Multi-Stage Based Feature Extraction Methods for Uyghur Handwriting Based Writer Identification

Kurban Ubul[1], Andy Adler[2] and Mamatjan Yasin[2]
[1]*School of Information Science and Engineering, Xinjiang University*
[2]*Department of Systems and Computer Engineering, Carleton University*
[1]*China*
[2]*Canada*

1. Introduction

Since starting of civilization era, it has been critical to the human society how to identify and verify the statuses of uncertain people. Consequently, personal identification is widely used in every aspect of society including governmental and commercial sections. However, traditional ways of personal identification, e.g., using identification cards or passwords, have their limitation and weakness that these surrogate representations of identity can easily be shared, lost, manipulated or stolen. Biometrics-based personal identification offers a natural and reliable solution to certain aspects of identity management by utilizing fully automated or semi-automated schemes to recognize individuals based on their characters (Jain et al., 2008). Biometric characteristics usually are physiological (e.g. face, fingerprint, palm print etc.), or behavioural (e.g. handwriting, gait, voice etc.). Among various kinds of biometrics, handwriting based personal recognition and verification have the advantage of easy to access, cheap, reliable (He et al., 2007), so it is widely used and welcomed by the public. As a result, writer identification is attractive enough to both industry and academia (He et al., 2007, 2008; Said et al., 2000; Schomaker & Bulacu, 2004; Srihari et al., 2002).

Writer identification is defined as a task of determining the author based on his/her handwriting from a set of writers (Plamondon & Lorette, 1989). According to the input method, there are generally two types: on-line and off-line writer identification. In on-line system, transducer equipment is connected to the computer that it can convert writing movement into a sequence of signals and then send the information to the computer, while handwriting materials are scanned into a computer in two dimensional image formats for processing in off-line system. The On-line system is get higher identification rate than the off-line system because extra features such as writing speed and pressure are extracted in on-line system. Therefore, off-line writer identification is more challenging task. Off-line research is further subdivided into text-dependent (or text-sensitive) and text-independent (or text-insensitive) approaches (Plamondon & Lorette, 1989; Said et al., 2000). Text-dependent methods refer to the study of one or a limited group of characters, so that they require the writers to write the same text. While text-independent approaches look at a feature set whose components describe global statistical features extracted from the entire image of a text. Text-dependent methods have a better performance on writer identification, but they are inapplicable in many practical applications because of their strict requirement on same the writing content. Off-line, text independent methods are studied in this paper.

1.1 Related work

A number of new approaches on off-line, text-independent writer identification have been proposed in recent years. Many researchers take the handwriting as an image containing some special texture, and they regard the writer identification as texture identification. Said et al. (2000) and Al-Dmemor (2007) treated the writer identification task as a texture analysis problem using multi-channel Gabor filtering and grey-scale co-occurrence matrix (GSCM) techniques. He et al. (2007, 2008) proposed wavelet based generalized Gaussian model (GGD) and hidden Markov tree (HMT) model in wavelet domain to replace the traditional 2-D Gabor filter. The approach which texture features extracted by Gabor and XGabor filters are combined with feature relation graph (FRG) and showed high efficiency for Persian writer identification (Helli & Moghaddam, 2010). Edge based directional probability distributions and connected component contours as features for the writer identification task are proposed (Schomaker & Bulacu, 2004). Li et al. (2009) proposed a text-independent method of writer identification based on grid-window microstructure feature for different multilingual handwritings. In order to reduce the dimensions of the features and to improve identification accuracy, feature selection is implemented by combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) (Fukunaga, 1990), and successfully used for text independent Chinese writer identification(Deng et al., 2008). The combined Gabor filter and Independent Component Analysis (ICA) method indicated high accuracy in texture segmentation and classification (Chen, Y & Wang, R, 2006, 2007). So these reports about writer identifications are mostly based on Latin handwriting (Plamondon et al., 1989; Said et al., 2000; Schomaker & Bulacu, 2004; Srihari et al., 2002), and Chinese handwriting (He et al. 2007, 2008; Li et al., 2009), Arabic handwritings (Al-Dmour & Zitar, 2007), even Persian handwritings (Helli & Moghaddam, 2010). However, there are only 4 reports about Uyghur handwriting based writer identification, in which two of them are our previous research (Ubul et al., 2008, 2009) indicated to using Gabor filter and Genetic algorithm (GA), and Gabor filter plus PCA and ICA methods for feature extraction, and get the 92.5% identification rate for 55 different people. Raxidin (2010) also used Gabor filter for feature extraction and achieved an accuracy rate of 79.8%. Li et al. (2009) used grid-window microstructure features for 120 different Uyghur persons handwritings, obtained 91.7% of identification rate. But the identification rate is still need to be improved comparing to other languages' identification. It is challenging task to find and develop methods suitable for Uyghur handwriting based writer identification. Therefore, there is still much research space for implementing existing algorithms or developing new effective algorithms and methods based on the nature of Uyghur handwriting.

1.2 Contribution

The approaches of writer identification methods are dependent on the languages, because letters in different languages have different patterns. In this chapter, we have proposed a method for texture feature extraction and selection by integrating Gabor filters, Genetic algorithm(GA), Principal component analysis (PCA), Kernel Principal component analysis (KPCA) and independent component analysis (ICA) for Uyghur handwriting based writer identification. Considering the diversity of Uygur handwriting, we conducted extensive handwriting data collection. The personal information of the selected people for data collection is different including their age, occupation and education level. A database of writer information was built based those information. Valid handwriting samples were scanned into computer to form a sample database of Uyghur handwriting. The handwriting images in the database were pre-processed based on the character of Uyghur handwriting. The Uyghur texture images were formed by specific pre-processing methods, in which some of the approaches were different from Latin and Chinese. Because the texture feature extraction method was used in this paper, and the style of texture image between different scripts (e.g. Latin, Chinese, and Arabic) is different, the characters of Uyghur handwriting, especially its stroke and local features were studied. Multi-channel Gabor filter suitable for the characters of Uyghur handwriting was designed. In order not to miss any feature of Uyghur handwriting image, 144 features were extracted. The high dimensionality features is computationally expensive, so some optimization (GA) and dimensionality reduction algorithms (PCA, ICA, KPCA) were used to find the best feature selection. Among these strategies, multi-stage based feature extraction and selection methods (such as Gabor + PCA + KPCA) were selected as the most appropriate for dimension reduction. In order to validate the performance of various feature extraction methods, this paper used in four classifiers (ED, WED, NN and KNN) to conduct experiments, and get 89.6% of accuracy for 65 different people. The experimental results shows the effectiveness of the proposed method which to extract more features using Gabor filters and reduce the dimensionality of the features using multi-stage based feature extraction and selection methods.

2. Data acquisition and pre-processing

The common steps of writer identification include data acquisition, pre-processing, feature extraction, and classification. The data acquisition is the first phase for writer identification and verification systems. Subjects are asked to write their handwritings on a paper with their natural writing style. Because of the handwriting images contaminated by noise and letters with different sizes, the efficient features are extracted only after they are pre-processed. Different methods and algorithms (Said et al. 2000; Schomaker & Bulacu, 2004) are used for pre-processing and feature extraction phases based on the handwriting styles. The handwriting is taken as an image containing some special texture, and writer identification is regarded as texture identification here. Uyghur handwriting texture (Fig. 1: e) has its own characters compare to other languages such as Latin, Chinese and Arabic (Fig. 1: b, c, d) so improved methods are implemented in pre-processing and feature extraction steps. The nature of Uyghur handwriting is indicated before description of data acquisition.

2.1 The nature of Uyghur handwriting text

The Uyghur are a Turkic-speaking ethnic group inhabiting Eastern and Central Asia. Today, Uyghurs mainly live in the Xinjiang Uyghur Autonomous Region (hereafter: Xinjiang) in China. Arabic based Uyghur script is an official writing system in Xinjiang, while Cyrillicbased Uyghur script is still used by Uyghurs in former Soviet Union Republics and Latin-based Uyghur script are also in use[1]. The handwriting of Arabic-based Uyghur script (hereafter: Uyghur) used widely in Xinjiang area is studied in this paper.

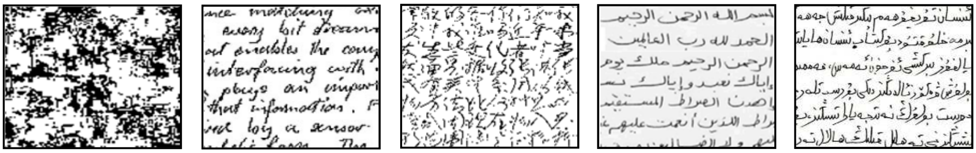

(a) Random texture (b) Latin (c) Chinese (d) Arabic (e) Uyghur

Fig. 1. Random texture image (a) and texture images of different handwriting

Uyghur character is composed of 32 letters including 8 vowel letters and 24 consonant letters, besides 4 kinds of different forms for each character. Thus, 32 letters become more than 120 character styles.

1. The writing direction of Uyghur character is from right to left, from left down right for the line progression. There are 4 different writing forms for Uyghur letter: (i) “initial form”: only the suffix is connected with the next character, (ii) “intermediate form”: initial and suffix are connected with adjacent letter, (iii) “final form”: only the initial is connected with the above letter, and (iv) “isolated form”: initial and suffix are not connected with adjacent letter.
2. The vocabulary of Uyghur character is composed of one or several letters. According to rules of writing, these letters will form one or several letter passages by initial and suffix connections. For a block letter or handwriting, the letters are connected along a certain level, which is called base line.
3. Unequal width of letters. This phenomenon happens not only on different letters, but also on the 4 different forms of certain letters. Furthermore, a straight line will be adopted to fill in the spaces among the letter to let a line of text distribute uniformly.
4. The vocabulary of Uyghur character is composed of syllable, which is generally constructed from the combination of vowel and consonant, where vowel is the centre. It is definite that the composition of syllable and vocabulary is regular. There is a blank space between two vocabularies (Fig. 2).

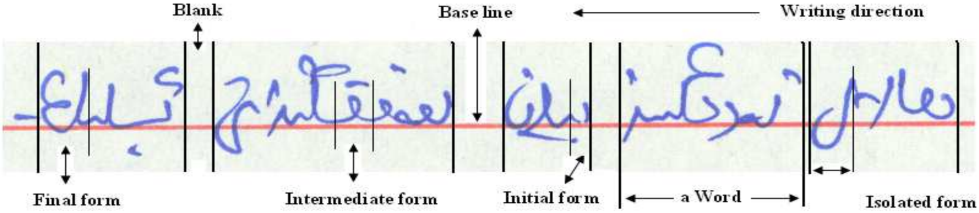
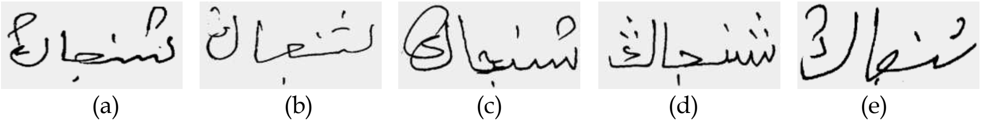


Fig. 2. An example of Uyghur Handwriting

[1] Uyghur alphabet. See http://en.wikipedia.org/wiki/Uyghur_alphabet

5. The stroke of Uyghur character is not fixed. The numbers of strokes for the same Uyghur word are different from person to person. Especially, they are different with position, size, longitude, slant angle and structure. Fig. 4 shows the same word written by different person.

In Fig. 3: (a) to (e), the strokes of the first letter"ﺷ" are consisted of 4, 3, 4, 6, 2 strokes respectively. The situation becomes more complicated if a word or a sentence is concerned without misspelling. Therefore, the local characters of Uyghur handwriting further increased the difficulty level for Uyghur handwriting based writer identification.



(a) (b) (c) (d) (e)

Fig. 3. The Uyghur word "Xinjiang" (in Uyghur - شىنجاڭ) written by 5 different people

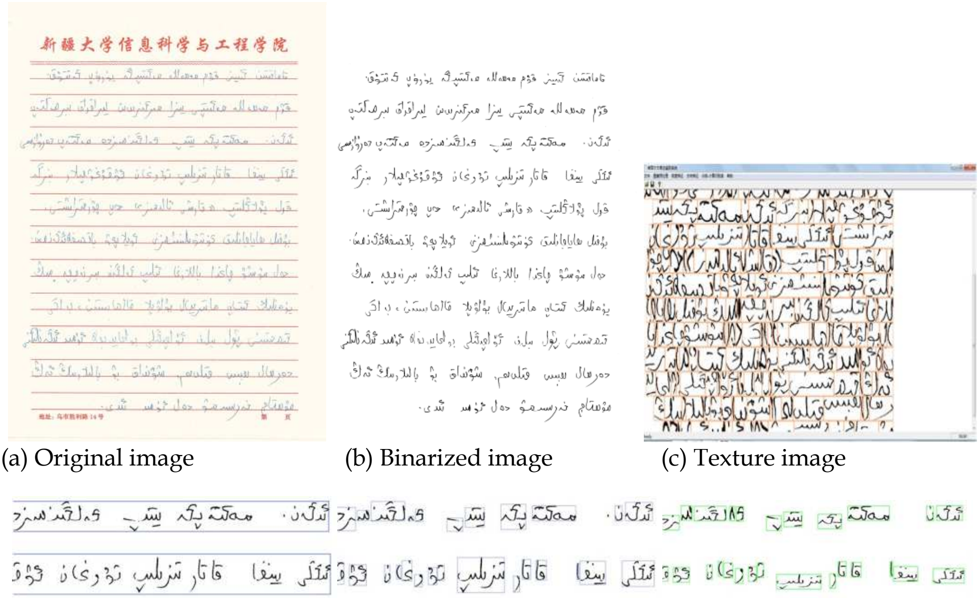
2.2 Data acquisition

We selected 353 Uyghur participants for data collection and asked each participant to write anything to fill 3 sheets of letter size paper. They were instructed to write in his/her natural handwriting just by keeping enough spaces between words and lines. They wrote in black or blue ink on the grid paper with red or green lines. To increase the diversity of Uyghur handwriting, the participants are selected respect of their age, gender, education, etc. One part of them is from elementary school to graduate students, while the remainder are adults of various professions. Among them, the oldest person was 80 years old and the youngest was 9. Databases of participants were established including all data collector's personal information (such as age, gender, education, job, etc.). The handwritten documents are digitally scanned using a HP scanner with a resolution of 300 dpi and saved in bmp image format. Thus, the sample database of Uyghur handwriting was set up.

2.3 Pre-processing

The input image contains noise and Uyghur characters of different sizes and spaces between text lines, so original image should be pre-processed before feature extraction. Based on the nature of Uyghur characters similar to Arabic and Persian, the minimum unit to be selected in pre-processing is connected components which are consist of one or several letters unlike Latin and Chinese. Inspired by the work of (He et al., 2008; Schomaker & Bulacu, 2004), we propose a procedure of automatic pre-processing as described below:

1. Removing the background and binarization. Collected handwriting samples were written on a plain graph paper. Although the grid of the graph paper does not affect the writing style, it will affect on accessing handwriting information. The selected cell lines of graph paper are in red colour, while the text colour is black or blue. Therefore, the method to remove the background in this paper is mainly the histograms of the red and green by setting the higher pixel component of the red and green to white, and others to black. It is not only to remove the background, but also to get binarized image as indicated in Fig. 4 (b).



(a) Original image (b) Binarized image (c) Texture image

(d) Locating the line (e) Connected component (f) Normalized image

Fig. 4. The steps of Uyghur handwriting image preprocessing

2. The binarized image contains a number of discrete noises. To avoid affecting the feature extraction, the discrete noise is removed. According to the real situation in handwriting samples, the size of the discrete noise threshold is set to 10. If the observation points related to the number of black spots are less than 10, it is considered as noise points and they are filled with white points.
3. Locating the line. Uyghur character has distinctive characteristics of connected letters with different width. A large number of additional components are exist in Uyghur text, so the blank space can be the space between two lines, or the space between additional part and the main part of a character in a line. Thus, a threshold has to be set, where the blank space exceeded the threshold value is the gap between text lines, or it is the gap within a line. Thus, the contour of the text image can be obtained from a text image. Writing along the baseline is a major feature of Uyghur, these characteristics in a line of text images are expressed with pixels concentrated around a particular horizontal line as well as on the baseline domain. When we take the horizontal projection along contour of text line images, two maximum values within the horizontal direction are set as upper and lower boundary in the text baseline domain and boundary of each line can be segmented. (Fig. 4(d))
4. Separating the connected components. After locating each line in Uyghur handwriting, directional statistics was performed for black pixels on each text line where the connected component is the place with small number of black pixels. The distance of blank space for text line images can be obtained through Vertical Projection Profile. From the text line, independent form of letters or their connection is extracted and be indicated as shown in Fig. 4(e). The connected component between letters is linked with relatively flat straight lines.

5. Normalization. The character normalization in writer identification is required for the tilted character, position of stroke and orientation to be stable. Unlike other languages, Uyghur text has significantly connected letters that each word in Uyghur writing is interconnected, so it is enough for adjusting the vertical height of the character. (Fig. 4(f))
6. Making texture image and dividing. After removing the spaces between connected components and lines in normalized image, the handwriting texture image in (Fig. 4(c)) is obtained. The size of texture image here is selected to be 1024 × 1024 pixels. In order to ensure the standard and accuracy of texture feature extraction, we divided the normalized image (size of 1024 × 1024 pixels) into 16 sub-images with the size of not more than 256 × 256 pixels. An example of image pre-processing is shown in Fig. 4.

3. Feature extraction

Multi-channel Gabor wavelet technique is becoming very popular in texture analysis, and has been successfully applied to a broad range of image processing tasks (Jain, & Farrokhnia, 1991). In this paper, we have proposed a method for texture feature extraction and selection by integrating Gabor filters, Genetic algorithm, Principal component analysis, Kernel Principal Component analysis and Independent Component analysis for Uyghur handwriting based writer identification.

3.1 Gabor filter

The two-dimensional Gabor filter (Jain & Farrokhnia, 1991; Plamondon et al., 1989) can be mathematically expressed as:

$$g(x, y) = \frac{1}{2\pi\sigma^2} \exp[-(\frac{x^2+y^2}{2\sigma^2})] \tag{1}$$

We can model each cortical channel by a pair of Gabor filters h_e (x, y) and h_o (x, y) as follows:

$$\begin{cases} h_e(x, y) = g(x, y)\cos[2\pi f(x\cos\theta + y\sin\theta)] \\ h_o(x, y) = g(x, y)\sin[2\pi f(x\cos\theta + y\sin\theta)] \end{cases} \tag{2}$$

where f and θ are the spatial frequency and the orientation of the Gabor envelope, and $h_e(x, y)$ and $h_o(x, y)$ denote the even and odd symmetrical Gabor filters respectively.

Texture feature extraction requires both radial frequency and orientation. Tan (1992) showed that, for any image of size N×N (where N is a power of 2) the important frequency components are within $f \leq N/4$ (cycles/degree). For Uyghur handwriting, N is set to 256, and so frequencies of 2, 4, 8, 16, 32, and 64 cycles/ degree are used. For each central frequency f, filtering is performed at values of 0°, 15°, 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°, 150°, and 165°, which gives a total of 72 output images (12 for each frequency, as shown in Fig. 5). The mean and standard deviation of the output images are chosen to represent texture features. In this way, a total of 144 features are extracted from a given image. They form a 144 dimensional feature vector which is reduced using the genetic algorithm.

3.2 Two-stage based feature extraction and selection

The two stage-based feature extraction and selection methods in here are indicated to take high dimensional feature vectors based on the character of Uyghur handwriting image first, and then effective features are select using some algorithms such as GA and PCA.

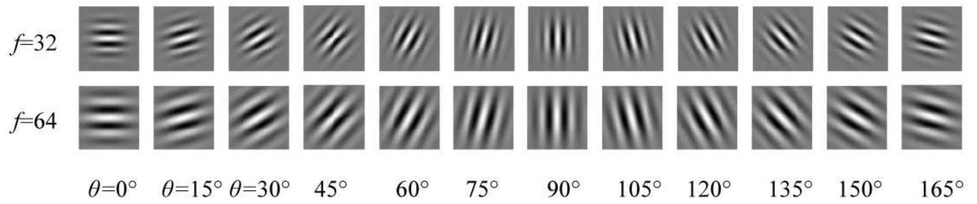


Fig. 5. Real part of Gabor filter with different frequencies (f) and orientations (θ)

3.2.1 Feature extraction using Gabor filter and Genetic algorithm

Genetic algorithms (GA), which are algorithmic models based on a Darwinian-type survival-of-the fittest strategy with sexual reproduction, were firstly introduced by John Holland in 1960s (Holland, 1992). Genetic algorithm is a kind of computerized procedure to search from a group of random initial solutions which are so called population (Goldberg, 1989). Each individual in the population is a possible solution of problem which is so called chromosome (or genes). It is represented in binary as strings of 0s and 1s. These chromosomes are evolved continuously in follow-up iteration, this is called genetic. Fitness is measurement of chromosomes' fair or foul in each generation, and new population created are called offspring which is born through crossover or mutation of previous chromosome. In a new population, the size of population is maintained as a constant based on the fitness selecting parts of offspring and weeding other parts of offspring out. The fitness value of individual with higher probability is selected. Thus, the algorithm converges to the best chromosome through evolving several generations, and it may be the optimum solution of the problem. The feature vectors extracted from Uyghur texture are reduced using the GA and feature selection algorithm (Siedlecki & Sklansky, 1989), as shown Fig. 6.

The input argument of the population is a vector of row indices from the training data. The fitness is a linear combination of the error rate and the posterior probability of the classifier:

$$q(f_j) = 100 \times e(f_j) + 1 - \frac{\sum_{i=1}^{M} r(f_j)}{M} \quad (3)$$

where M is the number of individual features f_j, $e(f_j)$ is the classification error rate, $r(f_j)$ are the maximum values along the columns of $u(f_j)$, and $u(f_j)$ is a matrix containing posterior probabilities that the *k*th training class is the source of the *i*th sample feature. The critical value of fitness $q\ (f_j^*)$ is set to 2.0×10^{-7} here.

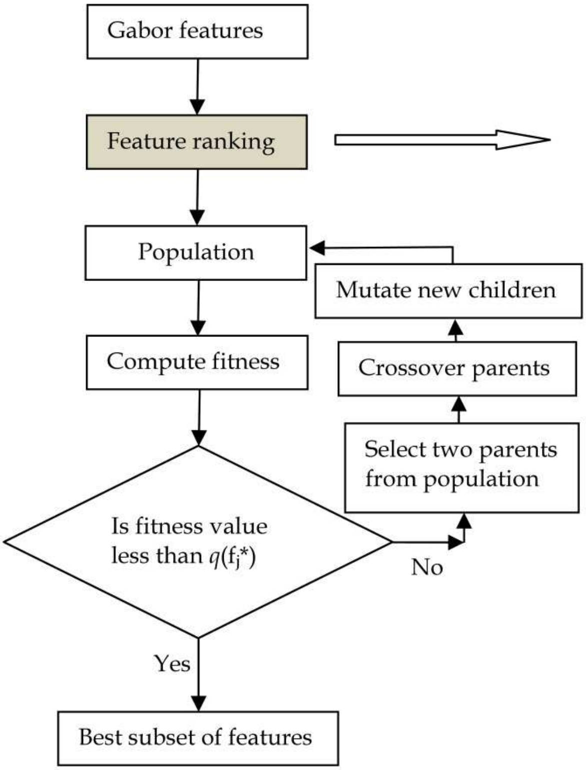
3.2.2 Feature extraction using Gabor filter and principal component analysis

Principal component analysis (PCA) is known as Karhunen-Loeve transform. The objective of the study is to perform appropriate linear combination of multi-dimensional data and orthogonal transformation (Dunteman, 1989). By controlling the mean square error in data, dimension reduction and compression are perform for high-dimensional linear space.

The high dimensional vector should be reduced, because Gabor function is non-orthogonal, and redundant information is present in the filtered image (Dunn, 1995), and from the calculation point of view, it is necessary to reduce the dimension of feature vector for classification (Chen & Wang, 2007). From above Gabor filter, we get a high-dimensional Gabor feature vector $F \in R^k$, where k is the dimension of feature vector:

$$k = f \cdot \theta \cdot \mu^2 \qquad (4)$$

where *f* and *θ* are the centre frequency of Gabor filter and its orientation respectively, and *u* is the size of filter window.



```
Input: Training samples
    $t = (t_1, t_2, \dots, t_M)$ with known
    class labels $C(c_1, c_2, \dots, c_N)$.
Initialize: Subset of survived
    features $f = [1, 2, \dots, m]$
  • List of ranked feature $l = [\ ]$
Repeat until $f = [\ ]$
  • Restrict training samples to
  good feature indexes $T = T_0(:, f)$
  • Train the SVM classifier
    $p = SVM\text{-}train\ (T, c)$
  • Count the weight vector of
  dimension length $v = \sum p_i c_i t_i$
  • Count the ranking criteria
    $k_i = (v_i)^2$
  • Find the feature with smallest
    ranking criterion $q = arg\ min(k)$
  • Update feature ranked list
    $l = [f(q), l]$
  • Eliminate the feature with
    smallest ranking criterion
    $f = f(1{:}\ q-1, q+1{:}\ \text{length})$
Output: Ranked feature list $l$
```

Fig. 6. The flow chart of selecting best features with Genetic algorithm

PCA is a classical dimension reduction method which has the important feature of using low-dimensional subspace to represent the original high-dimensional data based on the minimum mean square error (MSE). To perform PCA dimensionality reduction for high-dimensional Gabor feature vector F:

$$X = PF \qquad (5)$$

where $P = (p_1, p_2, \dots, p_k)^T \in R^{m \times k}$, $m < k$ is the characteristic matrix, output $X \in R^m$ is low-dimensional feature vector.

It can be seen from the previous procedure that PCA is a linear mapping algorithm, but PCA can only be used to remove the correlation between features (ignoring the non-linear correlation between the features), and it is not obtained the independent components of the feature, and it is a signal analysis method based on second-order statistical properties, cannot use the data in higher order statistics, so the transformed data may still exist between the higher order redundant information.

3.3 Three-stage based feature extraction and selection

Since PCA has some limitation mentioned above, it is necessary to use the combinations of PCA and ICA methods, to implement dimension reduction for high-dimensional feature vector, to remove redundant information in filtered images, and to obtain higher-order statistical characteristics of the handwriting texture in order to make better use of handwriting texture classification.

3.3.1 Feature extraction using Gabor filter plus PCA and ICA

Independent Component Analysis (ICA) is a higher-order statistics-based method of data analysis related to Blind Signal Separation (BSS) (Comon, 1994; Hyvärinen & Oja, 1997). BSS problem assumes that a linear combination of N independent targets (random variable) produces M observed variables, and the purpose of ICA is to identify the mixing matrix from the observed variables. When we perform texture analysis using ICA, the given texture image is considered to be a series of unknown statistical mixture of independent random variables. Jenssen & Eltoft (2003) proposed the concept of ICA filter bank for texture image segmentation and suggested that the performance of segmenting multi-texture image is close to or better than Gabor filters. The paper used ICA techniques for obtaining independent components of high-dimensional feature vector in features, and established an independent Gabor features used for handwriting classification.

We made the output of PCA, $X \in R^m$ as an observed vector for ICA analysis and assume that X is a linear combination of n unknown independent components $S = [s_1, s_2, ..., s_n]$, and then the linear relationship can be written as formula form as the model basic of ICA :

$$X=AS \tag{6}$$

where A is m × n unknown mixing matrix of full-rank, S is approximate independent component. The objective of ICA is to find a separation matrix W, making the output:

$$Y=WX \tag{7}$$

When the separation matrix W is the inverse of mixing matrix A, the independent component S can be accurately extracted, or it needs to sort and change the magnitude (Chen & Wang, 2006, 2007). Calculated Y is the last feature vectors used for texture classification. Hyvärinen (1997) proposed more popular matrix separation method according to very small mutual information and equivalence of bigger negative entropy, introduced non-linear monotonic estimated negative entropy function. This algorithm is used in this paper to get independent Gabor features, the formula is as follows:

$$\begin{cases} \vec{W}^{+} = \vec{W} - \dfrac{\mu\left[\vec{C}^{-1}E\{\vec{X}g(\vec{W}^{T}\vec{X})\} - \beta\vec{W}\right]}{E\{g'(\vec{W}^{T}\vec{X})\} - \beta} \\ \vec{W}^{*} = \dfrac{\vec{W}^{+}}{\sqrt{(\vec{W}^{+})^{T}\vec{C}\vec{W}^{+}}} \end{cases} \tag{8}$$

where W^+ and W^* respectively indicate the current and new value obtained by iteration. μ is the step size with initial value of 1, it decreases rapidly with the increased number of iterations, $\beta=E\{W^TXg(W^TX)\}$ is used for normalization to improve the robustness of the algorithm, g is the contrast function, the covariance matrix can be obtained for the observation vector X as $G(u) = u^4 / 4$, $g(u) = u^3.c = E\{XX^T\}$.

3.3.2 Feature extraction using Gabor filter plus PCA and KPCA

Kernel Principal Component Analysis (KPCA) algorithm uses kernel function to obtain the arbitrary high-order correlation between input variants, and the principal components needed through the inner production between input data. After SchÄolkopf et al. (1998) applied Kernel component analysis in feature extraction, it became more popular in image vision, content based retrieval and text classification fields. The PCA outputs are used as the selection vector of KPCA, assumed that the features of mapping is centralized as shown:

$$\sum_{i=1}^{N} \xi(x_i) = 0 \tag{9}$$

where ξ is a non-linear mapping, N is the total entries of mapping feature. After mapping, the features of the covariance matrix C such that:

$$C = \frac{1}{N}\sum_{i=1}^{N} \xi(x_i)\,\xi(x_i)^T \tag{10}$$

The characteristic equation can be found from

$$hV=CV \tag{11}$$

According to the theory of reproducing kernel, feature vector V must be in the space domain, $\{\xi(x_i),\dots,\xi(x_N)\}$ which is

$$V = \sum_{i=1}^{N} \alpha_i\,\xi(x_i) \tag{12}$$

Defining an N × N matrix K

$$K_{ij} = k(x_i, x_i) = \xi(x_i)^T\xi(x_i) \tag{13}$$

K is called the nuclear matrix. The substitution of (9), (10) and (11) into (13) produces:

$$K\alpha=Nh\alpha \tag{14}$$

By solving (11), the feature vectors V are transformed into the feature vector α of the characteristic equation (14). Based on the above definition, the kernel matrix K is symmetric, positive semi-definite matrix, and its eigenvalues are non-negative according to matrix theory. By solving (14), a group of non-zero eigenvalue h_j and the corresponding α^j (j = 1, 2, ..., p) are obtained that satisfy the normalization condition (15)

$$h_j\left(\alpha^j,\alpha^j\right)=1 \tag{15}$$

According to (11), the principal components V_j (j = 1, ..., p) are obtained from the projection of feature space. Let x to be test samples after PCA processing, its projection is V_j

$$\left(V^j\right)^T\xi(x) = \sum_{i=1}^{N} \alpha_i^j\,\xi(x_i)\xi(x) = \sum_{i=1}^{N} \alpha_i^j\,k(x_i, x) \tag{16}$$

Neither (13) nor (16) requires the $\xi(x_i)$ in explicit form, and they are only needed in dot products. Therefore, we are able to use kernel functions for computing these dot products without actually performing the map ξ (SchÄolkopf & Smola, 2001). The dot products are the features selected from KPCA.

4. Writer identification

The classification experiments using selected classifier are processed after extracting and selecting features. This chapter will describe the classifiers to be used and their experimental results based on the size of our database and feature extraction methods.

4.1 Classification

In theory, any kind of classifier can be applicable during the identification phase. Considering the smaller scale of sample data for Uyghur, we use the Euclidean distance classifier, the weighted Euclidean distance classifier, the nearest classifier and K-Nearest Neighbour classifier (Plamondon & Lorette, 1989). Definition of them as following:

1. Euclidean Distance (ED) classifier. Commonly used Euclidean distance is:

$$d_E(u,v) = [(u-v)^T(u-v)]^{\frac{1}{2}} = [\sum_{i=1}^{n}(u_i - v_i)^2]^{1/2} \tag{17}$$

where u_i and v_i are feature vectors and i=1,2,…n.

2. Weighted Euclidean Distance (WED) classifier. Unknown handwriting feature vectors are compared with the trained known samples of handwriting. When the weighted Euclidean distance between the feature vector and k class sample is smallest, the input handwriting is classified as k-class handwriting. Weighted Euclidean distance is calculated as:

$$\mathrm{WED}(k) = \sum_{i=1}^{M} \frac{\left(f_l - f_l^{(k)}\right)^2}{\left(\delta_l^{(k)}\right)^2} \tag{18}$$

where f_l is the lth feature of unknown sample, $f_l^{(k)}$ and $\delta_l^{(k)}$ are respectively the sample mean and sample standard deviation of lth feature of write k respectively.

3. Nearest Neighbour (NN) classifier. It classifies the handwriting by decision-making rule. For the nearest neighbour decision-making, assume there are c categories ω1, ω2, …, ωc for pattern identification, each class indicates Ni samples to be classified, i=1,2,…,c. The discriminate function for the class can be written as:

$$g_i(x) = \min_k \|x - x_i^k\|,\ k = 1,2,\dots,N_i \tag{19}$$

where the subscript i of x_i^k indicates ω_i class, k indicates ω_i class. According to above equation, the decision-making rules can be written as:

$$g_{j(x)} = \min_i g_i\,(x),\ \ i = 1,2,\dots,c \tag{20}$$

where decision maker $x \in \omega_j$.

4. K-Nearest (K-NN) Neighbour classifier. When we use K-NN classifier, the ideal characteristics of each trained class K is f_k, the detected characteristics of unknown writer is U. To determine class K of the writer, the similarity of each class is measured by calculating the distance between feature vectors f_k and U. The taken Euclidean distance, which is the distance d_k among the class for unknown writer is:

$$d_k = \left[\sum_{j=1}^{N}(U_j - f_{kj})^2\right]^{1/2} \tag{21}$$

The writer assigned to the class R such that:

$$d_R = \min(d_k) \tag{22}$$

where *k* is the class number, and k=1,2,..., *N*. We took k=5 in our experiment.

4.2 Experimental results

We selected 65 Uyghur participants' handwritings from our database. In our experiment, we select size as1024 ×1024 pixels. In order to create a number of small images which belong to the same class, each of the1024×1024 images created was divided into non-overlapping 256×256 sub-images, thus forming 16 sub-images from each image. 12 sub-images from the first document were used for the training data set, and 9 sub-images from the second document were used as the testing data set. Four classifiers (ED, WED, NN, and KNN) were used each set of experiments. We performed three kinds of experiments:

1. Result with Gabor filtering. Firstly, features were extracted with frequencies of 8, 16, 32, and 64 cycles /degree. For each central frequency f, filtering was performed at 0°, 45°, 90°, and 165°. Thus a 32 dimensional feature vector was obtained. The results show poor classification rates that indicated as a table 1. Secondly, extracted 144 features as explained in section 3.1, achieved higher identification rates than mentioned above. NN classifier got a top-10 classification rate of 81.1%. Others are illustrated in table 1.

Classifier	Identification accuracy using two types of feature vectors					
	Extractd 32 dimensional features			Extractd 144 dimensional features		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
ED	41.5%	47.1%	51.4%	49.8%	52.3%	54.7%
WED	42.6%	49.9%	58.7%	54.0%	61.5%	66.8%
NN	56.3%	64.2%	71.0%	67.4%	76.9%	81.1%
K-NN	56.8%	63.9%	70.2%	68.7%	75.2%	79.5%

Table 1. Identification accuracy using Gabor features

2. Result using two-stage based feature extraction methods. The 144 dimensional Gabor feature vectors were reduced to 48 and 55 by selecting the appropriate features using GA feature selection model and PCA respectively. The identification rates is higher than the method of using Gabor features directly, the best Top-1 identification accuracy is 78.2% and the Top-10 accuracy reaches 86.9% with K-NN classifier, as shoved in table 2.

Classifier	Identification accuracy using two-stage based feature extraction methods					
	Gabor+GA			Gabor+PCA		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
ED	57.6%	59.5%	62.3%	51.6%	60.7%	69.4%
WED	58.9%	60.8%	61.5%	63.8%	73.0%	76.3%
NN	77.4%	80.1%	82.6%	71.7%	76.5%	83.2%
K-NN	78.2%	81.4%	83.7%	76.2%	82.8%	86.9%

Table 2. Identification accuracy using three-stage based feature extraction methods

3. Result from three-stage based feature extraction methods. The 144 dimensional Gabor feature vectors were reduced to 42 by selecting the most appropriate features using a Gabor plus PCA & ICA method or Gabor plus PCA & KPCA method. We obtain a best Top-1 identification accuracy is 89.6% and the Top-10 accuracy reaches 91.8% with K-NN that is higher than other's in the paper as indicated table 3.

Classifier	Identification accuracy using three-stage based feature extraction methods					
	Gabor+PCA+ICA			Gabor+PCA+KPCA		
	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
ED	57.7%	63.9%	69.3%	61.0%	64.3%	70.2%
WED	70.1%	73.0%	79.6%	72.5%	75.4%	78.9%
NN	79.8%	82.2%	84.5%	83.2%	86.7%	88.1%
K-NN	87.5%	88.3%	90.0%	89.6%	90.2%	91.8%

Table 3. Identification accuracy using three-stage based feature extraction methods

In order to see the efficiency of the method, it is compared with methods used for Uyghur handwriting based writer identification and related methods used for different languages indicated as following Table 4. and Table 5.

Authers (year)	Used methods	Experimented person No.	Identification rate
Ubul et al. (2008)	Gabor + GA	23 person	88.0%
Ubul et al. (2009)	Gabor+PCA+ICA	55 person	92.5%
Li et al. (2009)	Microstructure feature	120 person	91.7%
Abdiryim (2010)	Gabor	17 person	79.8%
This paper	Gabor+PCA+KPCA	65 person	89.6%

Table 4. Identification accuracy comparing with the methods used for Uyhgur handwriting

Authers (year)	Used feature extraction methods and identification rates (person)				language
	Gabor	Gabor+GA	PCA+LDA	Gabor+PCA	
Said et al. (2000)	96% (40)				Latin
Deng et al. (2008)			88.41% (138)		Chinese
Al-Demor et al. (2007)	85% (22)	90% (22)			Arabic
This paper	81.1% (65)	83.7% (65)		86.9% (65)	Uyghur

Table 5. Comparing with the related methods used for other languages

Among various kinds of approaches used for Uyghur handwriting based writer identification as indicated Table 4, the Microstructure feature based identification method is indicated its high efficiency (91.7% of identification rate), but the three stage based feature extraction approach (Gabor+PCA+KPCA) proposed in this paper still showed higher recognition rates (89.6% of identification rate) than others. It is obvious in Table 5 that the Gabor based feature extraction method is more suitable for Latin (96% of identification rate) than Uyghur (81.1% of identification rate), but its identification rates can be increased by using Gabor plus GA method (83.7% of identification rate) and Gabor plus PCA (86.9% of identification rate) method.

It can be seen from the experimental result that to extract high dimensional feature vectors in texture image are play vital role to increase the identification rates, but needs to large amount of computation. The purposed method which to extract more features and reduce them to with feature selection algorithms is achieved better identification rates than traditional Gabor based method for Uyghur handwriting.

5. Conclusion and future work

Writer identification is a popular research field in many languages such as English, Chinese, Arabic, Uyghur, etc. The approaches of writer identification methods are dependent on the languages, because letters in different languages have different patterns. In this chapter, we have proposed a method for texture feature extraction and selection by integrating Gabor filters, Genetic algorithm(GA), Principal component analysis (PCA), Kernel Principal component analysis (KPCA) and independent component analysis (ICA) for Uyghur handwriting based writer identification. The texture image is firstly formed by connecting components via projection profile on the basis of Uyghur handwriting's nature. It is filtered by a given bank of Gabor filters, and then higher dimensional feature vectors are constructed from the filtered texture images. Next, the dimensionality of these vectors is reduced by means of GA, PCA and KPCA. Finally, the independent components in the resulting vectors with reduced dimensionality are analyzed and extracted. Four classification techniques are used: Euclidean Distance Classifier (ED), Weighted Euclidean Distance Classifier (WED), Nearest Neighbour Classifier (NN), and K-Nearest Neighbour (K-NN) Classifier. Experiments are performed using Uyghur handwriting samples from 65 different people and very promising results of 89.6% correct identification are achieved.

In the future work, the number of samples in Uyghur handwriting database will be further expanded. For the texture feature extraction, on one hand, further research will be carried on the local features of Uyghur handwriting. On the other hand, other effective methods, such as GLCM, X-Gabor, auto-correlation function, the wavelet transform etc will be validated to use the global feature extraction used for the Uyghur texture feature extraction. In addition, combination of the characteristics of Uyghur handwriting with the feature optimization and dimensionality reduction algorithms further improve the performance. About the classifiers, we will try to use other classifiers such as Support Vector Machine (SVM) classifier, artificial neural network classifier (ANN), Linear Discriminate Classifier (LDC), Bayesian classifier, Multi-class classifier or the combinations of classifiers and so on.

6. Acknowledgment

This work is sponsored by the National Natural Science Foundation of China (No. 61163028), the Minorities Special Training Plan Project of Science and Technology Bureau Xinjiang Uyghur Autonomous Region (No. 201023116) and the Open Project of Xinjiang Laboratory of Multi-language Information Technology (No. 049807).

7. References

Al-Dmour, A. & Zitar, R. A. (2007). Arabic Writer Identification based on Hybrid Spectral-statistical Measures. *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 19, No. 4, (December 2007), pp. 307–332, ISSN 1362–3079

Chen, Y & Wang, R. (2006). Texture segmentation using independent component analysis of Gabor features. *Proceeding of the ICPR 18th International Conference on Pattern Recognition*, Vol. 2, pp. 147-150, ISBN 0-7695-2521-0, Hong Kong, China, August 20-24, 2006

Chen, Y. & Wang, R. (2007). A method for texture classification by integrating Gabor filters and ICA. *Journal of Acta Electronica Sinca*, (in Chinese), Vol. 35, No. 2, (February 2007), pp. 299-303, ISSN 0732-2112

Comon, P. (1994). Independent component analysis, a new concept?. *Signal Processing*, Vol. 36, No. 3, (April 1994), pp. 287–314, ISSN 0165-1684

Deng, W.; Chen, Q & Yan,Y. et al. (2008). Off- line Chinese writer identification based on character-level decision combination. *Procedding of the International Symposiums on Information Processing,* pp. 762 – 765, ISBN 978-0-7695-3151-9, Moscow, Russia, May 23-25, 2008

Dunn, D. & Higgins. W. E. (1995). Optimal Gabor filters for texture segmentation. *IEEE Transactions on Image Processing*, Vol. 4, No. 7, (July 1995), pp. 947 - 964. ISSN 1057-7149

Dunteman, G. H. (1989). *Pricipal Component Analysis*. Sage Publications Ltd. ISBN 978-0-803-93104-6, London, United Kingdom

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (second edition). Academic Press, ISBN 0-32-269853-7, New York, United States

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning, Reading*. Addison-Wesley, ISBN 0201157675, Boston, Massachusetts, United States

He, Z. Y.; You, X. & Tang, Y. Y. (2007). Writer identification of Chinese handwriting documents using hidden Markov tree model. *Pattern Recognition*, Vol. 41, No. 4, (April 2007), pp. 1295 – 1307, ISSN 0031-3203 41

He, Z. Y.; You, X. & Tang, Y. Y. (2008). Writer identification using global wavelet-based features. *Neurocomputing*, Vol. 71, No. 4, (February 2008), pp. 1832–1841, ISSN 0925-2312

Helli, B & Moghaddam, M. E. (2010). A text-independent Persian writer identification based on feature relation graph (FRG). Pattern Recognition, Vol. 43, No. 6 (June 2010), pp. 2199–2209, ISSN 0031-3203

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems* (second edition), MIT press, ISBN 978-0-262-58111-0, Cambridge, Massachusetts , United States

Hyvärinen, A. & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, Vol. 9, No. 7, (September 1997), pp. 1483-1492, ISSN 0899-7667

Jain, A. K. & Farrokhnia, F. (1991). Unsupervised texture segmentation using Gabor filters, *Pattern Recognition*, vol. 24, No. 12 (May 1991), pp. 1167-1186, ISSN 0031-3203

Jain, A. K.; Flynn, P. & Ross, A. (2008). *Handbook of Biometrics*. Springer, ISBN 978-0-387-71040-2, New York, United States

Jenssen, R. & Eltoft, T. (2003). ICA filter bank for segmentation of textured images. *Proceeding of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 827-832, ISBN 4-9901531-0-3, Nara, Japan, April 1-3, 2003

Li, X.; Ding, X. & Peng, L. (2009). A microstructure feature based text-independent method of writer identification for multilingual handwritings. *Acta Automatica Sinica* (in *Chinese)*, Vol. 35, No. 9 (September 2009), pp. 1199-1208, ISSN 1874-1029

Plamondon, R. & Lorette, G. (1989). Automatic Signature Verification and Writer Identification – the State of the Art. *Pattern Recognition*, Vol. 22, No. 2, (January 1989), pp. 107–131, ISSN 0031-3203

Raxidin, A. (2010). Study on Gabor wavelet based feature extraction method for Uyghur handwriting. Journal of Hotan Teachers College, Vol. 25, No. 5, (October 2010), pp. 184-185, ISSN 1671-0908

Said, H. E. S.; Tan, T. N. & Baker, K. D. (2000). Personal Identification based on Handwriting. *Pattern Recognition*, Vol. 33, No. 1, (January 2000), pp. 149- 160, ISSN 0031-3203

SchÄolkopf, B. Smola, A. & Mäuller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, Vol.10, No.5, (July 1998), pp.1299-1310, ISSN 0899-7667

SchÄolkopf, B. & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, ISBN 0262194759, Cambridge, Massachusetts, United States

Schomaker, L. & Bulacu, M. (2004). Writer identification using connected component contours and edge-based featurs of uppercase western script. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 6, (June 2004), pp. 787–798, ISSN 0162-8828

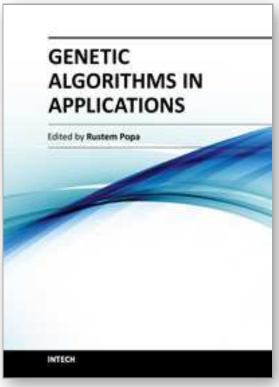
Siedlecki, W. & Sklansky, J. (1989). A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, Vol. 10, No. 5, (November 1989), pp. 335–347, ISSN 0167-8655

Srihari, S.; Cha, S.; Arora, H. & Lee, S. (2002). Individuality of handwriting. *Journal of Forensic Science*. Vol. 47, No. 4, (July 2002), pp. 1–17. ISSN 0022-1198

Tan, T. N. (1992). Texture feature extraction via visual cortical channel modeling. *Proceeding of the ICPR 11th International Conference on Pattern Recognition*, Vol. 3, pp. 607–610, ISBN 0-8186-2920, Hague, The Netherlands, August 31 - September 3, 1992

Ubul, K.; Hamdulla, A. & Aysa, A. et al. (2008). Research on Uyghur Off-line Handwriting-based Writer Identification. *Proceeding of the 9th International Conference on Signal Processing*, pp. 1656-1659, ISBN 978-1-4244-2178-7, Beijing, China, October 26-29, 2008

Ubul, K.; Tursun, D. & Hamdulla, A. et al. (2009). A feature selection and extraction method for Uyghur handwriting-based writer identification. *Proceeding of the 1st International Conference on Computational Intelligence and Natural Computing*, Vol. 2, pp. 345 –348, ISBN 978-0-7695-3645-3, Wuhan, China, June 6-7, 2009



Genetic Algorithms in Applications

Edited by Dr. Rustem Popa

ISBN 978-953-51-0400-1
Hard cover, 328 pages
Publisher InTech
Published online 21, March, 2012
Published in print edition March, 2012

Genetic Algorithms (GAs) are one of several techniques in the family of Evolutionary Algorithms - algorithms that search for solutions to optimization problems by "evolving" better and better solutions. Genetic Algorithms have been applied in science, engineering, business and social sciences. This book consists of 16 chapters organized into five sections. The first section deals with some applications in automatic control, the second section contains several applications in scheduling of resources, and the third section introduces some applications in electrical and electronics engineering. The next section illustrates some examples of character recognition and multi-criteria classification, and the last one deals with trading systems. These evolutionary techniques may be useful to engineers and scientists in various fields of specialization, who need some optimization techniques in their work and who may be using Genetic Algorithms in their applications for the first time. These applications may be useful to many other people who are getting familiar with the subject of Genetic Algorithms.

INTECH
open science | open minds

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

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