with $f=0.8$. This comparative result indicates that LP algorithm with generalized heat kernel can derive a more compact and efficient feature space than those of binary weights and heat kernel weights.

5.3. Further experiments on face recognition

In face recognition community, the locality measure based feature extraction methods such as Laplacianfaces (LPP) and UDP have reported excellent performance [19,30]. It is interesting to compare them with locality pursuit. For the fair comparison and the simplicity of parameter selection, the neighborhood graphs of all three methods are weighted by 0/1 value, and the number of neighbors is set to $\{1,2,\ldots,l-1\}$, where l is the number of training samples per class. All the three methods perform better with the cosine similarity measure than the Euclidean distance. We therefore compare their performance using the cosine similarity measure based nearest neighbor classifier.

Fig. 12 summarizes the comparative performance of the LPP, UDP and LP algorithms using different local neighborhood size k. The recognition rates of LPP and UDP are nearly identical at any projected dimension, which is consistent with our previous comment paper [33]. On both data sets, LP significantly outperforms LPP/UDP, especially when the feature dimension is low. Moreover, the performance of LP is not largely affected by the locality parameter, which indicates the local structure of the whitened space is stable across different numbers of preserved neighbors. In contrast, LPP/UDP perform unstably across different neighborhood size. Through integrating the high-dimensional whitening process and locality measure, the LP algorithm significantly improves the accuracy and stability of the manifold based methods.

Finally, we perform a comparative study on 10 state-of-the-art feature extraction algorithms, including

- **Fisherfaces** [34]
- **Gabor-Fisher Classifier** [23]
- **Null-Space LDA** [35]
- **Direct LDA** [36]
- **LDA/GSVD** [37]
- **ICA-II** [2]
- **Laplacianfaces** [19]
- **Unsupervised Discriminant Projections** [30]
- **Whitened Cosine Similarity Measure** [29]
- **Locality Pursuits**

The first five are supervised feature extraction methods that aims to generalize LDA in the small sample size problem, and the remaining four are unsupervised methods that are effective for face recognition. For all the two-stage methods except Fisherfaces, the intermediate PCA dimension is set to 200 for the fair

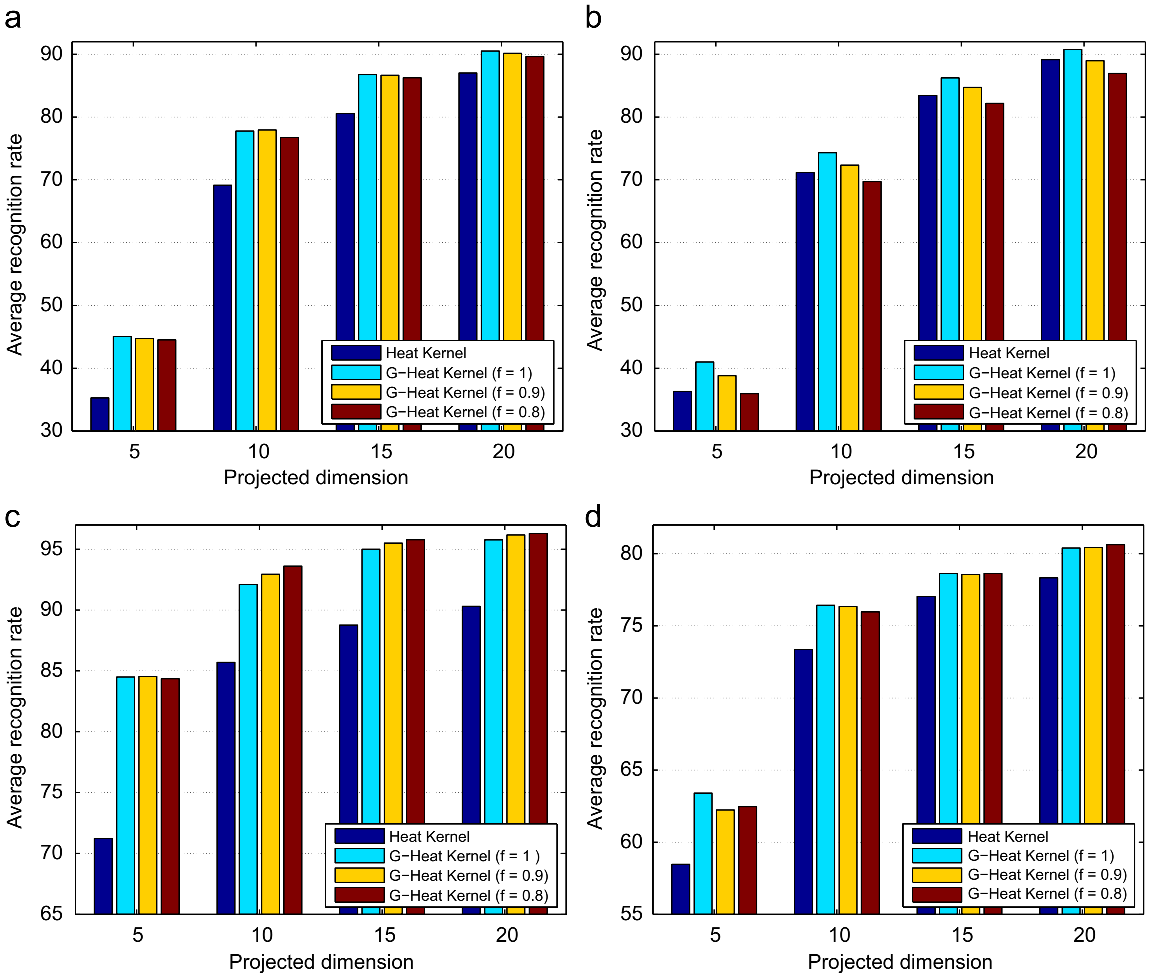

Fig. 11. The recognition performance of LP method with heat kernel weights and generalized heat kernel weights. The generalized heat kernel is tested using $f=1$, $f=0.9$, and $f=0.8$. (a) AR data set. (b) FERET data set. (c) COIL data set. (d) USPS data set.

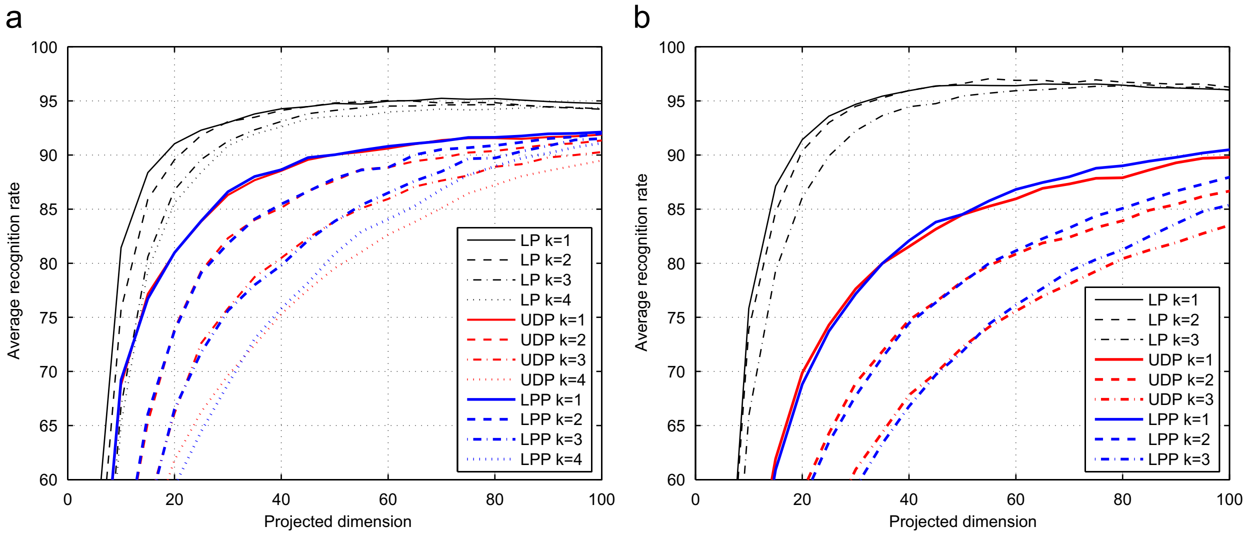

Fig. 12. Comparative face recognition performance of LPP, UDP and LP algorithm with different local neighborhood size k. (a) AR data set. (b) FERET data set.

comparison. The nearest neighbor classification is applied to evaluate the effectiveness of the low-dimensional feature space derived by the algorithms, and all the 10 methods perform better with the cosine similarity measure than the Euclidean distance. We therefore compare their performance using the cosine similarity measure based nearest neighbor classifier. Average recognition rate of 10 random training/test partitions is reported.

Fig. 13 shows the comparative face recognition performance, and one can see from the figure that generalized LDA algorithms perform better than the unsupervised ones in general. When the

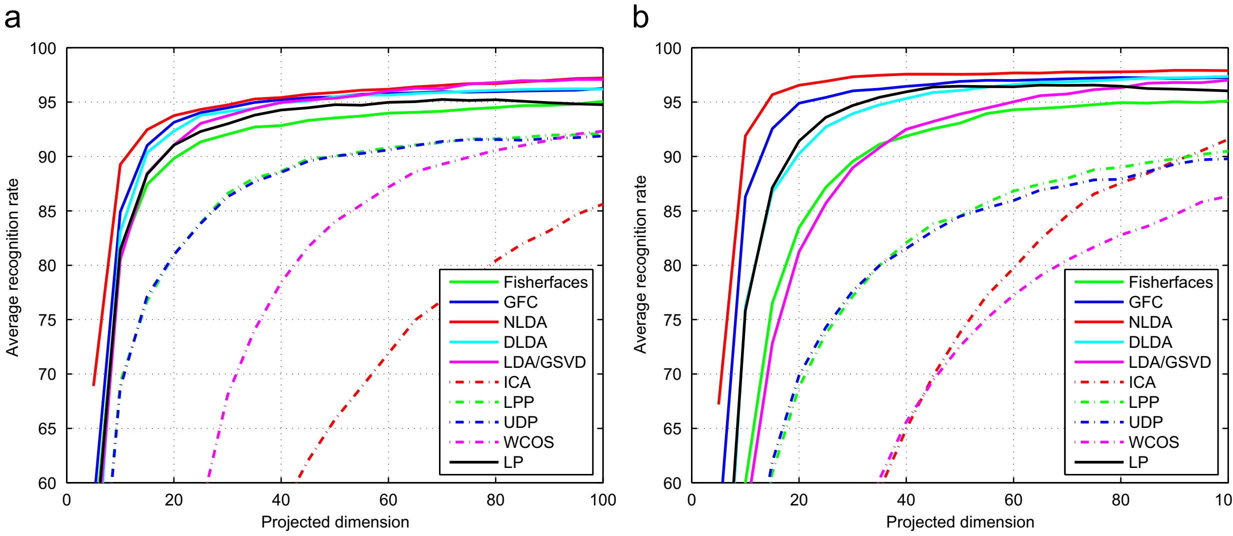

Fig. 13. Comparative face recognition performance of 10 state-of-the-art feature extraction methods. The average recognition rate are plotted a function of the projection dimension. (a) AR data set. (b) FERET data set.

feature dimension is low, Null-space LDA yields significantly better accuracy than other competing methods. As the dimension increasing, the performance of all supervised methods become similar. This phenomenon may be because that the face classes resides in some linear class-specific structures, and well clustered feature space can be derived by many approaches given a sufficiently high dimension. For supervised feature extraction, Laplacianfaces and UDP perform better than whitened PCA followed by ICA. The accuracy of whitened cosine similarity measure reaches to 85%–90% when the dimension is 100, suggesting that Assumption 1 is roughly fulfilled in the whitened PCA space. Locality pursuit outperforms all other unsupervised methods, and the accuracy gain is huge when the dimension is low.

Surprisingly, locality pursuit obtains competitive performance with the generalized LDA algorithms. This high performance can be directly explained by the subsequent combination of whitened PCA and locality measure, both which are known to be effective for face recognition. Furthermore, based on the link to Vapnik's statistical learning theory, the rationale behind locality pursuit could be much deeper. In lighting of Theorem 2.1, the optimality of whitened PCA space can be interpreted by Vapnik's statistical learning theory [38] (page 353), which proofs that the minimum margin over all dichotomies of $k \leq n$ points contained in a sphere in $\mathbb{R}^{n-1}$ can be maximized by placing these points on a regular simplex whose vertices lie on the surface of the sphere. Considering all dichotomies of one class (with k samples) and the rest classes, the full-dimensional whitened PCA space is an optimal feature space that maximizes the minimum one-against-the-rest margin between the classes. The reduced-dimensional whitened space is approximately optimal in light of Remark 1. At the same time, since sample vectors are nearly perpendicular to each other in the whitened space, the directions of intraclass differences and interclass margin tends to be uncorrelated. The projection directions that congregate the neighboring samples of the same class would simultaneously preserve the margins between classes. In this perspective, if Assumption 1 holds, locality pursuit is an unsupervised margin-preserving feature extraction method.

6. Conclusions

This paper studies the unsupervised feature extraction problem in high-dimensional whitened space, where the sample vectors display a special distributional phenomenon: they tend to be similarly distant, and perpendicular to each other. This newly found phenomenon on one hand benefits the one-sample pattern recognition problem, but on the other hand makes the subsequent feature extraction difficult. For instance, with this distributional property, the widely used ICA methods tend to extract the independent features simply by the projections that isolate single or very few samples apart and congregate all other samples around the origin, without any concern on the clustering structure. Our study further shows that ICA produces misleading features, whose generalization ability is almost always worse than those derived by random projections. To address this limitation, we suggest to apply the locality-based measures to pursues low-dimensional features in the whitened space. Experimental results show that the proposed "locality projections" method outperforms ICA by a large margin in various pattern recognition tasks. Moreover, the performance of locality projections can be further improved by a novel generalized heat kernel that effectively characterizes the neighborhood graph of the high-dimensional data points. Experimental results on face recognition show that the proposed LP method is significantly better than other locality measure based methods such as Laplacianfaces and UDP, and is even comparable to the supervised feature extraction methods.

Acknowledgments

The authors would like to thank the anonymous reviewers for their thoughtful and constructive remarks that are helpful to improve the quality of this paper. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61002051, No. 61005025, and No. 61073072, Major Project of National Science and Technology under Grant No. 2011ZX03002-005-01, and the Fundamental Research Funds for the Central Universities.

References

[1] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley-Interscience, 2000.
[2] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Transactions on Neural Networks 13 (6) (2002) 1450–1460.

[3] C. Liu, H. Wechsler, Independent component analysis of gabor features for face recognition, IEEE Transactions on Neural Networks 14 (4) (2003) 919–992.

[4] C. Liu, Enhanced independent component analysis and its application to content based face image retrieval, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 34 (2) (2004) 1117–1127.

[5] K.-C. Kwak, W. Pedrycz, Face recognition using an enhanced independent component analysis approach, IEEE Transactions on Neural Networks 18 (2) (2007) 530–540.

[6] A. Jain, J. Huang, Integrating independent components and linear discriminant analysis for gender classification, in: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society, 2004, pp. 159–163.

[7] B. Moghaddam, Principal manifolds and probabilistic subspaces for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6) (2002) 780–788.

[8] K. Baek, B. Draper, J. Beveridge, K. She, Pca vs. ica: a comparison on the feret data set, in: Joint Conference on Information Sciences, Citeseer, Durham, NC, 2002, pp. 824–827.

[9] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Computation 7 (6) (1995) 1129–1159.

[10] P. Comon, Independent component analysis: a new concept? Signal Processing 36 (3) (1994) 287–314.

[11] B. Moghaddam, Principal manifolds and probabilistic subspaces for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6) (2002) 780–788.

[12] B.A. Draper, K. Baek, M.S. Bartlett, J.R. Beveridge, Recognizing faces with pca and ica, Computer Vision and Image Understanding 91 (1) (2003) 115–137.

[13] J. Yang, D. Zhang, J. yu Yang, Is ica significantly better than pca for face recognition? In: International Conference on Computer Vision, 2005.

[14] M.A. Vicente, P.O. Hoyer, A. Hyvarinen, Equivalence of some common linear feature extraction techniques for appearance-based object recognition tasks, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 896–900.

[15] P. J. Phillips, P. J. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[16] P. J. Phillips, P. J. Flynn, T. Scruggs, K. Bowyer, W. Worek, Preliminary face recognition grand challenge results, in: Proceedings of the Seventh International Conference on Automatic Face and Gesture Recognition, 2006, pp. 15–24.

[17] W. Deng, J. Hu, J. Guo, Gabor-eigen-whiten-cosine: A robust scheme for face recognition, Lecture Notes in Computer Science, vol. 3723, 2005.

[18] W. Deng, J. Hu, J. Guo, W. Cai, D. Feng, Robust, accurate and efficient face recognition from a single training image: an uniform pursuit approach, Pattern Recognition 43 (5) (2009) 1748–1762.

[19] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.

[20] J. Ye, T. Xiong, Computational and theoretical analysis of null space and orthogonal linear discriminant analysis, The Journal of Machine Learning Research 7 (2006) 1204.

[21] G. Golub, C. Van Loan, Matrix Computations, Johns Hopkins University Press, 1996.

[22] G. Strang, Introduction to Linear Algebra, Wellesley Cambridge Press, 2003.

[23] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Transactions on Image Processing 11 (4) (2002) 467–476.

[24] A. Hyvarinen, Fast and robust fixed-point algorithms for independent component analysis, IEEE Transactions on Neural Networks 10 (3) (1999) 626–634.

[25] A. Hyvarinen, J. Karhunen, E. Oja, Indenpendent Component Analysis, John Wiley & Sons, 2001.

[26] K. Das, Z. Nenadic, An efficient discriminant-based solution for small sample size problem, Pattern Recognition 42 (5) (2009) 857–866.

[27] M. Davenport, P. Boufounos, M. Wakin, R. Baraniuk, Signal processing with compressive measurements, IEEE Journal of Selected Topics in Signal Processing 4 (2) (2010) 445–460.

[28] M. Bressan, J. Vitri?, On the selection and classification of independent features, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1312–1317.

[29] C. Liu, The bayes decision rule induced similarity measures, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1086–1090.

[30] J. Yang, D. Zhang, J.-Y. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4) (2007) 650–664.

[31] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, Advances in Neural Information Processing Systems 18 (2006) 507.

[32] C. Aggarwal, A. Hinneburg, D. Keim, On the surprising behavior of distance metrics in high dimensional space, Database TheoryICDT 2001, 2001, pp. 420–434.

[33] W. Deng, J. Hu, J. Guo, H. Zhang, C. Zhang, Comments on globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (8) (2008) 1503–1504.

[34] P.N. Belhumeour, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.

[35] R. Huang, Q. Liu, H. Lu, S. Ma, Solving the small sample size problem of lda, in: Proceedings of the 16th International Conference on Pattern Recognition, vol. 3. IEEE, 2002, pp. 29–32.

[36] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data–with application to face recognition, Pattern Recognition 34 (10) (2001) 2067–2070.

[37] P. Howland, J. Wang, H. Park, Solving the small sample size problem in face recognition using generalized discriminant analysis, Pattern Recognition 39 (2) (2006) 277–287.

[38] V. Vapnik, Statistical learning theory, 1998.

Weihong Deng received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2009, respectively. From October 2007 to December 2008, he was a postgraduate exchange student in the School of Information Technologies, University of Sydney, Australia, under the support of the China Scholarship Council. He is currently an associate professor in School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision, with a particular emphasis in face recognition.

Yebin Liu received the BE degree from Beijing University of Posts and Telecommunications, P.R. China, in 2002, and the PhD degree from the Automation Department, Tsinghua University, Beijing, P.R. China, in 2009. He is currently an assistant professor in the Automation Department, Tsinghua University, Beijing, People's Republic of China. His research interests include light field, image-based modeling, and rendering multicamera array techniques.

Jiani Hu received the B.E. degree in telecommunication engineering from China University of Geosciences in 2003, and the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2008. She is currently an associate professor in School of Information and Telecommunications Engineering, BUPT. Her research interests include information retrieval, statistical pattern recognition and computer vision.

Jun Guo received B.E. and M.E. degrees from BUPT, China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and the dean of School of Information and Communication Engineering, BUPT. His research interests include pattern recognition theory and application, information retrieval, content-based information security, and network management. He has published over 200 papers, some of them are on world-wide famous journals or conferences including SCIENCE, IEEE Trans. on PAMI, IEICE, ICPR, ICCV, SIGIR, etc. His book "Network management" was awarded by the government of Beijing city as a finest textbook for higher education in 2004. His team got a number of prices in national and international academic competitions including: the first place in a national test of handwritten Chinese character recognition 1995, the first place in a national test of face detection 2004, the first place in a national test of text classification 2004, the first place of paper design competition held by IEEE Industry Application Society 2005, the second place in the competition of CSIDC held by IEEE Computer Society 2006.