International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.1 (2014), pp.415-424
http://dx.doi.org/10.14257/ijsip.2014.7.1.38

Research on Uyghur Handwriting Identification Technology Based on Stroke Statistical Features

Askar Hamdulla, Guzaltaji Naby, Kurban Ubul and Kamil Moydin

Xinjiang University, Urumqi Xinjiang 830046, P.R. China
askar@xju.edu.cn

Abstract

The automatic handwriting identification is a hot topic in pattern recognition that it has been extensively studied in many languages. A Uyghur handwriting identification technique based on the stroke statistical features is proposed in this paper. Firstly, the handwriting image is preprocessed taking modified methods of grid line removal, noise reduction and thinning. Then novel stroke statistical features are extracted based on the structural character and writing styles of Uyghur handwriting. And this approach respectively achieves a top 1 and top 2 identification rates of 98.66% and 99.78% on the Uyghur handwriting data set from 224 different people. Finally, Comparison analysis of different stroke length and distance measurement method has been conducted through three different kinds of experiments, the optimal stroke length and distance measurement method is determined, and its effectiveness and stability are tested. The stroke statistical features can capture the structural character and writing style of Uyghur handwriting efficiently, and it is suitable for any languages theoretically.

Key words: *Uyghur; Handwriting identification; Thinning; Stroke statistical feature; Similarity measurement*

1 .Introduction

Handwriting Identification is a science and technology that aims to judge the identity of the writer according to handwriting styles [1]. The purpose of handwriting identification is to find a sample from the reference samples written by different people which closest to the characteristics of the test handwriting.

Handwriting identification technique is divided into online and offline categories. The offline handwriting identification includes text relevant technique and text independent technique [2]. Text relevant technique needs to know the characters contained in the text by character recognition or artificial calibration. However, text independent technique is not considered the text information, handwriting style information directly extracted from the preprocessed handwriting image; therefore, it has a wider range of use.

Today, Uyghur are primarily living in north western part of China. They are using Arabic script based Uyghur which has the nature of including complex character forms and frequent connection between characters. Uyghur offline handwritten character segmentation and recognition technology is not mature, and manual calibration is expensive; so the article is does not consider the text relevant method which takes a single character or word for the unit of the text, and dedicated to the study of text-independent writer identification technique.

The texture analysis is one of the identification methods that most commonly used for text independent handwriting identification. The texture analysis based text independent handwriting identification technique has successfully used in English, Chinese, Arabic and

ISSN: 2005-4254 IJSIP

other script's handwriting identification technology. Textural features are extracted by Gabor filter in [3-5]. Article [6] and [7] extracted wavelet feature from texture block and built a statistical probability framework model to describe handwriting styles. In Uyghur handwriting identification, Ubul *et al.*, [8] used Gabor filter and genetic algorithm for the extraction of handwriting characteristics, Linuan *et al.*, [9] also used Gabor wavelet transform to extract global features and using a mixture Mahalanobis distance to classify. The [10] also used Gabor wavelet for texture feature extraction. Handwriting texture image formation depends on the character or text stitching, while different combinations of characters or text make great changes in texture, this will affect the stability of the extraction of the handwriting features. Moreover, there are some difficulties in character segmentation of Uyghur handwriting, so, the texture analysis based method is not used in this paper.

Bulacu and Schomaker presented a series of probability distribution function (PDF) characteristics for text-independent writer identification method, and they have extremely successful application in English and Arabic handwriting. The Contour-Hinge PDF [11] and the Grapheme emission PDF [12] has the best performance among them. Learn from the thought of the probability distribution function, [13] proposed a text-independent method of writer identification based on grid-window microstructure feature for multi-language handwritings, and have a good identification results on Chinese, English, Tibetan and Uyghur handwriting. Extraction of the Contour-Hinge PDF needs to judge inner and outer boundaries of the character, Grapheme emission PDF needs clustering, and its time complexity is very high. Finding stable segmentation fragments for Uyghur handwriting, which has complex structure and frequent hyphenation, by a robust segmentation method is very difficult. Therefore, draw on the thinking of the probability distribution function, this paper presented a stroke statistical feature for text-independent Uyghur handwriting identification. Stroke statistical feature is a probability distribution function such as a series of features in [11-13]. In addition, we use the Euclidean distance, similarity, chi-square distance for distance measurement between the feature vectors, and comparatively analyzed identification rates. The experimental results show that the method has a higher identification result.

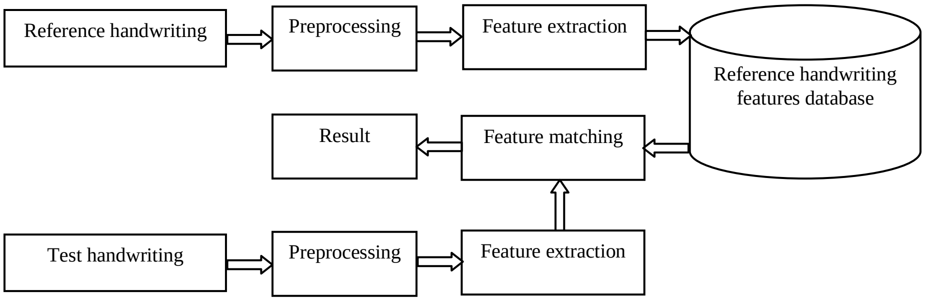


Figure 1. Handwriting Identification System Flow Chart

2. Handwriting Feature Extraction Algorithm

2.1. Preprocessing

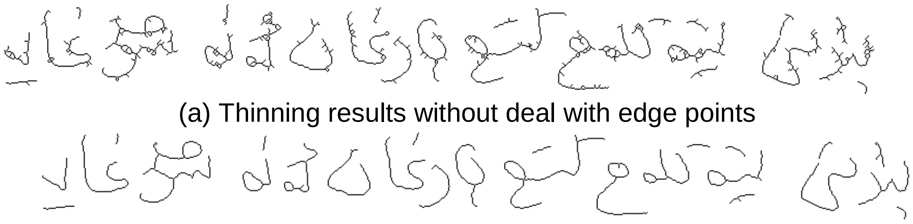
The 1344 Uyghur handwriting samples collected from 224 different people. Not limited to the contents of the text was interlacing written in the manuscript in their own style, and scanned the handwriting samples by using CanonMP810 type scanner, the scanning resolution of 300dpi, handwriting samples database stored in BMP format.

In order to maintain the original handwriting form, the normalizing operation no longer be used. Image preprocessing includes image binarization (simultaneously removed the non-handwriting information, graphics, grid, *etc.*), removal of the scatter noise and thinning, in order to obtain the binary image contains only the ink pixel skeleton.

Handwriting samples are written in the manuscript and the grid of the manuscript does not affect the writing style, but will affect the access of handwriting information. Grid of the selected manuscript is red, text color is mostly black or blue, therefore, the pixels which have higher red, green components in the red and green histogram have been set to white when remove background [14], the remaining pixels set to black to get a binary image.

Binary image usually contains some scatter noise. In order to reduce their affections to the feature extraction, these scatter noise has been removed. The scatter noise length threshold was set up based on the actual situation of the handwriting samples. If the numbers of black spots connected with the inspection point are less than the threshold, inspection point is estimated a noise point, and all of the scatter noise points filled with white.

The thicknesses of the character outline are not the same level, and these will make handwriting feature extraction more complex, so the stroke statistical features proposed in this paper are extracted from the thinned image. Skeleton extraction from the source image is the aim of thinning [15], that is, the lines which have greater than one pixel width in the original image are thinned into only one pixel wide lines to form a 'skeleton'. After the formation of the skeleton, it is relatively easy to extract the stroke statistical features, and thinning does not affect the writing style of handwriting. Commonly used thinning algorithms are Hilditch thinning algorithm, Pavlidis thinning algorithm, Zhang thinning algorithm, Rosenfeld thinning and the index table thinning algorithm. This article employed Rosenfeld thinning algorithm, but its result is not ideal. Through the analysis of the original binary image and thinning results, we summed up the causes and solutions. Some edge points in handwriting image have great effect on the thinning algorithm, thence, these edge points need to be preprocessed before thinning. Processing method is as follows: first, calculate the number of black pixels around the each pixel; if the numbers of black pixels are less than 2, the point is set to white. If the numbers of black pixels are greater than 6, the point is set to black. The handling edge points in before and after thinning are indicated as Figure 1.



(a) Thinning results without deal with edge points

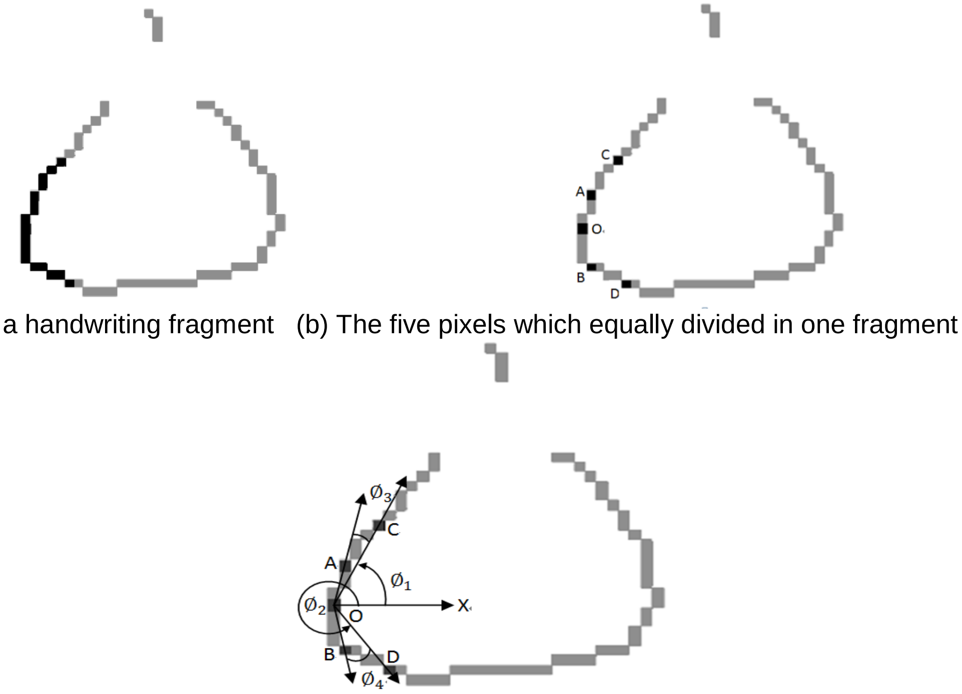
(b) Thinning results after deal with edge points

Figure 2. Thinning Effect Figure

2.2. Stroke Statistical Features Extraction

It is a crucial point to extract features from thinned handwriting image in this paper. Handwriting can be seen as numerous connected tiny fragments; in order to get for more details and more accurately describe the handwriting, expression of these fragments in a certain way and as handwriting features will greatly enhance the identification rate of the handwriting. There are some difficulties in segmentation of characters in Uyghur handwriting and the cost of manual segmentation is very high.

Firstly, describe stroke fragments in some way within the local area and statistics the probability within the global area to get the stroke statistical features. Image lines after the thinning are one pixel wide, therefore, each black pixel as a central point to find out 2m connected black pixels (m is a positive integer), this series of pixels are a 2m +1 length stroke, which is mentioned in this article, the ink fragments. Using all the pixel locations to describe the fragments is very informative, the computational complexity is high, and system uptime will be long, which directly affects the effectiveness of the system. We choose five pixels to describe the fragments which are equally divided; the five pixels are the center point O, two endpoints C and D, and two midpoints A and B shown in figure 3. Stroke statistical feature extraction diagram is shown in Figure 3.



(a) a handwriting fragment (b) The five pixels which equally divided in one fragment

(c) Five pixels' location relationships

Figure 3. Stroke Statistical Features Extraction Diagram

If the coordinate of center point is (x_k, y_k), and another point which has a distance of ε from the center point is ($x_{k+\varepsilon}$, $y_{k+\varepsilon}$), then the angle between the connection of the two points and the horizontal line is computed using equation (1):

$$\phi = \arctan\left(\frac{y_{k+\varepsilon} - y_k}{x_{k+\varepsilon} - x_k}\right) \tag{1}$$

In Figure 3(c), $\emptyset_1$and $\emptyset_2$are the angles from the positive direction of x-axis to the vectors connecting the center point and the endpoint respectively; they can determine the opening direction of the handwriting fragments and it's roughly curvature. $\emptyset_3$ and $\emptyset_4$ are the angles between two vectors which connecting the center point with midpoint and connecting the center point with the endpoint; they both indicate the curvature change within fragment. This algorithm can uniquely identify the handwriting fragments by these four angles.

As the algorithm runs over all the black pixels, the four angles of the each local fragment are computed and angle histograms are thereby built. The angle histograms are then normalized to a joint probability distribution function $P(\emptyset_1, \emptyset_2, \emptyset_3, \emptyset_4)$, which gives the probability of the fragments in the handwriting image.

$\emptyset_1$and $\emptyset_2$are randomly resides all quadrants, so these angles normalized to 18 equal portions trough experimentations, each segment of 20° gives a sufficiently detailed and, at the same time, sufficiently robust description of handwriting to be used in writer identification. By analyzing the range of $\emptyset_3$and $\emptyset_4$, we can know that the most angles are distributed in 0° to 30°, and some beyond this, but not much. Thus, when normalize the two angles, average normalization method no longer be used, these angles will be divided into three segments: 0° to 10°, 10° to 30° and 30°, so the dimension of the features are 18 × 18 × 3 × 3 = 2916. Although the dimensions of feature are more than two thousand, but its training and testing time is only about ten minutes for 1344 samples.

It should be noted that the stroke statistical features are dealing with text-independent handwriting samples. Different text content will cause a change of strokes categories in handwriting; it would also produce the differences of the stroke probability distribution. But when the text contained characters reach a certain number, the proportion of various stroke categories will reach statistical stability; then the stroke statistical features will be able to reflect the different style of handwriting between the writers.

3. Similarity Measurement

After capturing writer individuality by the above features, it is need to use some distance measurement method to compute (dis)similarity between any two feature vectors. In handwriting identification field, classifiers used widely are the neural network classifier, support vector machine classifier etc. But considering the handwriting sample database is not large, the paper selected similarity measurement methods for classification and comparative analysis of the experimental results, such as Euclidean distance, chi-square distance and similarity. Suppose there are two handwriting vectors ϑ_1 and ϑ_2, the Euclidean distance $d_{Euc}(\vartheta_1, \vartheta_2)$, chi-square distance $d_{Chi}(\vartheta_1, \vartheta_2)$ and the similarity $sim(\vartheta_1, \vartheta_2)$ between them are as follows:

$$d_{Euc}(\vartheta_1, \vartheta_2) = \sqrt{\sum_{i=1}^{n}(\vartheta_{1i} - \vartheta_{2i})^2} \tag{2}$$

$$d_{Chi}(\vartheta_1, \vartheta_2) = \sum_{i=1}^{n} \frac{(\vartheta_{1i} - \vartheta_{2i})^2}{\vartheta_{1i} + \vartheta_{2i}} \tag{3}$$

$$sim(\vartheta_1, \vartheta_2) = \frac{\sum_{i=1}^{n} \vartheta_{1i} \times \vartheta_{2i}}{\sqrt{(\sum_{i=1}^{n} \vartheta_{1i}^2)(\sum_{i=1}^{n} \vartheta_{2i}^2)}} \tag{4}$$

Where, the ϑ_{1i} and ϑ_{2i} are the dimensional elements of ϑ_1 and ϑ_2 respectively, n is the dimensionality of the feature.

4. Experimental Process and Results

Handwriting identification is to identify a most likely writer of query handwriting among a large number of candidate reference handwriting. A Uyghur handwriting identification system is implemented by C + +, and experiments are conducted on 1344 different Uyghur handwriting samples written by 224 individuals.

Choice of stroke length in stroke statistical features will directly affect the identification rate. Too short fragments cannot correctly describe the handwriting style. However, if the fragments are too long, segmentation effect is not ideal because of the interlaced points in Handwriting. So in order to determine the optimal stroke length, different stroke lengths are chosen during the feature extraction and classification experiments. Three types of experiments are conducted here.

Experiments 1: Different stroke lengths are taken in experiment 1. This experiment is conducted on 224 writers, six samples for per writer, four for training, and two for testing; experimental results of different stroke length and different distance measurement method are shown as the following Figure 4.

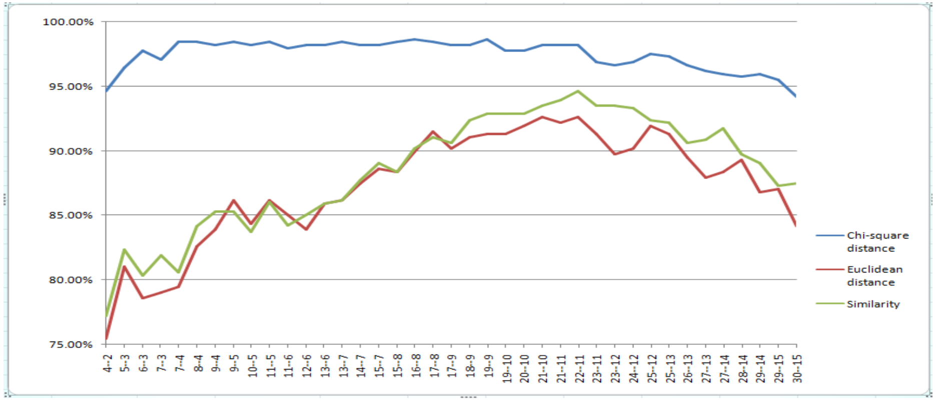


Figure 4. Results of Different Stroke Length and Different Distance Measurement Method

In Figure 4, the vertical axes indicated the identification rate, and the horizontal axes indicated the different stroke length to be selected. In the abscissa unit p- -q, p represents the distance from the center point to the two end points in a stroke, and q represents the distance between center point and two midpoints in a stroke. Take 4- -2 as an example, the stroke length is 9, the end points distance from the center point is 4, the mid-points distance from the center point is 2.

It can be seen from the figure above that identification rates are quite different at the different distances measurement methods, the classification results of the chi-square distance are the best, the similarity is slightly better than the Euclidean distance.

When using the chi-square distance, the identification results of the different stroke length have a slight difference, in which the identification rates are relatively stable and the 98.66% of highest identification rate is achieved with 16- -8 stroke length. In order to further determine the stability of the stroke length, two kinds of experiments indicated as the following (experiment 2 and experiment 3) are conducted.

Experiment 2: Training and testing samples are interchanged together in the experiment 2. In this test, there are six samples for per writer, two for training, and four for testing; experimental results are shown as the following Figure 5.

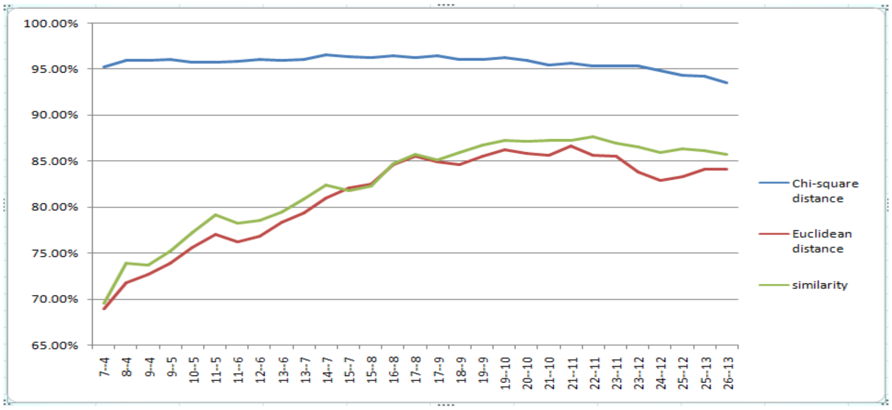


Figure 5. Training and Testing Samples Interchange Experimental Results

The results of the experiment 2 are slightly lower than the results of experiment 1, but they are still relatively stable. The highest identification rate of 96.54% is achieved when 14- -7 of stroke length is selected. Identification rate of chi-square distance is still much better than the Euclidean distance and similarity.

Experiment 3: The handwriting samples are randomly selected in the experiment 3. In this test, we randomly select four samples in each writer's samples for training, the rest of the 2 samples are used for testing. Five random experiments were conducted using chi-square distance, and the average results of them are taken as the random experiment results shown in Figure 6.

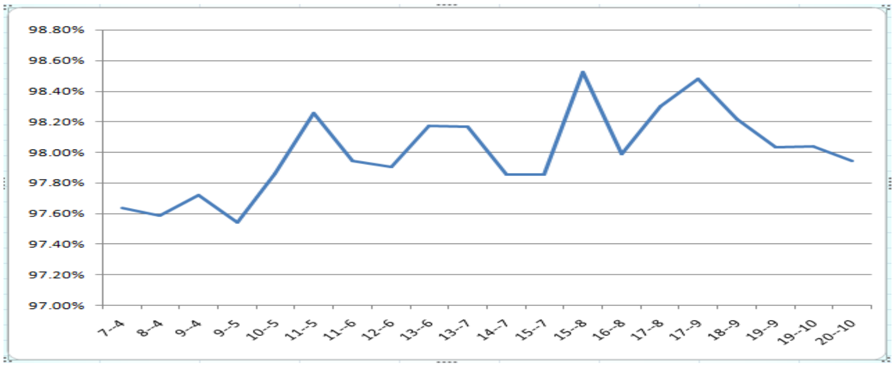


Figure 6. Random Experiment Results

Results of random experiment show that the identification rate is relatively stable. The highest identification rate reached 98.53% when 15- - 8 of stroke length is selected and chi-square distance is used.

Stroke statistical features presented in this paper are indicated high accuracy and stability that the 98.66% of highest identification rates and 99.78% of top 2 identification rate s are obtained respectively. So this feature is most suitable in Uyghur handwriting identification.

5. Conclusions

This paper presented stroke statistical features based on Uyghur handwriting identification method according to the characteristics of Uyghur handwriting. It shows higher and more stable identification rates in Uyghur handwriting database. Experimental results proved that this feature can reflect the handwriting style differences of different writers. It is also show that the performance of simple distance measurement method is very good, it is not need to use a weighted distance measurement method. Therefore, it can be avoided seeking process of standard deviation in [13], which greatly reduces the computational complexity. This feature does not require clustering, and do not need to determine the internal and external boundaries of the handwriting. Unlike other writer identification methods such as texture analysis, it is no need to construct the texture block and multi-channel filters, writer identification method proposed in this paper directly extract stroke statistical features from the thinned handwriting image. Therefore, the method has more advantages in practical applications. It does not require text content analysis and it can be avoided the text image segmentation and text marks, because this method is a text-independent writer identification method. For Uyghur handwriting, this is a practical writer identification method. Under the premise of increasing number of samples, we will need to further improve the feature and study an appropriate classification method to enhance the identification rate and its stability.

Acknowledgements

This work is supported by Program for New Century Excellent Talents in University (NCET-10-0969) and the National Natural Science Foundation of China (No. 61065001 and No. 61163028).

References

[1] R. Plamondon and G. Lorette, “Automatic signature verification and writer identification-the state of the art.Pattern Recognition”, vol. 107, no. 131, **(1989)**.

[2] H. Said, T. Tan and K. Baker, “Personal identification based on handwriting”, Pattern Recognition, vol. 149, no. 160, **(2000)**.

[3] Y. Zhu, T. Tan and Y. Wang, “Biometric personal identification based on handwriting”, Proceedings of 15th International Conference on Pattern Recognition, Barcelona, Spain, IEEE, **(2000)**, pp. 797-800.

[4] Z. Y. He and Y. Y. Tang, “Chinese handwriting-based writer identification by texture analysis”, Proceedings of 3rd International Conference on Machine Learning and Cybernetics, Shanghai, China, IEEE, **(2004)**, pp. 3488-3491.

[5] F. Shahabi Nejad and M. Rahmati, “A New Method for Writer Identification and Verification Based on Farsi/Arabic Handwritten Texts”, Proc of 9th International Conference on Document Analysis and Recognition, **(2007)**, pp. 829-833.

[6] Z. He, B. Fang, J. Du, Y. Tang and X. You, “A novel method for offline handwriting-based writer identification”, Proceedings of the 8th International Conference on Document Analysis and Recognition, Seoul, Korea, IEEE, **(2005)**, pp. 242-246.

[7] Z. He, X. You and Y. Tang, “Writer identification of Chinese handwriting documents using hidden Markov tree model”, Pattern Recognition, vol. 1295, no. 1307, **(2005)**.

[8] K. Ubul, A. Hamdulla, A. Aysa, A. Raxidin and R. Mahmut, “Research on Uyghur off-line handwriting-based writer identification”, Proceedings of the 9th International Conference of signal processing (ICSP2008), **(2008)**, pp. 1656-1659.

[9] L. I. Yuan and K. Moydi, “Uyghur handwriting distinction method research”, (in Chinese), Computer technology and development, vol. 9, no. 11, **(2008)**.

[10] A. Raxiding, “Research on Uyghur handwriting feature extraction method based on Gabor wavelet”, (in Chinese), Journal of Hotan teachers college, **(2010)**.

[11] M. Bulacu, L. Schomaker and A. Brink, “Text-Independent Writer Identification and Verification on Offline Arabic Handwriting”, Proceedings of the 9th International Conference on Document Analysis and Recognition, Curitiba, Brazil, IEEE, **(2007)**, pp. 769,773.

[12] M. Bulacu and L. Schomaker, “Text-Independent Writer Identification and Verification Using Textural and Allographic features”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **(2007)**, pp. 701-717.

[13] L. Xin, D. Xiaoqing and J. Peng Liangrui, “A microstructure feature based text-independent method of writer identification for multilingual handwritings”, Acta Automatica Sinica, **(2009)**, pp. 1199-1208.

[14] K. Ubul, “Research on features extraction and selection methods for Uyghur off-line handwriting-based writer identification”, (in Chinese), Xinjiang University, **(2009)**.

[15] R. Stefanelli and A. Rosenfeld, "Some parallel thinning algorithm for digital pectures", Journal of the association for computing machinery, **(1971)**, pp. 255-264.

Authors

Askar Hamdulla, received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA, tutored by Professor Biing-Hwang(Fred) Juang. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 90 technical papers on speech synthesis, natural language processing and image processing. He is an affiliate member of IEEE.

Guzaltaji Naby, received her B.E. and M.S. degree in electronics, signal and information processing from Xinjiang University, China, in 2009 and 2012, respectively. Her scientific interest includes handwriting identification. Currently, she is a research assistant at the Key Laboratory of Intelligent Information Processing, Xinjiang University, China. Her research interests include feature extraction and selection techniques, writer identification and verification.

Kurban Ubul, received his B.E. and M.S. degree in communication engineering, communication and information system from Xinjiang University, China, in 1997 and 2009, respectively. Since 1997, he has been working as a teacher in School of Information Science and Engineering, Xinjiang University, and became an associate professor in 2009. He was a visiting scholar (researcher) in Carleton University, Canada from 2011 to 2012. His research interests include image processing, pattern recognition, speech signal processing and digital signal processing.

Kamil Moydin, received his B.E. and M.S. degree in radio electronics and computer science from Xinjiang University, China, and Osaka Institute of Technology, Japan in 1983 and 1998, respectively. He has been working as a teacher in School of Information Science and Engineering, Xinjiang University since 1983. He was a visiting scholar in the Osaka Institute of Technology, Japan from 1994 to 1996. In 2002, he got the position of associate professor in Xinjiang University. His research interests include computer network, pattern recognition, and digital image processing.