Classical Mongolian Words Recognition in Historical Document

Guanglai Gao, Xiangdong Su, Hongxi Wei, Yeyun Gong
School of Computer Science
Inner Mongolia University
Hohhot, China 010021
csggl@imu.edu.cn

Abstract—**There are many classical Mongolian historical documents which are reserved in image form, and as a result it is difficult for us to explore and retrieve them. In this paper, we investigate the peculiarities of classical Mongolian documents and propose an approach to recognize the words in them. We design an algorithm to segment the Mongolian words into several Glyph Units(Glyph Unit abbr. GU). Each GU is consisted of no more than three characters. Then we used a three-stage method to recognize the GUs. At the first stage, all the GUs are classified into nine groups by decision tree using three features of the GUs. At the second stage, the GUs in each group are classified individually by five independent BP Neutral Networks whose inputs are other five feature vectors of the GUs. At the last stage, the five results of each GU group from the above five classifiers are combined to provide the final recognized result. The recognition rate of the Mongolian words in our experiment achieves 71%, indicating that our method is effective.**

Keywords- *Classical Mongolian, off-line Handwritten Recognition, Mongolian Segmentation, Multi-classifier Combination*

I. INTRODUCTION

Character Recognition (CR) is the mechanical or electronic translation of printed and handwritten text in image form or other data forms into machine-encoded text. Many works have been done in CR such as [1] on Latin, [2] on Arabic, [3] on Chinese and so on. Now the recognition of machine-printed text on clear scanned document images has already reached a very high degree of accuracy and many applied systems are used in official tasks (e.g. Tablet PCs). However, the recognition of handwritten text is still a challenging task. In Asia, classical Mongolian is used by so many people in China. Nevertheless, classical Mongolian recognition, especially the handwritten classical Mongolian, is barely examined. In this paper, we investigate the recognition of classical Mongolian words in historical documents which possessed the character of handwritten.

In our study, the research objects are the classical Mongolian words in woodblock-printed historical documents. (In woodblock printing, some craftsmen wrote the classical Mongolian words on wooden boards and others carved these words into these wooden boards according to the strokes on the boards, much like seals, and then the printers covered the wooden boards with ink to print the texts on the paper). Fig. 1 is a fragment of classical Mongolian historical document. The challenges in recognition the words are as follows:

Firstly, multiple writers and multiple xylographers create many variants of the same word. These differences among the variants are in the aspects of stretch, skew, relative size, and character appearance. Take the word in Fig. 2 for example.

Secondly, the words used to be low quality partly due to the ancient print techniques and materials. The words in documents can not be clearly recognized by computers, even if it is not difficult for people to read them.

Thirdly, the historical documents were produced in Qing Dynasty. After more than three hundred years, the words in these documents are not as clear as before because ink mark fell off.

The usual steps in performing document recognition include preprocessing, document decomposition, word recognition and post-processing. In this paper, we put much emphasis on the recognition of classical Mongolian words and simply describe the other steps in which we only used the common techniques.

The methods used in word recognition usually belong to either: segmentation-based approach with which a word is split into small pieces (usually characters) to be recognized and then the results are combined into the code of the original word, such as the methods proposed in [4], or segmentation-free approach with which the word is recognized holistically, such as the methods proposed in [5]. In this study, we use a segmentation-based method in order to recognize the classical Mongolian words.

Firstly, we explain why we use the segmentation-based method. Classical Mongolian word is consisted of one or more Mongolian characters. In a word, the characters are joined together along the base line. In the historical documents, the words show obvious variation in size. The longest word is consisted of more than 20 characters, the shortest words, only one character. Even the same word, its appearances on different pages vary tremendously. This makes holistic recognition more difficult. M. Zand et al. [6] tell us, although the holistic recognition methods release us from the segmentation and usually outperform the segmentation-based methods, their complexities grow as the vocabulary gets larger. According to our estimation, the vocabulary of the historical document is about 100,000 which is a very large number. As to segmentation-based methods, the number of classes of the recognition unit, here GU, is much smaller than the number of the words in vocabulary. Therefore, the segmentation-based methods have an advantage over the segmentation-free methods at

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.145

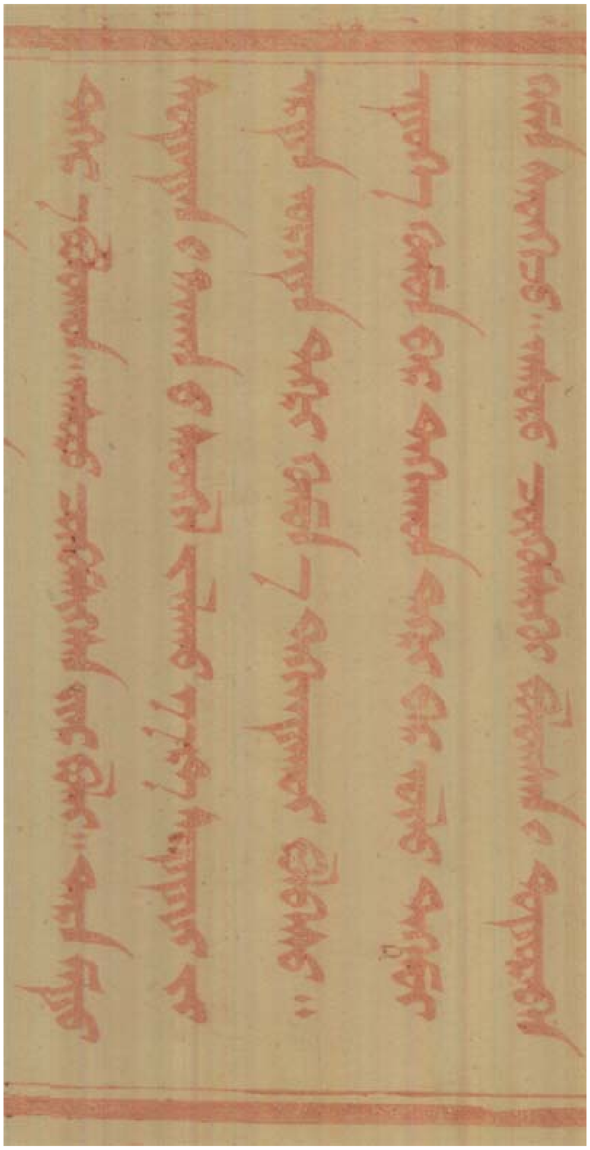


Figure 1. A Fragment of Classical Mongolian Historical Document

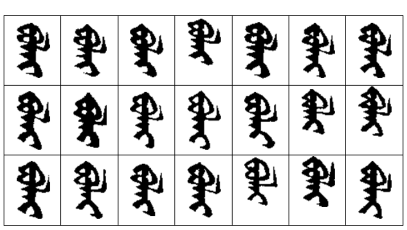


Figure 2. The Variation of

extendability. In addition, most of the holistic recognition methods require us to label every different word at least one time. This is a time-consuming task. Therefore, in this paper we propose a segmentation-based method to recognize the classical Mongolian words in historical documents.

We design an algorithm to segment the word into several GUs. Each GU is consisted of no more than three Mongolian characters. After segmentation, recognizing the GU is a crucial step. In previous studies, many classification techniques have been employed in recognizing the individual characters (strokes or primitives) after segmentation: support vector machine [7], Nearest Neighbor [8], Artificial Neutral Network [9] and so on. In our study, we design a three-stage method which integrates the decision tree and BP neutral network to recognize the GUs. Eight kinds of features of GU image are used in the recognition section, including RP, Euler number, BD, DCT, DWT, PCA, Con&Pro, and EPI. The meaning of these eight features is explained in Table II in section IV. At first stage, we apply a decision tree classifier to divide each GU into one of the nine groups by using RP, Euler number and BD. Then as to the GUs in each group, we apply five BP neutral networks which use the other five features as inputs to recognize them individually. At third stage, we combine the results from the above five BP neutral networks to provide the final recognizing result.

Machine-printed classical Mongolian recognition has been studied in [10][11]. However, the recognition of historical classical Mongolian document is more challenging than the former researches. The recognition rate of the classical Mongolian words in our experiment achieves 71%,, which is a satisfying performance.

The remainder of this paper is organized as follows. In the next section, we simply describe the techniques used in the preprocess,. In section 3, we detail the algorithm used in GU segmentation. In section 4, we describe the features used in the GU recognition module. We present the three-stage classifier used in GU recognition in section 5. Our experiment is demonstrated in section 6. Finally, we conclude our works.

II. Preprocess

In this paper, we recognize the classical Mongolian words in historical document by using a segmentation-based method. In the preprocess section, we firstly correct the slant according to the boundary which surrounds the Mongolian text. And secondly we use the OTSU method to binarize the Mongolian document images and use a method based on wavelet transform to denoise the document image. Thirdly, we split the document images into columns according to their vertical projection profiles. Because each word in the document image is a connect domain, we extract all the words by traverse all the connected domain in each column.

III. GU Segmentation

GU segmentation is a crucial step in classical Mongolian words recognition, because it directly affects the recognition rate. Before GU segmentation, we present the features of classical Mongolian words:

1. A classical Mongolian word is consisted of several characters which are jointed together along the base line.

2. The shapes of the characters differ depending on where they are found in a word. The same character at the heading, medial or end of a word can have a completely different appearance.

The machine-printed Mongolian word can be easily segmented along the base line. On the contrary, it is quite difficult to segment the woodblock-printed Mongolian words, because they have a serious problem of character intersection. Therefore, we segment the Mongolian words into GUs instead of characters. Each GU is consisted of no more than three characters. Take the word in Fig. 3 for example. The machine-printed Fig. 3 (a) is consisted of the five characters on the left of it and it can be easily segmented due to its standardization. On the contrary, it is difficult to segment the woodblock-printed into the expected segmental result listed in Fig. 3 (b), because of character intersection and transformation. Therefore, we segment woodblock-printed into three GUs as listed in Fig. 3 (c). In the segmental result in Fig. 3 (c), the first GU is consisted of three characters: , , ; the second GU and the third GU are consisted of one character individually. The algorithm used in GU segmentation is presented in Table□.

The first step in segmentation is to detect the base line. By analyzing the vertical projection, we found that there are significant decrease in the height near the left boundary and the right boundary of the base line in the Mongolian word projection as presented in Fig. 4. Therefore, there are two characteristics in the left boundary and right boundary of the base line: (1) the gradients of the projection near them are very big, and (2) the variations of the projection heights near them are sharp. According to these characteristics, we could detect most of words' base line. The words, whose heights are low, may not possess these two attributes. In this case, we used a rule-base method to detect their base lines. According to the method mentioned above, positional accuracy of the base lines achieves 98.7% in our experiment.

The second step in segmentation is to judge whether a line meets the segmentation condition. Therefore, it is necessary to compute following ratio of each row in the Mongolian word image:

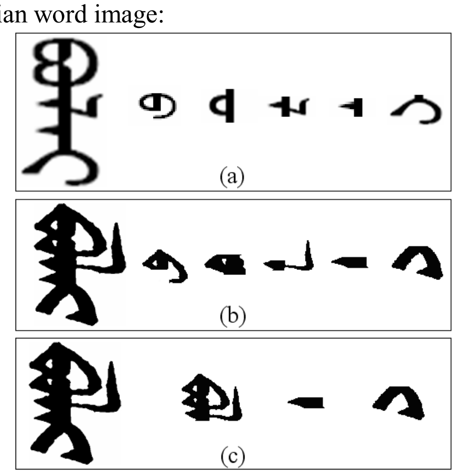


Figure 3. Machine-printed and woodblock-printed Mongolian word of

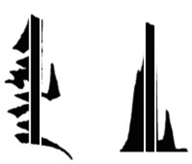


Figure 4. Mongolian Word and its Vertical Projection

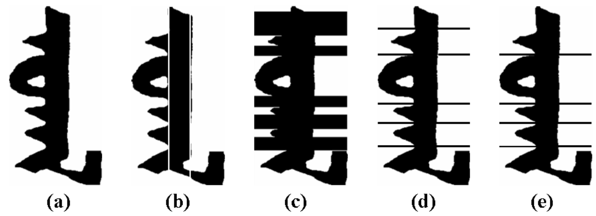


Figure 5. (a)Mongolian Word (b)Base Line (c)Optional Segmentation Line (d)Representative of the Optional Segmentation Line (e)Final Segmentation Result

TABLE I. ALGORITHM USED IN MONGOLIAN GU SEGMENTATION

Description: This algorithm is designed to implement the segmentation task on Mongolian words.
(1) Detecting the base line of the Mongolian word;
(2) Find the rows which conform to the segmentation condition, and use them as optional segmentation line;
(3) Select an optimum row from each group of the continuous optional segmentation line as segmentation line;
(4) Remove the irrational segmentation line

$$R1=abs(BN-(RM-LM))/(RB-LB) \quad (1)$$
$$R2=BN/(RB-LB) \quad (2)$$
$$R3=(RC-LC)/(RB-LB) \quad (3)$$
$$R4=(\min(RM,RB)-\max(LM,LB))/(RB-LB) \quad (4)$$
$$R5=(\min(RM,RB)-\max(LM,LB))/(RM-LM) \quad (5)$$
$$R6=(\min(RC,RB)-\max(LC,LB))/(RB-LB) \quad (6)$$
$$R7=(\min(RC,RB)-\max(LC,LB))/(RC-LC) \quad (7)$$

Where

LM is the x-coordinate of the leftmost black pixel in the specific row.

LC is the x-coordinate of a black pixel, such black pixel meet this condition: the distance between it and the first white pixel on the right of it is the longest in the specific row.

RM is the x-coordinate of the rightmost black pixel in the specific row.

RC is the x-coordinate of a black pixel, such black pixel meet this condition: the distance between it and the first white pixel on the left of it is the longest in the specific row.

BN is the number of black pixels in each row.

LB is the x-coordinate of the left boundary of base line.

RB is the x-coordinate of the right boundary of base line.

As to each row, when its R1, R2, R3, R4, R5, R6 and R7 are in the predetermined ranges, it becomes an optional segmentation line. The ranges are learned according to the training experiment.

Then we select an optimum row from each group of the continuous optional segmentation lines as segmentation line. The optimum row meets specific conditions as follows:

(1) The optimum row is between *m*-10 and *m*+10, where *m* is the middle line in the group of continuous optional segmentation lines.

(2) R6 of the optimum row is biggest, compared to the values gained from all the lines which are between *m*-

10 and m+10.

(3) R2 of the optimum row approximately equals 1.

The last step is to remove the irrational segmentation lines between which there is less than one character. Fig. 5 is example of this algorithm used in word segmentation.

IV. Feature Extraction

As for recognition, we should exactly describe the GU image in a form suitable for the classifier. The descriptions could be the binary image, transformation of the image, statistical features of the image, or the topological features of the image. Some of these numeric descriptions are scalars, and others are vectors. The description is good if it possesses the invariance, namely its values gained from the same GU in different appearances are same. In Table Ⅱ, we list the descriptions used in the recognition of GUs. The first three are scalar, and the rest are vectors.

TABLE II. Descriptions Used in GU Recognition

Name	Meaning
RP	Whether there is a stroke of the GU on the right of the base line.
Euler number	A scalar whose value equals the total number of objects in the image minus the total number of holes in those objects.
BD	Whether the GU's base line is discontinuous.
DCT	Data at the top left corner of metrics gained from discrete cosine transformation
DWT	The third layer's appropriate coefficient gained from the discrete wavelet transformation
PCA	The principal component analysis of the GU image
Con&Pro	The contour and projection of the normalized GU image
EPI	The epitome of the GU image which is obtained by normalizing GU image to 40×40

Here, we explain the feature PCA and Con&Pro futher. In this paper, the feature PCA represents the data resulting from principal component analysis on the GU image. Principal component analysis uses an orthogonal transformation to convert the sub-block of the GU image into a set of values of uncorrelated variables called principal components. In the procedure, we select 400 different GU images (size 400×400), cut them into the scale 40×40, and then use these sub-blocks as the observation data to compute covariance matrix and eigenvectors. We select 20 eigenvectors as basic vectors to transform each sub-block of the GU image into principle components. Then we combine the principle components of all the sub-block to form the PCA feature of the GU image.

The feature Con&Pro represents the contour and projection of GU image. It is a vector which is made up of the following data: (1)the distance between the left border and the first black pixel of each line, (2)the distance between the right border and the last black pixel of each line, (3)the distance between the top border and the first black pixel of each column, and (4)the distance between the bottom border and the last black pixel of each column, (5) the vertical projection of GU, and (6)the horizontal projection of GU.

In table III, we list the concrete dimension of each feature. As to DWT, EPI, and Con&Pro, their dimensions are determined by the GU image size. The dimension of DCT and PCA are chose according to experimental performance.

V. Recognition

Recognizing GU is one of the important steps in Mongolian word recognition. It is a problem of classification. At this step, the features extracted at last step are fed into the classifier to make a final decision. Therefore, classifier selection directly affects the recognition performance. G. Pirlo et al. [12] and M. Grafmüller et al. [13] discuss the strategy which improves the recognition rate by combining multiple classifiers. In order to recognize the GUs, we design a three-stage method which integrates the decision tree and BP neutral network. In the first stage, we use a decision tree classifier to divide each GU into one of the predefined nine groups. As to each GU group, five kinds of feature vectors are fed into five individual BP neutral networks so that training and validation are implemented in the second stage. In the last stage, we combine the results from the above five BP neutral networks to generate the final recognition result. Fig. 6 is the workflow of GU recognition.

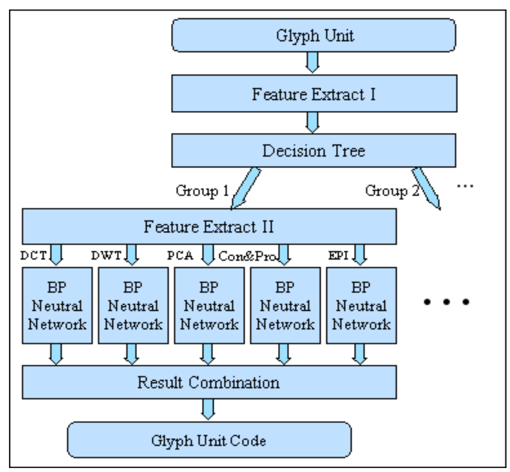


Figure 6. Workflow of GU Recognition

A. Decision Tree

Decision tree is a powerful and popular tool of classification and prediction. It is in the form of a tree structure, where each node is either following one. One is a leaf node which indicates the value of the target attribute (class) of examples. The other is a decision node which specifies some test carried out on a single attribute-value; in addition, such a decision node represents rule and has one branch and sub-tree for each possible outcome of the test. The decision tree learning algorithm creates hierarchical structure of classification rules through searching the attribute which best separates the training data. In our experiment, we use the information gain as the measure to select the attribute in each decision node.

The definition of information is as following:

$$Gain(A) = Info(D) - Info_A(D) \quad (8)$$

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (9)$$

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (10)$$

Where

D is the data partition contains a set of class-labeled training GUs.

m is number of dictinct values of the class label attribute

C_i (for i = 1, : : : , m) is distinct classes in GUs

$C_{i,D}$ be the set of tuples of class C_i in D.

$|D|$ and $|C_{i,D}|$ denote the number of GUs in D and $C_{i,D}$

p_i is the probability that an arbitrary GU in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$.

According to the information gain, we select three attributes to divide the GUs into nine groups. These three attributes are RP, Euler number and BD.

B. BP

In second stage, we use Back-propagation (BP) neutral network as the classifier. BP neutral network have been successful in a wide array of real-world data, including pattern recognition, image processing, adaptive control and so on. It uses the back-propagation learning algorithm to train the perceptrons. The advantages of neural networks, includes their high tolerance of noisy data and their ability to classify patterns on which they have not been trained. They can be used even if you may have little knowledge of the relationships between attributes and classes.

The BP neutral networks used in this study are fully connected and have four layers: an input layer, two hidden layers and an output layers. According to our experimental experience, we select the number of hidden layer's neurons. The number of the first hidden layer's neurons is 200, the number of the second layer's neurons is 25, and the number of output layer's neurons is 1. As for each BP neutral network, the number of input layer's neurons equals the dimension of input feature vector. TableⅢ lists the dimensions of the feature vectors.

TABLE III. THE DIMENSIONS OF THE FEATURE VECTORS

Feature	Con&Pro	DCT	DWT	PCA	EPI
Number	1200	1225	1444	2000	1600

C. Results Combination

In this stage, we combine the results from the five BP neutral networks to make the final decision. It is similar to the vote theory. The score of each category is computed by using (11). We select the category label which has the highest score as the final decision.

$$score_j = \sum_{i=1}^{5} Accuracy_i \times Sig_j \quad (j = 1.....m) \quad (11)$$

Where

m is the total number of the GU categories in the corresponding group.

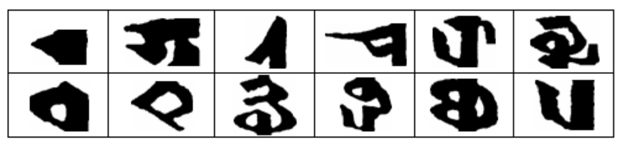
$Accuracy_i$ is the recognition accuracy of the ith BP neutral network individually.

We assign the code to each GU according to its category number in the corresponding group.

VI. EXPERIMENT

In our experiment, the collection is 20 document images of BMP format from the classical Mongolian sutra. The images are 17197 pixels by 5622 pixels in size. We totally extract 5037 Mongolian words from these images. These Mongolian words are segmented through the algorithm mentioned in section III. After segmentation, the number of small components is 27335. Among these small components, the number of correct GUs is 26296. If all the small components which consist of a word are correct GUs, the segmentation of the word is correct. On the contrary, if any component in the segmentation result of a word is not a correct GU, the segmentation of that word is wrong. The overall segmentation accuracy in our experiment is 95.9%, approximately equaling to 96.2%. (96.2% equal to 26296 divided by 27335). There are 159 kinds of GUs in the segmentation results. Part of 159 kinds of GUs are listed in TableⅣ. We divide the GUs into three kinds: Unigram which is consisted of only one character, Bi-gram which is consisted of two characters, and Trigram which is consisted of three characters. Statistically, the majority of the GUs in the segmentation result is the Unigram.

TABLE IV. PART OF 159 KINDS OF GUS



To compare the efficiency of the different features and the classification methods in the recognition of GU, we carried out three groups of experiments. The first group is All-Train-All-Test (ATAT) which means that we use all the GUs to train the classifiers and to validate the classifiers. The second group is 2-fold cross-validation which means that we use half of the QUs to train the classifiers and use the remnant to validate the classifiers. The third group is 5-fold cross-validation which use one group to train and four groups to test. In each group, comparisons are made among three different classification methods which are listed as follows.

B: represents the method only used in the BP neutral networks;

D+B: represents the method which integrates the decision tree and BP neutral network by only using one of the five kinds of features;

D+B+C: represents the three-stage method.

All the experimental results are listed in TableⅤ. Through the statistical analysis of the experimental results, we find:

(1) In each group, the three-stage method achieves the highest recognition accuracy.

(2) Con&Pro is the best feature and DCT is almost the worst feature in the GU recognition which use BP neutral network as classifier.
(3) Using decision tree greatly improves the recognition accuracy. Because it divides the GUs into several groups. In each group, the number of categories which are classified by BP is less.
(4) The result from D+B+C is more accurate than the result from D+B, which proves that multi-evidence theory is appropriate for GU recognition.

According to statistics of the 5-fold experiment, the recognition accuracy of tradition Mongolian word is 71%, which is a satisfying performance. It is lower than the product of GU segmentation accuracy and GU recognition accuracy because if a character in a word is recognized incorrectly, the whole word is recognized incorrectly. Here, GU segmentation accuracy equals 95.9% and GU recognition accuracy equals 78.35% (5-fold cross-validation).

VII. CONCLUSION

In this paper, we investigate the recognition of classical Mongolian words in woodblock-printed historical documents and use a segmentation-based approach to recognize the words in them. The experimental results show our methods used in GU segmentation and GU recognition are effective. Although there is still a gap between recognition accuracy of the machine-printed Mongolian words and recognition accuracy of woodblock-printed Mongolian words, the work in this paper help the digitalization of classical Mongolian historical document a lot and meet the fundamental need of the historical document retrieval. We will try different features to improve the recognition accuracy in the future.

ACKNOWLEDGMENT

This paper is supported by the National Science Foundation of China and the project number is 60865003.

REFERENCES

[1] D. Nasien, S. S. Yuhaniz, H. Haron, "Recognition of Isolated Handwritten Latin Characters using One Continuous Route ofFreeman Chain Code Representation and Feedforward Neural Network Classifier", World Academy of Science, Engineering and Technology, pp. 667-673, 2010.

[2] R. A.-H. Mohamad, L. Likforman-Sulem, C. Mokbel, "Combining Slanted-Frame Classifiers for Improved HMM-Based Arabic Handwriting Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 31, no. 7, pp. 1165-1177, July, 2009.

[3] H.-G. Zhang, J. Yang, W.-H. Deng, J. Guo, "Handwritten Chinese character recognition using Local Discriminant Projection with Prior Information," Proc. 19th International Conference on Pattern Recognition, 2008.

[4] D. Das and R. Yasmin, "Segmentation and Recognition of Unconstrained Bangla Handwritten Numeral," Asian J. Information Technology, Vol. 5(2), pp. 155-159, 2006.

[5] V. Lavrenko, T. M. Rath and R. Manmatha, "Holistic Word Recognition for Handwritten Historical Documents," Proc. the First International Workshop on Document Image Analysis for Libraries (DIAL'04), pp. 278 – 287, 2004.

[6] M. Zand, A. N. Nilchi, and S. A. Monadjemi, "Recognition-based Segmentation in Persian Character Recognition," Proc. World Academy of Science, Engineering and Technology 38, pp. 183-187, 2008.

[7] N. Shanthi and K. Duraiswamy, "Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM," J. Computer Science, Vol. 3, No. 9, pp. 760-764, 2007.

[8] S. Mozaffari, K. Faez and M. Ziaratban, "Structural Decomposition and Statistical Description of Farsi/Arabic Handwritten Numeric Characters," Proc. the 8th Eighth International Conference on Document Analysis and Recognition, Vol. 1, pp. 237- 241, 2005.

[9] O. Batsamhan, Y. P. Singh, "Mongolian character recognition using multilayer perceptron (MLP)," Proc. the 9th International Conference on Neutral nformation Processing (ICONIP'02), Vol. 2, pp.621-625, 2002.

[10] H.-X. Wei and G.-L. Gao, "Machine-Printed Traditional Mongolian Characters Recognition Using BP Neural Networks," Proc. Computational Intelligence and Software Engineering, 2009.

[11] L.-R. Peng, C.-S. Liu, X.-Q. Ding, J.-M. Jin, Y.-S. Wu, H. Wan, Y.-H. Bao, "Multi-font printed Mongolian document recognition system," International Journal on Document Analysis and Recognition, Vol. 13, pp. 93–106, June 2010.

[12] G. Pirlo, C.A.Trullo, and D. Impedovo, "A feedback-based multi-classifier system," Proc. 10th International Conference on Document Analysis and Recognition, Vol. 1, pp. 713 – 717, July 2009.

[13] M. Grafmüller, J. Beyerer and K. Kroschel, "Decision tree classifier for character recognition combining support vector machines and artificial neural networks", Proc. SPIE 7799, 77990B; August, 2010.

TABLE V. ALL THE EXPERIMENTAL RESULTS IN GU RECOGNITION

Group	Feature / Classifier	Con&Pro	DWT	DCT	PCA	EPI
ATAT	B	**97.56%**	97.55%	97.34%	**93.21%**	96.13%
	D+B	**99.99%**	99.99%	99.99%	**99.97%**	99.99%
	D+B+C	**99.99%**				
2-fold	B	**71.29%**	65.43%	**58.36%**	67.98%	66.84%
	D+B	**84.70%**	77.75%	**66.36%**	80.76%	79.41%
	D+B+C	**88.03%**				
5-fold	B	**47.09%**	42.78%	**36.44%**	45.31%	42.79%
	D+B	**74.39%**	67.58%	**57.72%**	71.58%	67.80%
	D+B+C	**78.35%**				