Feature Dimension Reduction of NaXi Pictographs Characters Recognition based on LDA

Hai Guo[1], Jinghua Yin[2], Jingying Zhao[3]

**[1] Harbin University of Science and Technology,
Harbin, 150001, P.R. China
And
Department of Computer Science and Engineering, Dalian Nationalities University,
DaLian, 11660, P.R.China**

**[2] Harbin University of Science and Technology,
Harbin, 150001, P.R. China**

**[3] Department of Computer Science and Engineering, Dalian Nationalities University,
DaLian, 11660, P.R.China**

Abstract

As a kind of pictographic character, Naxi pictographs character has received little academic attention. Proposing dimension reduction method of Naxi pictographs characters on the basis of LDA (Linear Discriminant Analysis), this paper thus makes an in-depth study of feature dimension reduction, an important issue in the recognition of Naxi pictographs characters. By constructing a recognition sample library involving four features of grid feature, permeability number feature, moment invariant feature, and directional element feature (DEF), 50% of data are selected from sample library as training set and testing set respectively. Two dimension reduction methods of LDA and FA (Factor Analysis) are applied to dimension reduction experiment of features of Naxi pictographs characters. The experiment result proves LDA method to be significantly superior to FA method, as LDA method could still maintain a 99% recognition precision when the dimension is reduced to 10% of the original dimension.

Keywords: NaXi pictographs character, Linear Discriminant Analysis (LDA), Dimension reduction, features extraction, characters.

1. Introduction

The optical character recognition (OCR) process includes image acquisition, image preprocessing, layout analysis, text line segmentation, feature extraction, character recognition, and post processing. For any character recognition method, it's critical to extract the class correlated features from character images to maximize the mutual information[1]. However, this is a challenge owing to many factors, including huge numbers of characters, noise, different fonts, various character types, and complicated document layouts.

Generally, character recognition includes text information collection, information analysis and processing, information classification and discrimination, and so on. Information collection means that gray of characters on paper will be converted into electrical signals, which can be input into computers. Information collection is based on paper feeding mechanisms and photoelectric conversion devices in character recognition reader, flying-spots scanners, video camera, photosensitive components, laser scanners, and other photoelectric conversion devices. Information analysis and processing eliminate the noises and disturbance caused by printing quality, paper quality, writing instruments and other factors [2-3]. It can normalize size, deflexion, shade and thickness. Information classification and discrimination can remove the noises, normalize the character information, classify the character information and output the recognition results. Feature extraction is a crucial part of character recognition.

As a kind of pictographic character, most of Naxi pictographs character have numerous strokes, with typical Naxi pictographs characters shown in Figure 1[4-6]. The feature dimension of Naxi pictographs characters is also much higher than that of other characters. Therefore, dimension reduction constitutes an important research direction in the recognition of Naxi pictographs characters. This paper makes a dimension reduction of features of Naxi pictographs characters through LDA (Linear Discriminant Analysis). It is mainly composed of five parts. Part one is introduction. Part two is a brief review of

information processing of Naxi pictographs characters. Part three talks about LDA dimension reduction algorithm. Part four includes experiment setting and result. Part five is conclusion.

Fig. 1 NaXi Pictograph script

2. Overview of NaXi Pictographs

Naxi (a.k.a. Nakhi, Naqxi, Nasi) belongs to the Yi branch of the Tibeto-Burman language family. About 300,000 people speak Naxi in the Chinese provinces of Yunnan and Sichuan, particularly in the Lijiang region. There are also Naxi people in Tibet and possibly in Burma/Myanmar. The NaXi Pictographs (Naxi Dongba script) was reputedly invented by King Moubao Azong in the 13th century. It is used exclusively by the Dongba (shamans/priests) as an aid to the recitation of ritual texts during religious ceremonies and shamanistic rituals[7].

Use of the Naxi language and script was discouraged after the Communist victory in 1949, and they were actively suppressed during the Cultural Revolution in the 60s when thousands of manuscripts were destroyed. Today there are about 60 Dongba priests who can read and write the Dongba script. Most are over 70, though at least three are under 30. In an effort to revive the script, the younger Dongbas frequently visit local schools in the Lijiang region to teach classes on it.

A newspaper was published during the 1980s printed in the Dongba script and the Latin alphabet in an attempt to increase the level of literacy among the Naxi people in their own language. Over 30 books were also published. Their efforts were successful at first - in 1982, 200 people could read Naxi in the Latin alphabet. By 1985, 1,700 could do so. The Chinese government phased out Naxi language teaching in the late 80s, but recently effort have been made to start teaching the language again. The script is also used on road signs and shop names in Lijiang[4,5].

3. Linear Discriminant Analysis

3.1 Overview of Linear Discriminant Analysis

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events[6]. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. Linear discriminant analysis (LDA) is a classical statistical approach for supervised dimensionality reduction and classification.[8] LDA computes an optimal transformation (projection) by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination. The optimal transformation in LDA can be readily computed by applying an eigendecomposition on the so-called scatter matrices. It has been used widely in many applications involving high-dimensional data However classical LDA requires the so-called total scatter matrix to be nonsingular. In many applications involving high-dimensional and low sample size data, the total scatter matrix can be singular since the data points are from a very high-dimensional space, and in general the sample size does not exceed this dimension. This is the well-known singularity or undersampled problem encountered in LDA[9].

3.2 Linear Discriminant Dimensionality Reduction

In this section, we will integrate Fisher score and Linear Discriminant Analysis in a unified framework. The key idea of our method is to find a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion. It can be mathematically formulated as follows,

$$\arg\max_{W,p} tr\begin{Bmatrix}\left(W^T diag(p) S_t diag(p) W^{-1}\right)\\ \left(W^T diag(p) S_b diag(p) W\right)\end{Bmatrix}, \quad (1)$$
$$s.t. p \in \{0,1\}^d, p^T 1 = m$$

which is a mixed integer programming. Eq. (1) is called as *Linear Discriminant Dimensionality Reduction* (LDDR) because it is able to do feature selection and subspace learning simultaneously.

Recent studies established the relationship between LDA and multivariate linear regression problem, which provides a regression-based solution for LDA. This motivates us to solve the problem in Eq. (1) in a similar manner. In the following, we present a theorem, which

establishes the equivalence relationship between the problem in Eq. (1) and the problem in Eq. (2).

Theorem 1. *The optimal* **p** *that maximizes the problem in Eq. (1) is the same as the optimal* **p** *that minimizes the following problem*

$$\arg\min_{P,W} \frac{1}{2}\left\|X^T diag(p)W - H\right\|_F^2 \quad s.t. p \in \{0,1\}^d, p^T 1 = m \tag{2}$$

Where $H = [h_1, \cdots, h_c] \in \Re^{n \times c}$, and h_k *is a column vector whose i-th entry is given by*

$$h_{ik} = \begin{cases} \sqrt{\frac{n}{n_k}} - \sqrt{\frac{n_k}{n}}, if & y_i = k \\ -\sqrt{\frac{n_k}{n}}, & otherwise \end{cases} \tag{3}$$

In addition, the optimal W_1 *of Eq. (1) and the optimal* W_2 *of Eq. (2) have the following relation*

$$W_2 = [W_1, 0]Q^T \tag{4}$$

under a mild condition that

$$rank(S_t) = rank(S_b) + rank(S_w) \tag{5}$$

and Q *is a orthogonal matrix.*

Note that the above theorem holds under the condition that X is centered with zero mean. Since $rank(S_t) = rank(S_b) + rank(S_w)$ holds in many applications involving high-dimensional and under-sampled data, the above theorem can be applied widely in practice.

According to theorem 1, the difference between W_1 and W_2 is the orthogonal matrix Q Since the Euclidean distance is invariant to any orthogonal transformation, if a classifier based on the Euclidean distance is applied to the dimensionality-reduced data obtained by W_1 and W_2, they will achieve the same classification result.

Suppose we find the optimal solution of Eq. (2), i.e., W^* and p^*, then p^* is a binary vector, and $diag(p)W$ is a matrix where the elements of many rows are all zeros. This motivate us to absorb the indicator variables **p** into **W**, and use $L_{2,0}$-norm on **W** to achieve feature selection, leading to the following problem

$$\arg\min_W \frac{1}{2}\left\|X^T W - H\right\|_F^2 \quad s.t. \|W\|_{2,0} \le m \tag{6}$$

However, the feasible region defined by $\|W\|_{2,0} \le m$ is not convex. We relax $\|W\|_{2,0} \le m$ to its convex hull, and obtain the following relaxed problem,

$$\arg\min_W \frac{1}{2}\left\|X^T W - H\right\|_F^2 \quad s.t. \|W\|_{2,1} \le m \tag{7}$$

Note that Eq. (7) is no longer equivalent to Eq. (1) due to the relaxation. However, the relaxation makes the optimization problem computationally much easier. In this sense, the relaxation can be seen as a tradeoff between the strict equivalence and computational tractability.

Eq. (7) is equivalent to the following regularized problem,

$$\arg\min_W \frac{1}{2}\left\|X^T W - H\right\|_F^2 + \mu\|W\|_{2,1} \tag{8}$$

4. Experiments and Results

4.1 Data

In order to verify the validity of the method in this paper, we construct sample database of NaXi pictography. An automatic generation program is developed. In Figure 2, we can use this tool to extract images of NaXi pictograph character directly from the sample database. 100 images of different sizes are extracted for each character. After pretreatment, these pictures are normalized into 64x64 sizes. The sample recognition database consists of 210,000 NaXi pictograph characters.

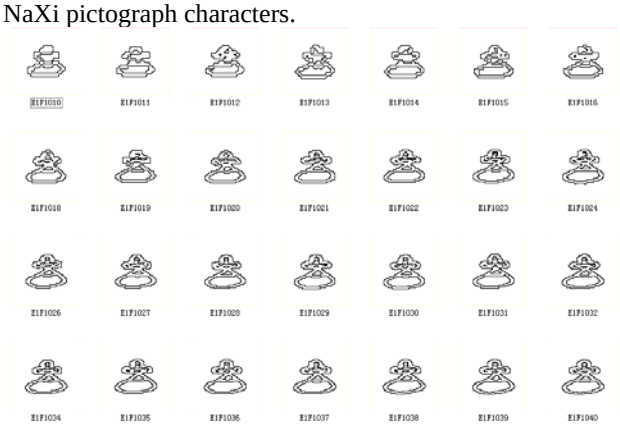


Fig.2 NaXi Pictographs character database

4.2 Feature Extraction of Naxi Pictographs

Recognition features of Naxi pictographs characters are mainly divided into four types: coarse grid, permeability

number, moment invariant feature and directional element feature (DEF). These four kinds of features are most commonly used in character recognition.

Coarse grid feature belongs to a local feature in the statistics feature, reflecting the distribution of the overall shape of the character. It divides sample character into M*M grids, and then makes statistics of the number of pixels in every grid. During the recognition stage, it could combine statistic feature of every grid as character statistic feature to realize character recognition, as the pixel in every grid could reflect a part of features of character.

We have also selected moment invariant feature in the recognition of Naxi pictographs characters. The moment invariant is widely applied to image retrieval and image recognition. As certain moments in the image region has some invariable characteristics of the geometry changes of translation, rotation and scaling, the representation of moment is of great significance in object classification and recognition. In image processing, geometric moment invariant can be used as an important feature to represent objects, and this feature could be applied to operations like image classification. Common moments include HU moment and Zernike moment.

In model recognition, an important problem lies in the recognition of directional change of target. Zernike moment is orthogonal, characterized by rotation invariance. In other words, the rotation target does not change its mode value. As Zernike moment can construct any high order moment, the recognition effect of Zernike moment is superior to other methods.

Directional element, as a kind of effective feature, is widely applied to the recognition of various characters and has achieved good effects. Naxi pictographs characters are pictographic, without various radicals and strokes in Chinese characters. Instead, Naxi pictographs characters are combined by a variety of primitives and basic dollars, which are the constituents of Naxi pictographs characters. The primitive and basic dollar of every Naxi pictographs character has its specific structure, of which the feature could be reflected from three aspects of level, local part, and detailed. The directional element is exactly the effective means to portray these structural features .

4.3 Classifier

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

The Euclidean distance between points p and q is the length of the line segment connecting them ($\overline{pq}$). In Cartesian coordinates, if $p=(p_1,p_2,\cdots,p_n)$ and $q=(q_1,q_2,\cdots,q_n)$ are two points in Euclidean n-space, then the distance from p to q, or from q to p is given by:

$$\begin{aligned} &d(p,d)=d(q,p)\\ &=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} \\ &=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \end{aligned} \quad (9)$$

The position of a point in a Euclidean n-space is a Euclidean vector. So, p and q are Euclidean vectors, starting from the origin of the space, and their tips indicate two points. The Euclidean norm, or Euclidean length, or magnitude of a vector measures the length of the vector:

$$\|p\|=\sqrt{p_1^2+p_2^2+\cdots+p_n^2}=\sqrt{p\cdot p} \quad (10)$$

A vector can be described as a directed line segment from the origin of the Euclidean space (vector tail), to a point in that space (vector tip). If we consider that its length is actually the distance from its tail to its tip, it becomes clear that the Euclidean norm of a vector is just a special case of Euclidean distance: the Euclidean distance between its tail and its tip. The distance between points p and q may have a direction (e.g. from p to q), so it may be represented by another vector, given by

$$q-p=(q_1-p_1,q_2-p_2,\cdots,q_n-p_n) \quad (11)$$

In a three-dimensional space (n=3), this is an arrow from p to q, which can be also regarded as the position of q relative to p. It may be also called a displacement vector if p and q represent two positions of the same point at two successive instants of time. The Euclidean distance between p and q is just the Euclidean length of this distance (or displacement) vector:

$$\|q-p\|=\sqrt{(q-p)\cdot(q-p)} \quad (12)$$

which is equivalent to equation 1, and also to:

$$\|q-p\|=\sqrt{\|p\|^2+\|q\|^2-2p\cdot q} \quad (13)$$

In general, for an n-dimensional space, the distance is

$$d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_i-q_i)^2+\cdots+(p_n-q_n)^2} \quad (14)$$

4.4 Method and Experimental Results

In order to validate the effects of LDA dimension reduction method on the character recognition of Naxi pictographs characters, we have made pre-treatment, extracted outline and constructed sample library about 210000 images through the four features of grid feature,

Table 1. LDA dimension reduction

permeability number+coarse grid+directional harness+moment invariant					moment invariant+ permeability number +directional element				
Dimension reduction proportion	Feature dimension	Training set classification precision	Testing set classification precision	Generalization precision	Dimension reduction proportion	Feature dimension	Training set classification precision	Testing set classification precision	Generalization precision
100%	284	97.502948	96.468897	0.989395	100%	220	97.39682	96.33255	
90%	256	99.9794	99.8541	0.9987	90%	198	99.941	99.7767	0.9984
80%	227	99.9823	99.8511	0.9987	80%	176	99.9381	99.7745	0.9984
70%	199	99.9794	99.8474	0.9987	70%	154	99.9233	99.7642	0.9984
60%	170	99.9823	99.8438	0.9986	60%	132	99.8939	99.7597	0.9987
50%	142	99.9823	99.832	0.9985	50%	110	99.8703	99.7126	0.9984
40%	114	99.9499	99.7818	0.9983	40%	88	99.77	99.5718	0.998
30%	85	99.8703	99.661	0.9979	30%	66	99.4045	99.3595	0.9995
20%	57	99.6875	99.4001	0.9971	20%	44	99.8732	99.6212	0.9975
10%	28	93.4758	92.7624	0.9924	10%	22	99.605	99.0603	0.9945

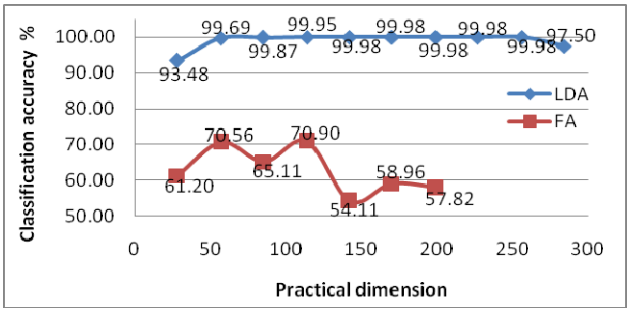
Table 2. Factor Analysis dimension reduction

permeability number +coarse grid+ directional element +moment invariant					moment invariant +permeability number +directional element				
Dimension reduction proportion	Feature dimension	Training set classification precision	Testing set classification precision	Generalization precision	Dimension reduction proportion	Feature dimension	Training set classification precision	Testing set classification precision	Generalization precision
100%	284	-	-	-	100%	220	97.39682	96.33255	-
90%	256	-	-	-	90%	198	61.5094	56.2006	0.9137
80%	227	-	-	-	80%	176	63.1751	57.9312	0.917
70%	199	57.8154	52.2067	0.903	70%	154	61.5153	55.88	0.9084
60%	170	58.9593	53.4095	0.9059	60%	132	71.1232	66.5817	0.9361
50%	142	54.1126	48.5952	0.898	50%	110	68.2134	63.492	0.9308
40%	114	70.8962	75.4752	1.0646	40%	88	70.467	75.2311	1.0676
30%	85	65.1144	70.1769	1.0777	30%	66	74.4257	78.8149	1.059
20%	57	70.5649	75.3042	1.0672	20%	44	72.0024	76.6132	1.064
10%	28	61.204	65.8821	1.0764	10%	22	59.2335	64.1333	1.0827

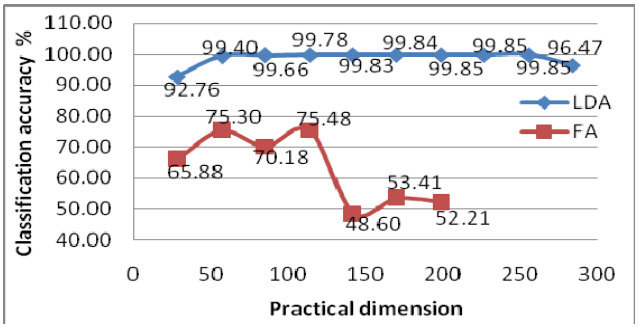
permeability number feature, moment invariant, and directional element feature. Among them, the combination of three features of permeability number +coarse grid+ directional harness +moment invariant is sample library 1 and the combination of three features of moment invariant +permeability number +directional element is sample library 2. 50% of each sample library is extracted as the training sample, and 50% as testing sample. Two methods of Factor Analysis dimension reduction and LDA dimension reduction are applied to the treatment of dimension reduction. The two groups of experimental data in Table 1 and Table 2 are gained.

Through the data analysis in Figure 3 and Figure 4, it is found that in sample library 2 that applies LDA method, when the dimension reduces to 10%, the recognition precision is only 99.0603%. However, when LDA dimension reduction is not applied, the recognition precision is only 96.33255%. Analyzing from the curve in

the figure, when the dimension reduction is 10%-90%, the recognition precision has no great changes, maintaining roughly above 99%. The experiment results above indicate that the method of LDA dimension reduction could not only reduce feature dimension of Naxi pictographs characters, but also effectively improve recognition precision.

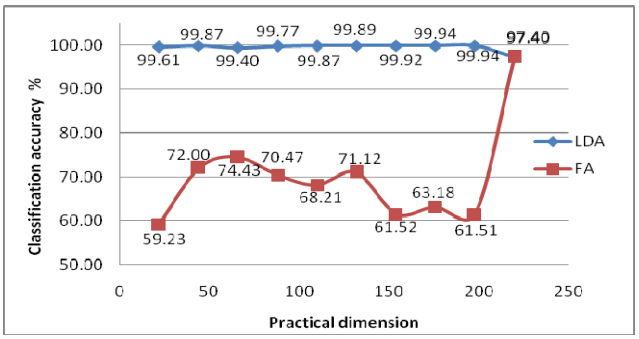


a Contrast of classification precision after dimension reduction using training set

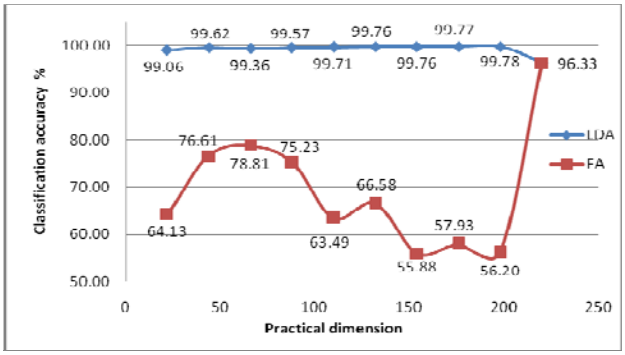


b. Contrast of classification precision after dimension reduction using testing set

Figure 3. Contrast of feature dimension reduction recognition precision of LDA and FA in terms of features of permeability number, coarse grid, directional harness



a Contrast of classification precision after dimension reduction using training set



b. Contrast of classification precision after dimension reduction using testing set

Figure 4. Contrast of feature dimension reduction recognition precision of LDA and FA in terms of features of moment invariant, permeability number and directional element feature

Through its contrast between the method of Factor Analysis dimension reduction, it is found that in both sample libraries, LDA dimension reduction effect and recognition precision are both higher than the traditional method of Factor Analysis dimension reduction.

5 Conclusions

Recognition dimension reduction of Naxi pictographs characters is a both significant and difficult job. This paper studies this issue and applies LDA dimension reduction method to make dimension reduction experiment of a variety of feature combinations of Naxi pictographs characters. The experiment indicates that the method in this paper could effectively reduce dimension of Naxi pictographs characters. When the dimension is reduced to 10%, it could still maintain a recognition precision of 99.0603%. In the future, we will continue to apply more advanced dimension reduction methods such as ISOMAP and LLE to further reduce feature dimension of Naxi pictographs characters and improve recognition speed and recognition efficiency.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No. 60803096 and Fundamental Research Funds for the Central Universities.

References

[1] M. Ayatullah Faruk, M. Nabamita, B. Subhadip, N. Mita, "Design of an optical character recognition system for camerabased handheld devices", International Journal of Computer Science Issues, 2011, Vol.8, No.4 , pp 283-289.
[2] P. Giuseppe, and I. Donato, "Adaptive membership functions for handwritten character recognition by Voronoi-based

image zoning", IEEE Transactions on Image Processing, 2012, Vol. 21, No. 9, pp. 3827-3837.
[3] G.P. Liu , D. J. Zhao, H. Huang , "Character recognition of license plate using autoencoder neural network reconstruction", Guangdianzi Jiguang/Journal of Optoelectronics Laser, 2011, Vol. 22, No.1, pp. 144-148.(in Chinese)
[4] H. Guo, J.Y. Zhao, "Research on feature extraction for character recognition of NaXi pictograph", Journal of Computers, 2011, Vol. 6, No. 5, pp. 947-954. .
[5] H. Guo, J.Y. Zhao, "Segmentation Method for NaXi Pictograph Character Recognition", Journal of Convergence Information Technology, 2010, Vol.5, No.6, pp.87-98.
[6] H. Guo, J.Y. Zhao, M. J. Da, X.N. Li, "NaXi Pictographs Edge Detection Using Lifting Wavelet Transform", Journal of Convergence Information Technology, 2010,Vol.5, No.5, pp.203-210.
[7] Joseph Francis Charles Rock, A Nakhi-English encyclopedic dictionary, Rome :I.M.E.O, 1963.
[8] L. Zhai, Z.Y. Ding, Y. Jia,B. Zhou, "A word position-related LDA model", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 25, No. 6, 2011, pp. 909-925.
[9] G. Giorgi, and K. Yamashita, "Amphoteric behavior of Ge in GaAs: An LDA analysis", Modelling and Simulation in Materials Science and Engineering, 2011, Vol.19, No.3, pp. 035001-14.
[10] Y. Zhao, J. Wang, Q.J. Lu, R. Jiang, "Pattern recognition of eggshell crack using PCA and LDA", Innovative Food Science and Emerging Technologies, 2010, Vol.11, No.3, pp. 520-525.

Hai Guo born in 1979, received the M.Sc. degree (with distinction) in pattern recognition and intelligent system from the Kun Ming University of science and technology, China, in 2004. And now he works in the Department of Computer Science and Engineering of Dalian Nationalities University, Dalian, P.R.China as an associate professor. He has been interested in image processing, pattern recognition.. He has been in charge of the project of National Natural Science Foundation of China. He has published more than twenty research papers which are included in SCI, EI and so on.

Prof. Jinghua YIN born in 1957, received his doctor's degree in Harbin Institute of Technology (HIT) in 2001. And now she works in Harbin University of Science and Technology, Harbin, P.R.China as professor and Ph.D supervisor. She is an Academic Leader of the key discipline of Heilongjiang Province as well as Harbin University of Science and Technology, an executive director of Heilongjiang Physics Society, a Distinguished Teacher of China as well as Heilongjiang Province. She has been interested in Structure and Property of Inorganic Nanohybride Films, Structure and Failure of VDMOS Devices, Packaging Structure and Reliability of Electronic Devices. She has been in charge of the project of National 863 Research Project and Natural Science Foundation of China. She has published more than sixty research papers, twenty of which are included in SCI and EI.

Jing-ying Zhao received the M.Sc. degree (with distinction) in computer application from the Chang Chun University of science and technology, China, in 2003.She has been teaching at Dalian Nationalities University since 2003. For many years, she has done research in image processing, pattern recognition. He has published more than 20 papers.