Neurocomputing 113 (2013) 168–176

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

From the idea of "sparse representation" to a representation-based transformation method for feature extraction

Yong Xu [a,*], Qi Zhu [a], Zizhu Fan [a,d], Yaowu Wang [b], Jeng-Shyang Pan [c]

[a] Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[b] Shenzhen Key Laboratory of Urban Planning and Decision-Making Simulation, Shenzhen, China
[c] Innovative Information Industry Research Center, Shenzhen Graduate School, Harbin Institute of Technology, China
[d] School of Basic Science, East China Jiaotong University, Nanchang, China

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form
26 December 2012
Accepted 10 January 2013
Communicated by D. Tao
Available online 1 March 2013

Keywords:
Biometrics
Face recognition
Feature extraction
Sparse representation

ABSTRACT

Transformation methods have been widely used in biometrics such as face recognition, gait recognition and palmprint recognition. It seems that conventional transformation methods seem to be "optimal" for training samples but not for every test sample to be classified. The reason is that conventional transformation methods use only the information of training samples to obtain transform axes. For example, if the transformation method is linear discriminant analysis (LDA), then in the new space obtained using the corresponding transformation, the training samples must have the maximum between-class distance and the minimum within-class distance. However, it is hard to guarantee that the transformation also maximizes the between-class distance and minimizes the within-class distance of the test samples in the new space. Another example is that principal component analysis (PCA) can best represent the training samples with the minimum error; however, it is not guaranteed that every test sample can be also represented with the minimum error. In this paper, we propose to improve conventional transformation methods by relating the training phase with the test sample. The proposed method simultaneously uses both the training samples and test sample to obtain an "optimal" representation of the test sample. In other words, the proposed method not only is an improvement to the conventional transformation method but also has the merits of the representation-based classification, which has shown very good performance in various problems. Differing from conventional distance-based classification, the proposed method evaluates only the distances between the test sample and the "closest" training samples and depends on only them to perform classification. Moreover, the proposed method uses the weighted distance to classify the test sample. The weight is set to the representation coefficient of a linear combination of the training samples that can well represent the test sample.

1. Introduction

Biometrics such as face recognition has attracted much attention [1–4]. As we know, to properly represent the sample data is very important for pattern classification. A number of methods to represent the biometric sample have been proposed [5–7]. Among these methods, transformation methods have been widely used. Typical transformation methods include principal component analysis (PCA) [8–11], linear discriminant analysis (LDA) [12–17], minimum squared error (MSE) method [7,18], kernel LDA (KLDA) [19,20] and kernel (KPCA) [21–23]. A common characteristic of the transformation methods is to transform samples into a new space in which some properties might hold. For example, when transforming samples into the new space, PCA can make the samples in the new space have the most variance. LDA will enable the samples in the new space to have the maximum ratio of the between-class distance to the within-class distance. The MSE method aims at obtaining a mapping that can well transform the sample data into its class label. As KLDA and KPCA are the extensions of LDA and PCA, respectively, KLDA and KPCA have goals similar to those of LDA and PCA. However, KLDA and KPCA usually perform nonlinear transformations, whereas LDA and PCA produce linear transformations. Two-dimensional PCA (2DPCA) and two-dimensional LDA (2DLDA) have also been widely used as linear transformation methods [24–27]. Recently, transformation methods have been extended to the complex space in which the sample is denoted by a complex matrix or vector [10,16]. The recently proposed quaternion-based discriminant analysis can be viewed as an extension of the complex space based transformation methods and achieves good performance in color face recognition [28]. Other important extensions of conventional transform methods

* Corresponding author. Tel.: +86 755 2603 2458; fax: +86 755 2603 2461.
E-mail address: laterfall2@yahoo.com.cn (Y. Xu).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.01.036

include probabilistic principal component analysers [29], tensor-based transform methods such as tensor PCA and tensor LDA [30–34]. One noticeable advantage of the tensor-based transform methods is that they can cope with high-dimensional matrices [35–37], whereas previous transformation methods can work for only one-dimensional vectors or two-dimensional matrices. Actually, tensor PCA and tensor LDA can be viewed as a unified framework of PCA and LDA, respectively. The simple exponential family PCA (SePCA) is a generalized family of probabilistic principal component analysers and can well handle general types of observations [29].

Compared with conventional transformation methods, matrix factorization uses a very different way to represent the sample. For example, non-negative matrix factorization considers that the image should be represented by non-negative numbers and factorizes the sample matrix into two non-negative matrices to represent the image [38–40]. Moreover, Manhattan non-negative matrix factorization (MahNMF) proposed in [41] can robustly estimate the low-rank part and the sparse part of a non-negative matrix. Especially, it can do very well for the data contaminated by outliers. Another advantage of MahNMF is that its implementation is computationally efficient. The manifold learning method such as the graph embedding learning is also competent in data representation [42,43]. The max–min distance analysis based dimension reduction method can also perform well in representing the data [44]. Differing from conventional data representation methods, the method proposed by Bian et al. aimed at modifying the distance between samples for achieving better classification performance [45]. They also proposed a constrained empirical risk minimization framework [45]. This framework outperforms the representative distance metric learning algorithms and shows very good performance in both classification and image retrieval.

We note that transformation methods exploit only the training samples to implement their training phases. Generally, they work as follows: they first use the training samples to produce a number of transform axes and then exploit the transform axes to convert each training sample and testing sample into lower-dimensional representation. The transform axes obtained are "optimal" for the training samples; however, they are not optimal for the test sample to be classified. For example, PCA can best represent the training samples with the minimum error; nevertheless, it is not sure that every test sample can be also represented by PCA with the minimum error. Another example is that LDA allows the training samples to have the maximum ratio of the between-class distance to the within-class distance; however, this does not mean that each test sample is always very close to its class center and far from the centers of the other classes. Moreover, for the biometrics issue such as face recognition, the test sample is usually very different from the training sample from the same subject owing to varying illumination, facial expression and pose. Thus, it seems that the transform axes obtained using only the training samples is usually hard to maximize the between-class distance and minimize the within-class distance of the test samples in the new space. In this sense, we conclude that conventional transformation methods seem to be optimal for only training samples but not for every test sample to be classified. As a consequence, there is room to improve conventional transformation methods for pattern classification applications.

In this paper, inspired by the idea of "sparse representation" [1], we propose representation-based transformation methods by relating the feature extraction procedure with each test sample. The representation-based transformation method is not only optimal for training samples but also can well represent the test sample to be classified. In other words, the representation-based transformation method inherits the advantages of both transformation methods and "sparse representation", being able to exploit the statistical information of the training set and to identify and use the training samples that are most "related" to the test sample. Our method is also computationally much efficient than naïve sparse representation methods such as those in [46,47]. The experimental results show that the representation-based transformation methods can obtain a significant improvement in classification accuracy. The code of the proposed method can be downloaded at http://www.yongxu.org/lunwen.html.

The remainder of the paper is organized as follows. Sections 2 and 3 present the main steps and rationales of representation-based transformation methods, respectively. Section 4 shows the experimental results. Section 5 offers our conclusions.

2. The representation-based transformation methods

The representation-based transformation methods (RBTM) not only make the training samples in the new space hold the same property as those in the conventional transformation methods but also take advantages of the representation method. For example, the improvement to LDA not only enables the ratio of the between-class distance to the within-class distance of the training samples to be maximized but also tries to make the test sample to be well represented by a small number of training samples.

The proposed method consists of three steps. First of all, the first step exploits the training samples to represent the test sample. It then obtains the "closest" training samples of the test sample and the representation coefficient corresponding to each "closest" training sample. The second step implements the conventional transformation method to obtain the transform axes. The third step uses the transform axes, the "closest" training samples and the corresponding representation coefficients to produce features for all the samples and to calculate the distances between the test sample and the "closest" training samples. This step also uses the nearest neighbor classifier to classify the test sample. In Sections 2.1–2.3 we describe the first, second and third steps, respectively.

2.1. The first step of RBTM: determine the "closest" training samples

The first step of RBTM determines k nearest neighbors for each test sample from the set of training samples. We refer to them as the "closest" training samples of the test sample.

The first step of RBTM determines the k nearest neighbors for test sample y as follows: Let $x_i (i=1,\ldots,N)$ denote all the training samples in the form of column vectors. It is assumed that $y=\sum_{i=1}^{N} c_i x_i$ is approximately satisfied. The first step defines $C=[c_1 \ldots c_N]^T$ and obtains the solution of C using $\widehat{C}=(X^TX+\gamma I)^{-1}X^Ty$, where $X=[x_1 \ldots x_N]$, I is the identity matrix, and γ is a small positive constant. If $\widehat{c}_i$ is the i-th entry of $\widehat{C}$, then the "distance" between y and x_i is calculated using $e_i=\|y-\widehat{c}_i x_i\|$. Hereafter $\|.\|$ stands for the l_2 norm. The k training samples associated with the first k smallest "distances" are selected as k "closest" training samples of test sample y. The first step of RBTM also includes the following procedure. Let $x'_j (j=1,\ldots,k)$ denote the k "closest" training samples. The first step of RBTM also expects that a weighted sum of $x'_1,\ldots,x'_k$ can approximate test sample y. Thus, it also obtains k coefficients that approximately satisfy the following equation:

$$y=\sum_{j=1}^{k} b_j x'_j \quad (1)$$

Eq.(1) is solved using

$$\tilde{b}=(X'^TX'+\gamma I)^{-1}X'^Ty, \quad (2)$$

where $X' = [x'_1 \ldots x'_K]$, $\tilde{b} = [\tilde{b}'_1 \ldots \tilde{b}'_k]$. It seems that the above procedure is somewhat similar to the regression-based face representation proposed in [48].

We can view coefficient $\tilde{b}'_j$ $(j=1,\ldots,k)$ as the weight that denotes the importance of the j-th "closest" training sample x'_j. We also refer to $\tilde{b}'_j$ as the representation coefficient to show the connection between y and $x'_j (j=1,\ldots,k)$. A large representation coefficient usually means a strong connection.

2.2. The second step of RBTM: perform LDA or PCA

In this section we take the improvement to LDA or PCA as an example to describe the second step of RBTM. In this case, the second step of RBTM indeed implements conventional LDA or PCA to obtain a number of transform axes. For simplicity of presentation, we use x^i_j to denote the j-th training sample from the i-th class. The eigen-equation of conventional PCA is $S_t w = \lambda w$, where $S_t = 1/\mathrm{Ln}_i \sum_{i=1}^{L} \sum_{j=1}^{n_i} (x^i_j - \overline{m})(x^i_j - \overline{m})^T$. n_i stands for the number of the training samples from the i-th class. L and $\overline{m}$ denote the number of all the classes and the mean of all the training samples, respectively.

The eigen-equation of conventional LDA is $S_b w = \lambda S_w w$. S_b and S_w are the so-called between-class scatter matrix and within-class scatter matrix, respectively. They are defined as $S_b = \sum_{i=1}^{L} (m_i - \overline{m})(m_i - \overline{m})^T$ and $S_w = \sum_{i=1}^{L} \sum_{j=1}^{n_i} (x^i_j - m_i)(x^i_j - m_i)^T$. m_i stands for the mean of the training samples of the i-th class.

Conventional PCA and LDA solve the eigen-equation and take the eigenvectors corresponding to the first p largest eigenvalues as transform axes. Conventional PCA and LDA directly use these transform axes to transform every sample into a p-dimensional vector. However, as shown in Section 2.3, RBTM will simultaneously exploit the transform axes and representation coefficients to perform transform.

2.3. The third step of RBTM: feature extraction and classification

The third step of RBTM works as follows. Let $w_1, \ldots, w_p$ be p transform axes obtained using a conventional transformation method and $W = [w_1 \ldots w_p]$. The third step extracts features from the "closest" training samples and test sample y using

$$z_j = \tilde{b}'_j W^T x'_j \quad (j=1,\ldots,k), \tag{3}$$

$$z = W^T y. \tag{4}$$

z and z_j are the features of test sample y and the j-th "closest" training sample x'_j, respectively. As the feature extraction of the training sample is influenced by the corresponding representation coefficients, we also refer to the above feature extraction as adaptive feature extraction. The third step calculates the distances between test sample y and the "closest" training samples using

$$d_j = \|z - z_j\|. \tag{5}$$

If $q = \arg\min d_j$, then test sample y is classified into the same class as x'_q.

3. Analysis and rationale of RBTM

In this section we show the rationale of RBTM. First, RBTM shares the advantage of conventional transformation methods, i.e. in the new space the training samples satisfy some statistical properties which are beneficial to pattern classification. Second, as RBTM performs adaptive feature extraction, the "distance" between the test sample and training sample will be evaluated in a better way. In other words, RBTM regards that it is not necessary to evaluate the distances between the test sample and all the training samples and only the distances between the test sample and the "weighted" "closest" training samples should be calculated. The representation coefficient is taken as the weight which somewhat stands for the importance of the training sample in representing the test sample. We note that a similar idea of exploiting the training samples that are close to the test sample to perform classification has been applied by previous literatures [49–51].

If RBTM also adopts the conventional feature extraction procedure, then it extracts features of test sample y and the "closest" training samples using $g_j = W^T x'_j$ and $z = W^T y$.

As a result, Eq. (5) can be rewritten as

$$d_j = \|z - \tilde{b}'_j g_j\|, \quad j=1,\ldots,k. \tag{6}$$

If $q = \arg\min d_j$, then test sample y will be classified into the same class as x'_q. Thus, we can also view RBTM as a combination of a conventional feature extraction procedure and an improved nearest neighbor classifier. The distance defined in Eq.(6) is referred to as weighted distance. When implementing RBTM, we need to just carry out the second step one time whatever there are how many test samples. On the other hand, the first and third steps should run one time for each test sample.

Though the weights are obtained in the original space, the weights in the original space are somewhat consistent with those in the new space. The underlying evidence of the consistence can be shown as follows. Suppose that in the original space test sample y can be accurately represented by a weighted sum of all the training samples. In other words, $y = \sum_{i=1}^{N} a_i x_i = X\alpha$ is satisfied. $\alpha = [a_1 \ldots a_N]^T$ and $X = [x_1 \ldots x_N]$. x_i is the i-th training sample and a_i is its weight. $\|y - a_i x_i\|$ somewhat shows the "distance" relationship between the test sample and all the training samples. If $\|y - a_i x_i\|$ is small, we say that the test sample is "close" to the i-th training sample. Let $W = [w_1 \ldots w_p]$. $w_1, \ldots, w_p$ are p transform axes obtained using a conventional transformation method. It is easy to know that $W^T y = W^T X\alpha$ is also satisfied. $W^T y$ stands for the test sample in the new space. $W^T X = [W^T x_1 \ldots W^T x_N]$ and $W^T x_i$ is the i-th training sample in the new space. $y = X\alpha$ and $W^T y = W^T X\alpha$ indeed implies that the weights in the original space are consistent with those in the new space. Thus, it is reasonable to apply the weights obtained in the original space to the new space. Moreover, when we use Eq.(5) to evaluate the distance between the test sample and training sample, we indeed simultaneously exploit the information of the conventional transform method and representation method. Actually, our method uses the "closest" training samples and the "optimal" features extracted from conventional transform methods to classify the test sample.

The proposed method is also computationally efficient. Its computational cost is similar to that of the two norm based representation method. It is easy to know that the main computational cost of the two norm based representation method proposed in [53] is caused by calculating $(X^T X + \gamma I)^{-1} X^T y$. Suppose that the sample is a R-dimensional vector. For each test sample, to solve $(X^T X + \gamma I)^{-1} X^T y$ (X is the same as the X in Section 2) has the computational cost of $O(N^2 R + N^3)$. N is the number of all the training samples. R is usually greater than the number of all the training samples. If there are N' test samples, the computational cost of this method will be $O(N'N^2R + N'N^3)$. Moreover, under the same condition, the computational cost of the two norm based representation method proposed in [54] is much greater than $O(N'N^2R + N'N^3)$. Compared with the two norm based representation method proposed in [53] our method just needs extra computational operations to perform conventional transform, which is not high. Thus, it seems that our method has a comparable computational cost as the two norm based representation

method proposed in [53]. As we know, the naïve sparse representation method such as the one proposed by Wright et al. [46] has a very much higher computational cost than the two norm based representation method proposed in [53]. As a result, our method is computationally more efficient than the naïve sparse representation method.

Fig. 1 shows an illustration of the original Euclidean distances and weighted distances. In this illustration, all the training samples were used to represent the test sample. The original Euclidean distance is calculated using $\|y-x_i\|$, $i=1,\ldots,3$. x_1,x_2,x_3 stand for the first, second and third training samples, respectively y denotes the test sample. In this figure, the weighted distance is calculated using $\|y-c_ix_i\|$, $i=1,\ldots,3$, $[c_1c_2c_3]^T=(X^TX+\gamma I)^{-1}X^Ty$, $X=[x_1x_2x_3]$. γ was set to 0.01. This figure shows that the weighted distances might be very different from the original Euclidean distance. Especially, if a training sample has a negative weight coefficient, then the difference between the weighted distance and original distance might be very great. This figure tells us that in terms of the original distance, the first, second and third training samples are, respectively, the first, second and third nearest neighbors of the test sample. However, in terms of the weighted distance, the first, second and third training samples are, respectively, the first, third and second nearest neighbors of the test sample.

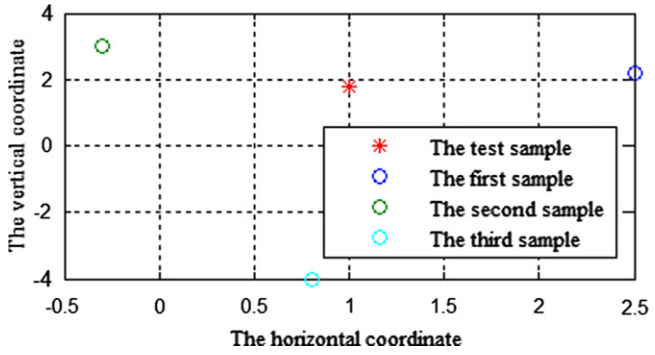
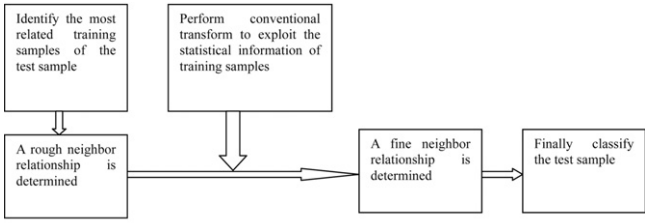


Fig. 1. An illustration of the original Euclidean distances and weighted distances. The horizontal and vertical coordinates show the values of the first and second components of the two-dimensional sample, respectively. The original Euclidean distances between the test sample and the first, second and third training samples are 1.5524, 1.7692 and 5.8034, respectively. The weighted distances between the test sample and the first, second and third training samples are 0.8203, 1.7680 and 1.7400, respectively. The weight coefficients of the first, second and third training samples are 0.4496, 0.1226 and −0.1106, respectively.

We also use the following flowchart to summarize the proposed method.



The flowchart shows that when Step 1 of the proposed method identifies the most related training samples of the test sample, it indeed determines a rough neighbor relationship between the test sample and training samples. In other words, Step 1 of the proposed method considers that it is only necessary to exploit the k "closest" training samples of the test sample to perform classification of the test sample. Moreover, the effect, on the classification of the test sample, of the first "closest" training sample is the greatest and the effect of the k-th "closest" training sample is the least! Then Step 2 of the proposed method performs

Table 1
Classification accuracies of different methods on the AR database. k is set to $k=0.1*N$. N is the total number of the training samples. The number of the transform axes used in conventional PCA is equal to the total number of the training samples minus one. The number of the transform axes used in conventional LDA is equal to the total number of the subjects minus one.

Number of training samples	2	3	4
Conventional PCA (%)	55.28	54.20	52.20
RBTM on PCA (%)	65.80	65.36	63.60
Conventional LDA (%)	60.83	59.82	60.76
RBTM on LDA (%)	66.18	65.47	63.03
INNC proposed in [53] (%)	60.07	59.60	58.33
SAFR proposed in [54] (%)	64.83	65.04	63.03

Fig. 2. Some face images of two subjects in the AR database.

conventional transform, exploiting the statistical information of training samples to obtain the transform axes. As a result, the transform results of the samples will be statistically "optimal". Take LDA as an example, the transform results of the samples will statistically maximize the between-class separability of the training samples and also minimize the within-class separability! This, of course, will statistically make the test samples easily be classified. We also say that Step 1 and Step 2 of the proposed method pay attention to making the consequent classification optimal for the test sample to be classified and for all the training samples, respectively. These two steps, of course, are very complement for each other. As a result, the simultaneous use of the information from the previous steps will obtain a better neighbor relationship between the test sample and training samples and obtains more accurate classification results.

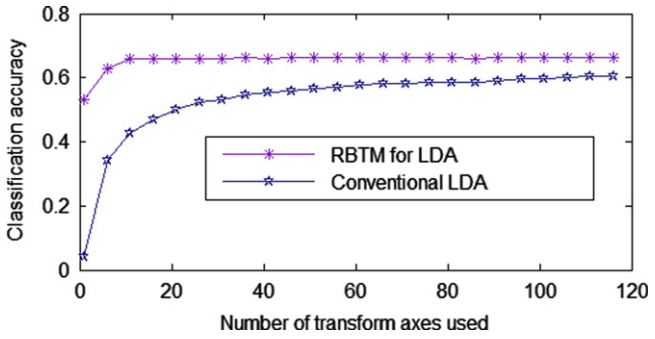


Fig. 3. The variation, with the number of transform axes used, of the classification accuracy of conventional LDA and RBTM on LDA on the AR database. The first two face images of each subject were used as training samples and the remaining images were used as test samples.

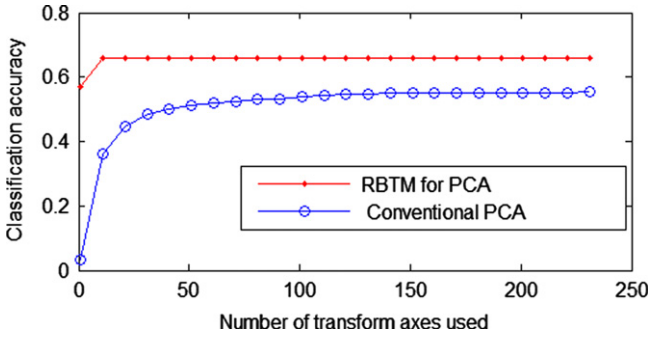


Fig. 4. The variation, with the number of transform axes used, of the classification accuracy of conventional PCA and RBTM on PCA on the AR database. The first two face images of each subject were used as training samples and the remaining images were used as test samples.

4. Experimental results

In this section we mainly conducted experiments to test RBTM on PCA and LDA. Conventional PCA and LDA are also tested. Moreover, the experiments on two representation-based classification methods, i.e. the improvement to the nearest neighbor classifier (INNC) [53] and simple and fast representation (SAFR) [54] were also performed. When we implemented conventional LDA we replaced the naïve within-class scatter matrix S_w by $S_w+0.01I$, where I is the identity matrix. We did this for avoiding the small sample size problem. In addition, in the case where only one training sample was available for each subject, we took $S_bw=\lambda w$ rather than $S_bw=\lambda S_w w$ as the eigen-equation of LDA. Actually, in this case S_w cannot be obtained.

4.1. Experimental results on the AR database

From the AR face database, we used 3120 images from 120 subjects. These images were taken in two sessions and each subject provided 26 images [52]. The face images of this database were obtained under the condition of varying pose, facial expression, or lighting. Occluded face images are also included in the AR face database. Fig. 2 shows some face images of two subjects in the AR database. We resized each face image from the AR database to a 40 by 50 matrix. We used only the first two, three, four samples from each class as the training samples and the others as the test samples, respectively. Before all the methods were implemented, all the training and test samples were normalized as vectors with length of 1.

Table 1 shows the experimental results. It tells us that RBTM can obtain an improvement in classification accuracy. The accuracy of RBTM on LDA is always greater than that of conventional LDA. The accuracy of RBTM on PCA is also clearly greater than that of conventional PCA. Moreover, the accuracy of RBTM is also higher than those of the representation methods, AINNC and SAFR. Figs. 3 and 4 show the variation, with the number of transform axes used, of the classification accuracy of conventional LDA (PCA) and RBTM on LDA (RBTM on PCA) on the visible light face images of the AR database. We also see that RBTM on LDA (RBTM on PCA) outperform the conventional LDA (PCA) again.

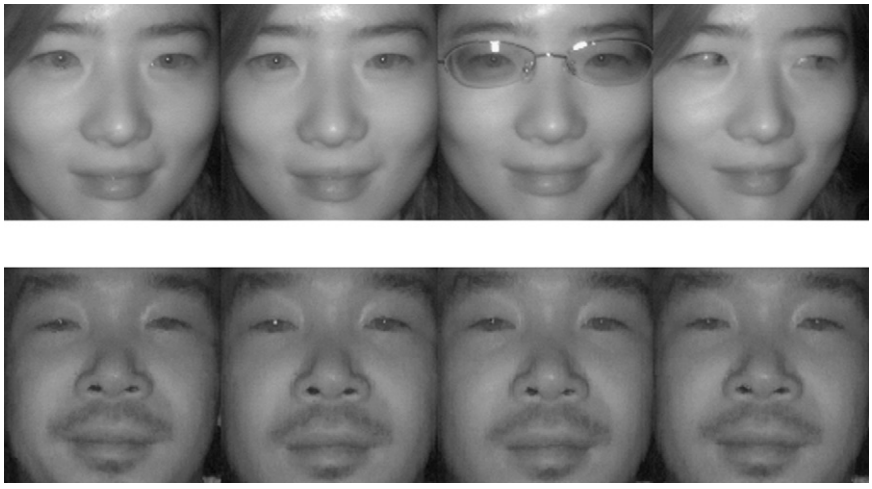


Fig. 5. The infrared face images of two subjects in the HFB database.

We see that besides the RBTM method can obtain higher classification accuracy than conventional transform methods, the classification accuracy of the RBTM method is also less sensitive to the number of the transform axes used in comparison with conventional transform methods. There are two main reasons. The first reason is that the RBTM method can better represent the test sample. Actually, since the RBTM method relates the training phase with the test sample, the representation of the test sample is obtained under the conditions that both the representation error is minimized and the statistical information of the training samples can be best exploited. As a result, the obtained representation of the test sample will be more reasonable, which is helpful for achieving high accuracy. The second reason is that the RBTM method somewhat inherits one performance characteristic of conventional transform methods, i.e. the classification accuracy might be very stable when the number of the transform axes used is large enough.

4.2. Experiment on the heterogeneous face biometrics (HFB) database

In this subsection, we used the visible light and near infrared face images of the heterogeneous face biometrics (HFB) database to test our method and the other methods. The original HFB database includes visible light, near infrared and three-dimensional (3D) face images [55,56]. There are 100 subjects and each subject provides four visible light and near infrared face images. Figs. 5 and 6 show the near infrared and visible light face images of two subjects in the HFB database, respectively. We first resized each original 128×128 face image into 64×64 image. The experiments were conducted on the visible light and near infrared face images, respectively. We took the first m near infrared or visible light face images of each subject as training samples and treated the remaining samples as test samples. m were set to 1 and 2, respectively. Before all the methods were implemented, every sample had been converted into a vector with the length of 1.

Table 2 shows the experimental results on the near infrared face images of the HFB database. Figs. 7 and 8 show the variation, with the number of transform axes used, of the classification accuracy of conventional LDA (PCA) and RBTM on LDA (RBTM on PCA) on the visible light face images of the HFB database. We see that RBTM on LDA and RBTM on PCA can achieve a much higher accuracy than conventional LDA and conventional PCA, respectively. Moreover, when a very small number of transform axes were used, conventional LDA and conventional PCA obtained a very low accuracy. However, RBTM on LDA and RBTM on PCA still achieved a high accuracy.

4.3. Experiment on the YALE database

The Yale face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 face images from 15 individuals each providing 11 images. These face images have various facial expressions and lighting conditions. In the experiment, each image was resized to 60×46 pixels. Fig. 9 shows sample images of one subject in the Yale database. The first three, four and five face images of each subject were used as training samples and the remaining images were taken as test samples, respectively.

Table 2
Classification accuracies of different methods on the near infrared face images of the HFB database. k is set to $k = 0.15{*}N$. N is the total number of the training samples. The number of the transform axes used in conventional PCA is equal to the total number of the training samples minus one. The number of the transform axes used in conventional LDA is equal to the total number of the subjects minus one.

	Visible face images		Near infrared face images	
Number of training samples	1	2	1	2
Conventional LDA (%)	55.0	74.0	72.7	80.0
RBTM on LDA (%)	87.0	93.0	93.0	95.0
Conventional PCA (%)	63.0	63.0	89.7	91.0
RBTM on PCA (%)	87.7	93.0	93.3	92.5%
INNC proposed in [53] (%)	89.00	93.00	93.3	90.5
SAFR proposed in [54] (%)	86.00	93.00	89.33	91.5

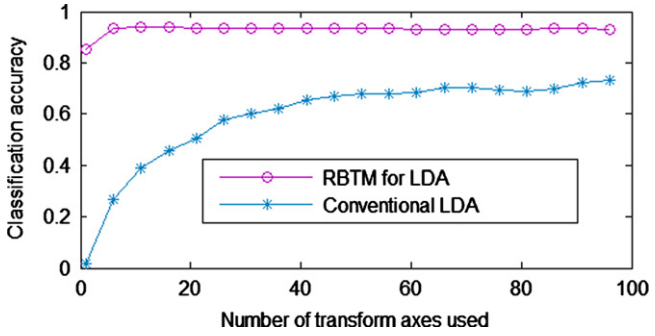


Fig. 7. The variation with the number of transform axes used for the classification accuracy of conventional LDA and RBTM on LDA on the visible light face images of the HFB database.

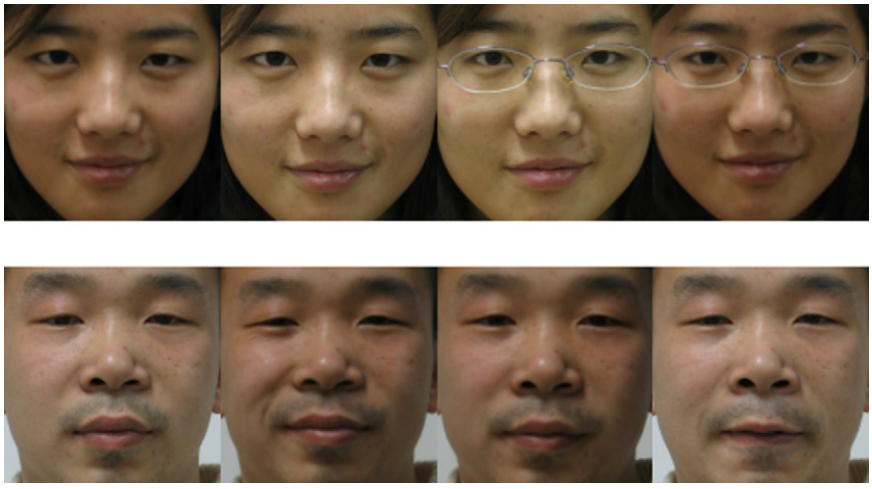


Fig. 6. The visible light face images of the same two subjects shown in Fig. 5.

Table 3 shows the experimental results on the YALE database. We see that our method also outperforms the others methods, i.e. conventional PCA, conventional LDA, INNC proposed in [53] and SAFR proposed in [54].

4.4. Experiment on the YALEB database

In this subsection we used the face images with the frontal pose from the YALEB database (http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html) to conduct the experiment. As a result, there were 450 face images and each subject provided 45 images. Each image was first resized to a 32 by 32 matrix. Fig. 10 shows sample images of two subjects in the YaleB database. The first 15, 20, 25 and 30 face images of each subject were used as training samples and the remaining images were taken as test samples, respectively. Table 4 shows the experimental results on the YALEB database. We see that our method also outperforms the others.

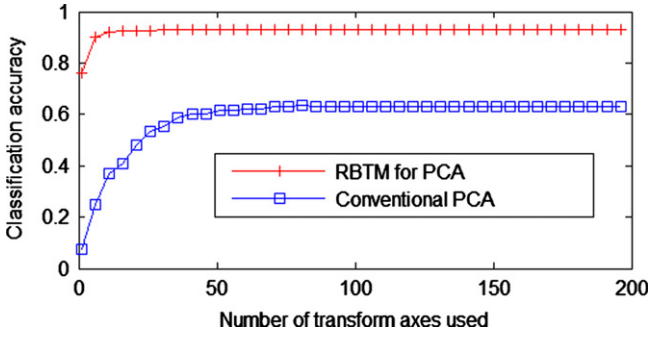


Fig. 8. The variation with the number of transform axes used for the classification accuracy of conventional PCA and RBTM on PCA on the visible light face images of the HFB database.

5. Conclusions

RBTM is motivated by the idea of "sparse representation" and is able to relate each test sample with the training samples. In other words, RBTM uses both the training samples and test sample to obtain an "optimal" representation of the test sample that is very beneficial to classification. RBTM has the merits of both transformation methods and "sparse representation". Actually, the use of the transformation method allows the

Table 4
Classification accuracies of different methods on the YALEB database. k is set to $k=0.3*N$. N is the total number of the training samples. The number of the transform axes used in conventional PCA is equal to the total number of the training samples minus one. The number of the transform axes used in conventional LDA is equal to the total number of the subjects minus one.

Number of training samples	15	20	25	30
Conventional PCA (%)	79.00	78.80	86.50	93.33
RBTM on PCA (%)	89.00	91.60	98.00	98.00
Conventional LDA (%)	48.33	36.00	38.50	37.33
RBTM on LDA (%)	88.00	92.40	97.00	98.67
INNC proposed in [53] (%)	87.00	84.80	87.50	95.33
SAFR proposed in [54] (%)	80.00	81.20	89.00	95.33

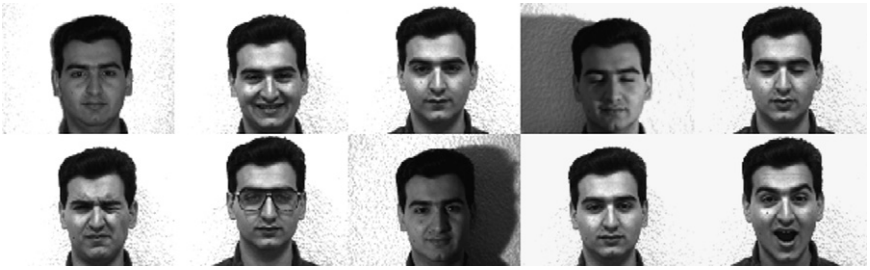


Fig. 9. Sample images of one subject in the Yale database.

Table 3
Classification accuracies of different methods on the YALE database. k is set to $k=0.2*N$. N is the total number of the training samples. The number of the transform axes used in conventional PCA is equal to the total number of the training samples minus one. The number of the transform axes used in conventional LDA is equal to the total number of the subjects minus one.

Number of training samples	3	4	5
Conventional PCA (%)	81.67	88.57	91.11
RBTM on PCA (%)	89.17	91.43	95.56
Conventional LDA (%)	68.33	74.29	71.11
RBTM on LDA (%)	88.33	91.43	95.56
INNC proposed in [53] (%)	82.50	91.43	94.44
SAFR proposed in [54] (%)	85.00	86.67	92.22

Fig. 10. Sample images of two subjects in the YaleB database. The first and second rows show the images of these two subjects, respectively.

representation-based transformation method to take advantages of the statistical information of the training samples and to denote the sample by a lower-dimensional vector. Moreover, the use of "sparse representation" enables the representation-based transformation method to determine and use the training samples that are most "related" to the test sample to perform classification. As only the distances between the test sample and the weighted "closest" training samples should be calculated and the "closest" training samples denotes those who are very close to the test sample, RBTM seems to be able to reduce the influence, on classification of the test sample of, the outliers. For conventional transformation methods, the features of the outliers might be very close to those of the test sample, though the outliers and the test sample are not from the same class. Another advantage of RBTM is that it is computationally much efficient than the naïve sparse representation methods. The experimental results show that RBTM obtains a significant improvement in classification accuracy. Though in semi-supervised learning the idea of linking training samples and test samples has been widely used [57–59] RBTM is very different from these methods.

Acknowledgments

This paper is partly supported by the Key Laboratory of Network Oriented Intelligent Computation, Program for New Century Excellent Talents in University (Nos. NCET-08-0156 and NCET-08-0155), National Nature Science Committee of China under Grant Nos. 61071179, 61203376, 61263032 and 61202276, as well as the Fundamental Research Funds for the Central Universities (HIT.NSRIF. 2009130). Jiangxi Provincial Natural Science Foundation of China under Grant 2010GQS0027, and the Science and Technology Foundation of Jiangxi Educational Committee of China (GJJ12309).

References

[1] M. Yang, L. Zhang, Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary, in: ECCV vol. 6, 2010,pp. 448–461.
[2] S.D. Lin, J.-H. Lin, C.-C. Chiang, Using gradient features from scale-invariant keypoints on face recognition, Int. J. Innovative Comput. Inf. Control 7 (4) (2011) 1639–1649, April.
[3] W. Yang, C. Sun, L. Zhang, A multi-manifold discriminant analysis method for image feature extraction, Pattern Recognition 44 (8) (2011) 1649–1657.
[4] C. Zhou, X. Wei, Q. Zhang, B. Xiao, Image reconstruction for face recognition based on fast ICA, Int. J. Innovative Comput. Inf. Control 4 (7) (2008) 1723–1732.
[5] S.-T. Pan, Application of integer FFT on performance improvement of speech recognition chip implementation, J. Inf. Hiding Multimedia Signal Process. 2 (4) (2011) 294–302.
[6] C.-Y. Chang, C.-W. Chang, C.-Y. Hsieh, Applications of block linear discriminant analysis for face recognition, J. Inf. Hiding Multimedia Signal Process. 2 (3) (2011) 259–269.
[7] C.-W. Tsai, K.-M. Cho, W.-S. Yang, Y.-C. Su, et al., A support vector machine based dynamic classifier for face recognition, Int. J. Innovative Comput. Inf. Control 7 (6) (2011).
[8] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: IEEE Conference on Computer Vision and Pattern Recognition, 1991, pp. 586–591.
[9] K.I. Diamantaras, S.Y. Kung, Principal Component Neural Networks: Theory and Applications, John Wiley & Sons, Inc., 1996.
[10] Y. Xu, D. Zhang, J.-Y. Yang, A feature extraction method for use with bimodal biometrics, Pattern Recognition 43 (2010) 1106–1115.
[11] S. Shan, B. Cao, Y. Su, L. Qing, X. Chen, W. Gao, Unified principal component analysis with generalized covariance matrix for face recognition, in: CVPR, 2008.
[12] S. Yan, D. Xu, Q. Yang, L. Zhang, et al., Multilinear discriminant analysis for face recognition, IEEE Trans Image Process. 16 (1) (2007) 212–220.
[13] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476.
[14] Y. Xu, F. Song, Feature extraction based on a linear separability criterion, Int. J. Innovative Comput. Inf. Control 4 (4) (2008) 857–865.
[15] J.Geoffrey McLachlan, Discriminant analysis and statistical pattern recognition, in: Wiley Series in Probability and Statistics, 2005.
[16] Y. Xu, D. Zhang, Represent and fuse bimodal biometric images at the feature level: a complex-matrix-based fusion scheme, Opt. Eng. 49 (3) (2010), March.
[17] M. Kan, S. Shan, Y. Su, X. Chen, W. Gao, Adaptive discriminant analysis for face recognition from single sample per person, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2011, pp. 193-199.
[18] S.A. Billings, K.L. Lee, Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm, Neural Networks, 15 (2) (2002) 263–270.
[19] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (2000) 2385–2404.
[20] Y. Xu, J.-Y. Yang, J. Lu, D.-J. Yu, An efficient renovation on kernel fisher discriminant analysis and face recognition experiments, Pattern Recognition 37 (2004) 2091–2094.
[21] B. Scholkopf, A. Smola, K.R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[22] B. Scholkopf, A. Smola, K.R. Muller, Kernel Principal Component Analysis, Artificial Neural Networks – ICANN'97, Berlin, 1997, pp. 583–588.
[23] J. Li, X. Li, D. Tao, KPCA for semantic object extraction in images, Pattern Recognition 41 (10) (2008) 3244–3250.
[24] J. Yang, D. Zhang, A.F. Frangi, J.-Y. Yang, Two dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. Pattern Anal. Machine Intell. 26 (1) (2004) 131–137.
[25] M. Li, B.g Yuan, 2D-LDA: a statistical linear discriminant analysis for image matrix, Pattern Recognition Lett. 26 (2005) 527–532.
[26] Y. Xu, D. Zhang, J. Yang, J.-Y. Yang, An approach for directly extracting features from matrix data and its application in face recognition, Neurocomputing 71 (2008) 1857–1865.
[27] Y. Pang, D. Tao, Y. Yuan, X. Li, Binary two-dimensional PCA, IEEE Trans. Syst. Man Cybern. Part B 38 (4) (2008) 1176–1180.
[28] Y. Xu, Quaternion-based discriminant analysis method for color face recognition, PLoS ONE 7 (8) (2012) e43493, http://dx.doi.org/10.1371/journal.pone.0043493.
[29] J. Li, D. Tao, Simple Exponential Family PCA, J. Mach. Learning Res. 2010 9:453–460 (Proceedings Track 2010 Journal of Machine Learning Research - Proceedings Track).
[30] J. Wen, X. Gao, Y. Yuan, D. Tao, J. Li, Incremental tensor biased discriminant analysis: a new color-based visual tracking method, Neurocomputing 73 (4–6) (2010) 827–839.
[31] Y. Mu, W. Ding, M. Morabito, D. Tao, Empirical discriminative tensor analysis for crime forecasting, in: KSEM, 2011, pp. 293–304.
[32] J. Wen, X. Gao, X. Li, D. Tao, Incremental learning of weighted tensor subspace for visual tracking, in: IEEE International Conference on Systems, Man and Cybernetics, 2009, pp. 3688–3693.
[33] D. Tao, X. Li, X. Wu, et al., Tensor rank one discriminant analysis – a convergent method for discriminative multilinear subspace selection, Neurocomputing 71 (10-12) (2008) 1866–1882.
[34] J. Sun, D. Tao, S. Papadimitriou, P.S. Yu, Christos Faloutsos: incremental tensor analysis: theory and applications, ACM Trans. Knowl. Discovery Data 2 (3) (2008).
[35] D. Tao, X. Li, X. Wu, S.J. Maybank, General tensor discriminant analysis and Gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1700–1715.
[36] D. Tao, X. Li, Xindong Wu, et al., Supervised tensor learning, Knowl. Inf. Syst. 13 (1) (2007) 1–42.
[37] D. Tao, M. Song, X. Li, et al., Bayesian tensor approach for 3-D face modeling, IEEE Trans. Circuits Syst. Video Technol. 18 (10) (2008) 1397–1410.
[38] N. Guan, D. Tao, Z. Luo, B. Yuan, NeNMF: an optimal gradient method for nonnegative matrix factorization, IEEE Trans. Signal Process. 60 (6) (2012) 2882–2898.
[39] D. Guillamet, J. Vitrià, B. Schiele, Introducing a weighted non-negative matrix factorization for image classification, Pattern Recognition Lett. 24 (14) (2003) 2447–2454.
[40] N. Guan, D. Tao, Z. Luo, et al., Non-negative patch alignment framework, IEEE Trans. Neural Networks 22 (8) (2011) 1218–1230.
[41] N. Guan et al., MahNMF: Manhattan Non-negative Matrix Factorization CoRR abs/1207.3438, 2012.
[42] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 328–340.
[43] B. Geng, et al., Ensemble manifold regularization, IEEE Trans. Pattern Anal. Mach. Intell. 34 (6) (2012) 1227–1233.
[44] W. Bian, et al., Max-min distance analysis by using sequential SDP relaxation for dimension reduction, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 1037–1050.
[45] W. Bian, et al., Constrained empirical risk minimization framework for distance metric learning, IEEE Trans. Neural Networks: Learn. Syst. 23 (8) (2012) 1194–1205.
[46] J. Wright, A.Y. Yang, A. Ganesh, et al., Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[47] J. Yang, L. Zhang, Y. Xu, J.-Y. Yang, Beyond sparsity: the role of L_1-optimizer in pattern classification, Pattern Recognition 45 (3) (2012) 1104–1118.
[48] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 2106–2112.
[49] Y. Xu, D. Zhang, J. Yang, J.Y. Yang, et al., A two-phase test sample sparse representation method for use with face recognition, IEEE Trans. Circuits and Systems for Video Technology 21 (9) (2011) 1255–1262.
[50] Z. Fan, Y. Xu, D. Zhang, Local linear discriminant analysis framework using sample neighbors, IEEE Trans. Neural Networks 22 (7) (2011) 1119–1132.

[51] M. Loog, D.d. Ridder, Local discriminant analysis, in: ICPR 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, 2006, pp. 328–331.
[52] 〈http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html〉.
[53] Y. Xu, Q. Zhu, Y. Chen, J.-S. Pan, An improvement to the nearest neighbor classifier and face recognition experiments, Int. J. Innovative Comput. Inf. Control 8 (12) (2012), December.
[54] Y. Xu, Q. Zhu, A simple and fast representation-based face recognition method, Neural Comput. Appl., 10.1007/s00521-012-0833-5.
[55] S.Z. Li, Z. Lei, M. Ao, The HFB face database for heterogeneous face biometrics research, in: Sixth IEEE Workshop on Object Tracking and Classification Beyond and in the Visible Spectrum (OTCBVS, in conjunction with CVPR 2009). Miami, Florida, June, 2009.
[56] 〈http://www.cbsr.ia.ac.cn/english/HFB%20Databases.asp〉.
[57] D. Zhou, J. Huang, et al., Learning from labeled and unlabeled data on a directed graph, ICML (2005) 1036–1043.
[58] X. Zhu, Z. Ghahramani, J.D. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, ICML (2003) 912–919.
[59] D. Zhou, O. Bousquet, T.N. Lal, et al., Learning with local and global consistency, in: NIPS 2003.

Yong Xu was born in Sichuan, China, in 1972. He received his B.S. degree, M.S. degree in 1994 and 1997, respectively. He received the Ph.D. degree in Pattern Recognition and Intelligence system at NUST (China) in 2005. Now he works at Shenzhen Graduate School, Harbin Institute of Technology. His current interests include pattern recognition, biometrics, machine learning and video analysis.

Qi Zhu obtained his Master degree in 2009 from Shenzhen graduate school, Harbin Institute of Technology. He is working for his Ph.D. degree at Shenzhen graduate school, Harbin Institute of Technology. His current interests include pattern recognition and biometrics.

Zizhu Fan received the M.S. degree in computer science from Hefei University of Technology, Hefei, China, in 2003. He is currently pursuing the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. He has published more than 10 journal papers. His current research interests include pattern recognition and machine learning.

Yaowu Wang is with Shenzhen Graduate School, Harbin Institute of Technology. He is the vice director of Shenzhen Key Laboratory of Urban Planning and Decision-Making Simulation. His main research interest is ecological urban design as well as landscape planning and design.

Jeng-Shyang Pan received the B.S. degree in Electronic Engineering from the National Taiwan University of Science and Technology in 1986, the M.S. degree in Communication Engineering from the National Chiao Tung University, Taiwan in 1988, and the Ph.D. degree in Electrical Engineering from the University of Edinburgh, UK in 1996. Currently, he is the Doctoral Advisor in the Harbin Institute of Technology and Professor in the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan. He has published more than 400 papers in which 110 papers are indexed by SCI. He is the IET Fellow, UK and the Tainan Chapter Chair of IEEE Signal Processing Society. He was Awarded Gold Prize in the International Micro Mechanisms Contest held in Tokyo, Japan in 2010. He was also awarded Gold Medal in the Pittsburgh Invention & New Product Exposition (INPEX) in 2010, Gold Medal in the International Exhibition of Geneva Inventions in 2011 and Gold Medal of the IENA, International "Ideas – Inventions – New products", Nuremberg, Germany. He was offered Thousand-Elite-Project in China. He is on the editorial board of International Journal of Innovative Computing, Information and Control, LNCS Transactions on Data Hiding and Multimedia Security, and Journal of Information Hiding and Multimedia Signal Processing. His current research interests include soft computing, robot vision and signal processing.