2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

A New Database for Online Handwritten Mongolian Word Recognition

Long-Long Ma, Ji Liu, Jian Wu
National Engineering Research Center of Fundamental Software
Institute of Software, Chinese Academy of Sciences
Beijing, P. R. China
{longlong, wujian}@iscas.ac.cn

***Abstract*—A new online handwritten Mongolian word database, MRG-OHMW, is introduced in this paper. This database contains 946 Mongolian words produced by 300 persons from Mongolian ethnic minority. These Mongolian words are composed of one to fourteen Mongolian characters, and selected from large-scale Mongolian text corpus according to the frequencies of usage. The current version of this database is collected using Anoto pen on paper. The database is further annotated using Mongolian word-level string alignment strategy. We partition the samples into training and test sets, and evaluate the database using the CNN-based recognizer as a baseline. Experimental results reveal a big challenge to higher recognition performance. To our knowledge, MRG-OHMW is the first publicly available database for online handwritten Mongolian research. It provides a basic database to compare empirically different algorithms for online handwritten Mongolian word recognition.**

Keywords—MRG-OHMW; online handwritten Mongolian word recognition; annotation; CNN; evaluation

I. INTRODUCTION

During recent years, online handwritten character recognition has attracted more and more attention from academic and industrial community. The input method based on the recognition technology has been widely used to many portable devices such as tablet PC, electronic whiteboard and mobile phone. Mongolian language is very popular among the Mongolia people in China, Mongolia, and Russia. So the research on online handwritten Mongolian recognition will facilitate the engagement of Mongolians with modern technologies such as the handwriting input.

Considering the characteristic of Mongolian script, and people usually write a Mongolian word as a meaning structure unit. Accordingly, the Mongolian handwriting input also selects Mongolian word as the basic input unit. Online handwritten Mongolian word recognition (OHMWR) is a key technology for the handwriting input method. For developing a high performance online handwritten Mongolian word recognizer, a large scale online handwritten Mongolian word database will play a crucial role.

Compared to the existing handwritten recognition work on CJK (Chinese, Japanese, and Korean) and Arabic, online and offline Mongolian recognition is a relatively unexplored and studied lately field. Researchers from only several institutes are devoted to related work. Gao et al. [1] presented a statistical and structural recognition method for printed Mongolian character recognition. Batsaikhan et al. [2] used multilayer perceptron classifier to recognize noisy Mongolian characters with single font. Two-stage classification method based on glyph segmentation was used to enhance Mongolian character recognition performance [3]. A segmentation-based approach segments classical Mongolian words into several GUs (Glyph Unites) and then recognizes the GUs [4]. These GUs are combined to form word recognition results. Multi-font printed Mongolian documents are recognized by integrating character segmentation and character recognition with support mixed script of Mongolian, Chinese and English [5]. For OHMWR, only several papers report related research works. An OHMWR system is developed based on related research of preprocessing, feature extraction and classification method [6]. Recurrent neural network is used to improve the performance of OHMWR [7].

Even though some researchers have reported the results on Mongolian word recognition, it is difficult to compare and evaluate the performance of different algorithms because of the unavailability of a public online handwritten Mongolian word database. So it is necessary to build an annotated online handwritten Mongolian word database to evaluate the recognition algorithms.

A number of handwritten databases for other languages have been published in the literature since 1990s. A comprehensive survey of the handwritten databases during the last two decades is provided [8]. For offline handwritten databases, there are English databases CENPARMI [9]、CEDAR [10] and IAM [11], Indian database ISI [12], Japanese Kanji character databases ETL8B and ETL9B, Korean database PE92 [13], Chinese databases HCL2000 [14] and HIT-MW [15], Farsi database FHT [16], Arabic [17] and so on. In the field of online handwritten recognition, several databases exist. Some widely used databases are Japanese databases Kuchibue and Nakayosi [18][19], UNIPEN project [20], SCUT-COUCH2008 [21], CASIA-OLHWDB1.0 and CASIA-OLHWDB1.1 [22]. We created an online handwritten Tibetan character database named MRG-OHTC, which was collected using electronic pen on digital tablet [23]. Up to now,

there is no public online handwritten Mongolian database. So our aim is to build an online handwritten Mongolian word database, promote the development of related research, and facilitate the comparison of different recognition algorithms.

Due to different writing style of Mongolian script and unnatural handwriting on digital tablet, our former sampling method [23] isn't suitable for collecting Mongolian words. Our building of online handwritten Mongolian word database is motivated by the work of [22], where samples are collected using Anoto pen on paper. We establish an online handwritten Mongolian word database, MRG-OHMW (MRG is the abbreviation of Multitech Research Group, and OHMW is the abbreviation of online handwritten Mongolian word). The database contains Mongolian word samples of 946 classes. Online samples are written by 300 persons. Mongolian word-level labeling is processed using word string alignment based on dynamic programming (DP). In preliminary experiments on the database, a baseline accuracy of 91.20% is achieved using the CNN (convolutional neural network)-based recognizer.

The rest of the paper is organized as follows. Section II simply introduces the characteristic of Mongolian scripts. The overview on how to design、build and annotate the MRG-OHMW database is presented in Section III. Section VI presents the benchmark testing of the database using the CNN-based recognizer, and Section V offers concluding remarks.

II. CHARACTERISTICS OF MONGOLIAN WORDS AND SCRIPTS

Mongolian is the official national language of Mongolia, and the official provincial language of China's Inner Mongolia Autonomous Region. The Mongolian writing system is different from that of Chinese、English and other Latin languages. Mongolian scripts are written vertically from top to bottom with the column order from left to right (see Fig.1). A blank space is used to separate two consecutive words. Mongolian script consists of 35 letters, including 7 vowels and 28 consonants.

Figure 1. Examples of Mongolian scripts

Similar to Arabic, Mongolian is a context script where letters are cursively joined and have initial, medial and final presentation forms for the same letter. For example, the Mongolian letter u has different forms in a Mongolian word as shown in Fig.2. In addition to these positional forms, many letters also have variant forms used in accordance with spelling and grammatical rules [24]. These unique characteristics bring new challenges to OHMWR research.

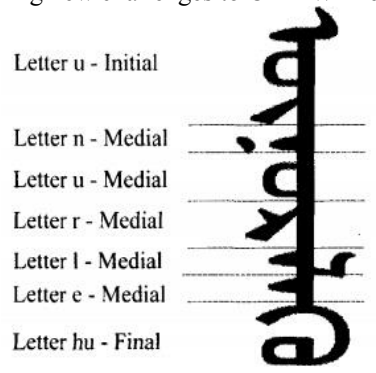

Figure 2. Different forms of letter u in a traditional Mongolian word

III. OVERVIEW OF MRG-OHMW

A. Choosing the Mongolian Word Corpus

The vocabulary of Mongolian words is extremely large. It would be impossible to collect all the words. Therefore, we select the collected Mongolian words according to the frequencies of usage. We get 3000 web pages in the domain of Mongolian arts, news, notice, religious, traditional culture and history from the related Mongolian website, and get 26,438 Mongolian sentences after preprocessing. The statistics on the usage frequency is obtained from these Mongolian sentences. Mongolian words with the highest usage frequencies are selected as word sampling objects.

Ultimately, 946 frequently used Mongolian words were picked as an appropriate number. Table I gives the statistics of Mongolian word type, where words is respectively composed of one to fourteen Mongolian characters. From Table I, we can see the number of three to nine-character Mongolian words account for a larger proportion (78.2%).

TABLE I. STATISTICS OF WORD TYPE

Mongolian Word Type	#Word
one-character	32
two-character	54
three-character	149
four-character	137
five-character	210
six-character	133
seven-character	111
eight-character	56
nine-character	33
ten-character	13
eleven-character	7
twelve-character	6
thirteen-character	3
fourteen-character	2
total	946

B. Sampling Layout Design

There exist many available devices and pens which can record the online handwritten trajectories. The Mongolian writing system is a very special writing system. After analyzing and comparing these devices and pens, we finally decided to use Anoto pen on paper [22], not selecting our former sampling method using electronic pen on digital tablet [23]. Layout design is similar to that of CASIA-OLHWDB database [22]. Mongolian words are printed on papers with dot-matrix pattern, and persons are required to write below the printed characters. Fig.3 shows a part for one of pages of a printed form. In total, the layout with 15 pages is designed to include the selected 946 Mongolian words.

Figure 3. An illustration of layout for collecting samples

According to the sampling layout, the writers are asked to write the Mongolian words in their accustomed writing manner. After collecting all samples for each writer, online data is saved page by page. Each page is thought as an online document including mainly point and stroke information. Meanwhile, Mongolian document transcript of each page is also saved. Online handwritten Mongolian word samples corresponding with the layout of Fig.3 are shown in Fig.4.

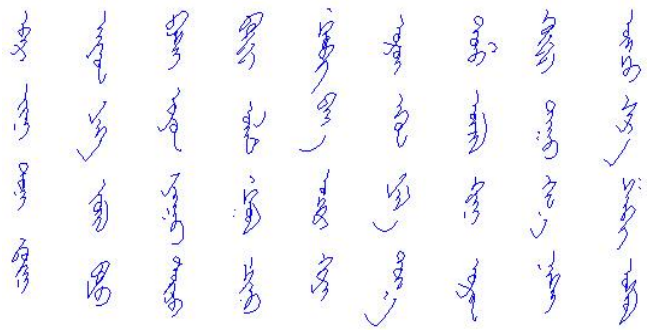

Figure 4. Examples of online handwritten Mongolian word samples

C. Mongolian Word Annotation

Online handwritten Mongolian word samples and corresponding transcript for each page are saved separately. To evaluate the performance, the annotation of Mongolian words is important for OHMWR research. Our objective is to obtain word-level annotation by aligning online handwritten Mongolian document with corresponding transcript for each page.

The annotation process is depicted in Fig.5. The input includes online handwritten Mongolian documents and corresponding transcripts for each page. Generally, Annotation process involves two different level alignments: a text-line level alignment and a character/word alignment. Because most of Mongolian words are well separated, words can be segmented correctly using existing segmentation method. Our annotation process directly involves segmentation and alignment at word level. Firstly, we segment online handwritten Mongolian document into candidate Mongolian word sequences using overlapping degree [25] and rule-based combination. Secondly, word string alignment can be viewed as candidate Mongolian word sequence matching with word-based lexicon in the corresponding transcript. The optimal matching result can be found by DP. During alignment, we did not use Mongolian word classifier as in [26]. After aligning automatically, if the number of segmented words is different from the word number in the transcript of one page, the user need find and correct the segmentation error. Fig.6 shows the annotation results at Mongolian word level.

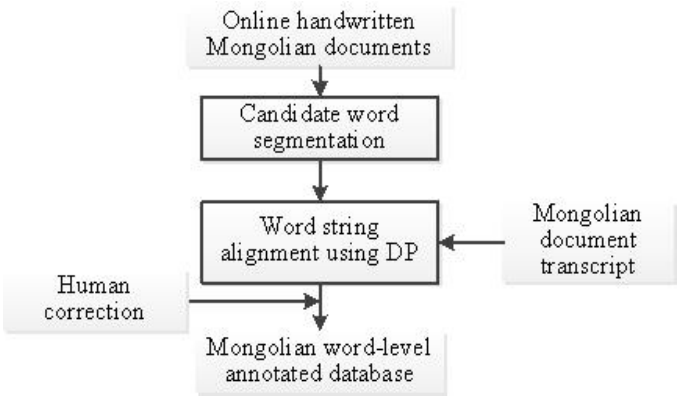

Figure 5. Annotation process for Mongolian word-level annotation

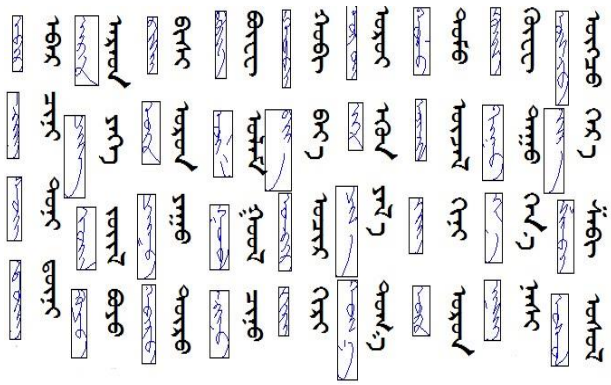

Figure 6. Mongolian word-level annotation results

Note that, we don't further label the annotation of Mongolian words at character level. On the one hand, it is very difficult to correctly segment the characters from online handwritten Mongolian words using computer segmentation

algorithms. On the other hand, this is determined by our research goal and method. Segmentation-free strategy is suitable to OHMWR. After annotation, each Mongolian word with the handwritten point trajectories corresponds to a string of Mongolian character codes.

D. Dataset Statistics

According to the above-mentioned sampling method, in total 300 writers complete their handwriting Mongolian words. We only describe the fundamental statistics including writer distribution and statistics of missing samples.

1) Writer distribution

Considering the particularity of Mongolian script, all the writers are from Mongolian ethnic minority. Firstly, we selected the writers from students in higher school of inner Mongolian autonomous region. College students are enrolled from different city over the autonomous region. Secondly, we selected some clerks from banks、government departments and companies. Most of these clerks are older than twenty-five.

According to the information of all the writers, we calculate the writer distribution from age and gentler. Table II and III give their distributions.

TABLE II. GENDER DISTRIBUTIONS OF ALL THE WRITERS

Items	#Persons	Percentage
Male	131	43.7%
Female	169	56.3%
Total	300	100%

TABLE III. AGE DISTRIBUTIONS OF ALL THE WRITERS

Items	#Persons	Percentage
Below 20	99	33%
Between 21 and 25	167	55.7%
Between 26 and 30	16	5.3%
Older than 30	18	6%
Total	300	100%

2) Statistics of missing samples

Although the policies of collecting samples are employed in the collecting procedure, sample sets of the 300 writers still have some missing characters. By statistics, the total number of valid Mongolian word samples is 282,954 and 846 samples are missing. The statistics of missing samples is shown in Table IV.

TABLE IV. STATISTICS OF MISSING SAMPLES

Items	#dataset	Percentage
0	239	79.7%
Between 0 and 9	52	17.3%
Above 10	9	3%
Total	300	100%

E. Dataset Analysis

MRG-OHMW is also an unconstrained database. We allow the writers to write these word samples in their own writing style. We briefly introduce the characteristic of our database in handwritten diversities.

Each writer has his or her own writing style. Even if written by the same person, different styles may be seen. Fig.7 shows some Mongolian words are written by a writer. Varied handwritten style from different writers is shown for one Mongolian word in Fig 8.

Figure 7. Mongolian word samples from a writer

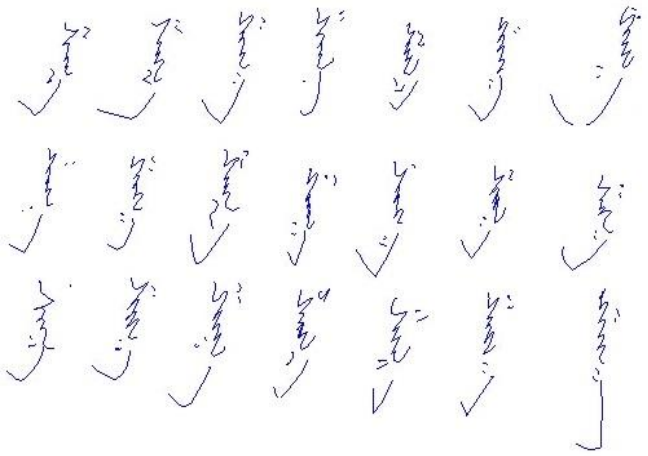

Figure 8. Varied handwritten style of one Mongolian word from different writers

IV. BENCHMARK EVALUATION

To evaluate the MRG-OHMW database, we have done some experiments using the CNN-based recognizer. We select randomly 250 samples per class for training, and the remaining 50 samples per class for evaluating the recognition performance. By this partition, there are 235,826 samples in the training set and 47,128 in the testing set, respectively.

There are two paradigms in word recognition: segmentation-based and segmentation-free. Segmentation-based method segments a word into characters, primitives or strokes. All possible segmentation hypotheses are evaluated, and the optimal hypotheses are regarded as the recognition

results. While segmentation-free method attempts to recognize the whole word without explicit segmentation. The main advantage of segmentation-based method is that the large number of words can be modeled using a finite set of sub-structure units. Segmentation-free method is more pragmatic in the case of cursive handwriting where characters are physically connected with each other and segmentation turns out to be impractical. Considering the characteristic of Mongolian words, OHMWR is more suitable for the segmentation-free method.

With the blooming of deep learning in recent years, CNN brings about new breakthrough technology for handwritten character recognition with great success [27][28]. We use CNN to construct the CNN-based recognizer on the MRG-OHMW database. Online Mongolian word samples are composed of point trajectory sequences, and different samples usually includes different point numbers. To satisfy the requirement that CNN needs the same node number at the input layer, we transform these online handwritten Mongolian words into offline binary word images with the same size of 64×64 image.

The CNN-based recognizer takes the same architecture (see Fig.9) as Lenet-5 [29], which is composed of ten layers (not including the input layer). The input of the CNN-based recognizer is a 64×64 image that contains a 30×30 image at the center. The feature maps of each convolution layer are fully connected to all feature maps of the previous layer. The detail of the network structures is presented in Table V. We conducted our experiments on an open CNN platform called Caffe [30] using GPU-based parallel processing.

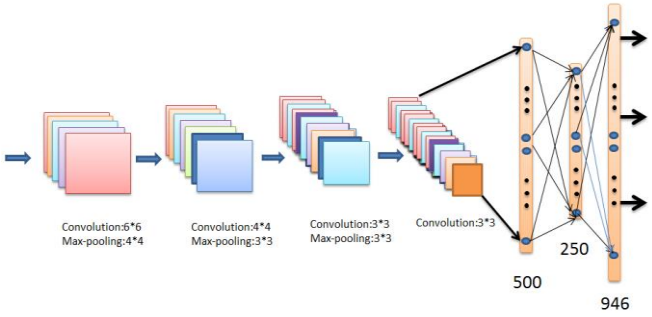

Figure 9. The architecture of CNN-base recognizer for OHMWR

TABLE V. STRUCTURE OF MONGOLIAN RECOGNIZER

Layers	Type	#of feature maps	Feature map size	Window size	Stride
C1	Convolution	32	30×30	6×6	2
P2	Max-pooling	32	14×14	4×4	2
C3	Convolution	81	11×11	4×4	1
P4	Max-pooling	81	9×9	3×3	1
C5	Convolution	210	9×9	3×3	1
P6	Max-pooling	210	5×5	3×3	2
C7	Convolution	360	3×3	3×3	1
F8	Fully connected	500	1×1		
F9	Fully connected	250	1×1		
O10	Output	946	1×1		

Table VI gives the recognition results using the CNN-based recognizer. From table VI we can see the test accuracy is lower than 92%, and there is a big accuracy difference between top1 and top2. This mainly attributes to similar Mongolian word confusion. Future works will be done to improve similar Mongolian word discrimination. Furthermore, the baseline CNN-based recognizer can only recognize the Mongolian words with predefined classes, which appears in the training set. A lexicon-free Mongolian word recognition method is expected to realize and satisfy the application requirement for the handwritten input.

TABLE VI. RECOGNITION RATE USING THE CNN-BASED RECOGNIZER

Method	Top1	Top2	Top3	Top5
Baseline(CNN)	91.20%	97.80%	98.74%	99.23%

V. CONCLUSION

In this paper, we describe a publicly available database, MRG-OHMW, for the research on online handwritten Mongolian word recognition. Our aim is to facilitate the comparison of different Mongolian word recognition algorithms and promote the development of OHMWR research. This database contains 946 frequently used Mongolian words written by 300 different writers. The Mongolian word-level annotation is generated using word string alignment strategy. Preliminary experiments using the CNN-based recognizer demonstrate the challenge of recognition.

Our purpose of collection is not only for research need, but also promises more substantial application. MRG-OHMW is available for academic researches by contacting with the first author. And ultimately we hope this database can improve the performance of online handwritten Mongolian word recognition.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) (no.61540057), Qinghai Natural Science Foundation (no. 2016-ZJ-Y04 and 2016-ZJ-740).The authors would like to thank Minghua Nuo, Inner Mongolia University, and Lei Chen, Institute of Intelligent Machines, Chinese Academy of Sciences, for providing sampling selection and template design.

REFERENCES

[1] G.L. Gao, W. Li, Multi-agent based recognition system of printed Mongolian characters. *Proc. ICAMT*, 2003, pp. 376-381.

[2] O. Batsaikhan, Y.P. Singh, Mongolian character recognition using multilayer perceptron(MLP), *Proc. 9th ICDAR*, 2005, pp. 621-625.

[3] H.X. Wei, G.L. Gao, Machine-printed traditional Mongolian characters recognition using BP neural networks, *Proc. ICCISE*, 2009, pp. 1-7.

[4] G.L. Gao, X.D. Su, et al. Classical Mongolian words recognition in historical document, *Proc. 11th ICDAR*, 2011, pp. 692-697.

[5] L.R. Peng, C.S. Liu, et al. Multi-font printed Mongolian document recognition system, *Int. J. Document Analysis and Recognition*, 13(2): 93-106, 2010.

[6] W.R. Bai, Research of the techinology of online handwriting Mongolia words recognition(in chinese), Master Thesis, 2007.

[7] W. Wu, G.L. Gao, Online handwriting Mongolia words recognition with recurrent neural networks, *Proc. 4th ICCSCIT*, 2009, pp. 165-167.

[8] R. Hussain, A. Raza, I. Siddiqi, K. Khurshid, C. Djeddi, A comprehensive survey of handwritten document benchmarks: structure, usage and evalution, *EURASIP Journal on Image and Video Processing*, 46: 1-24, 2015.

[9] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam, Computer recognition of unconstrained handwritten numerals, *Proc. IEEE*, 80(7): 1162-1180, 1992.

[10] J. Hull, A database for handwritten text recognition research, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(5): 550-554, 1994.

[11] U.V. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, *Int. J. Document Analysis and Recognition*, 5(1): 39-46, 2002.

[12] U. Bhattacharya, B.B. Chaudhuri, Databases for research on recognition of handwritten characters of Indian scripts, *Proc. 8th ICDAR*, 2005, pp. 789-793.

[13] D.H. Kim, Y.S. Hwang, S.T. Park, E.J. Kim, P. S.H, S.Y. Bang, Handwritten Korean character image database PE92, *IEICE Trans. Information and Systems*, E79-D(7): 943-950, 1996.

[14] H.G. Zhang, J. Guo, G. Chen, C. Li, HCL2000-A large-scale handwritten Chinese character database for handwritten character recognition, *Proc. 11th ICDAR*, 2009, pp. 286-290.

[15] T.H. Su, T.W. Zhang, D.J. Guan, Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. *Int. J. Document Analysis and Recognition*, 10(1): 27-38, 2007.

[16] M. Ziaratban, K. Faez, F. Bagheri, FHT: A unconstraint Farsi handwritten text database, *Proc. 11th ICDAR*, 2009, pp. 281-285.

[17] H. Alamri, J. Sadri, C.Y. Suen, N. nobile, A novel comprehensive database for Arabic off-line handwriting recognition, *Proc. 11th ICFHR*, 2008, pp. 664-669.

[18] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L. Higashigawa, K. Akiyama, On-line handwritten character pattern database sampled in a sequence of sentences without any writing instructions, *Proc. 4th ICDAR*, 1997, pp. 376-381.

[19] K. Matsumoto, T. Fukushima, M. Nakagawa, Collection and analysis of on-line handwritten Japanese character patterns, *Proc. 6th ICDAR*, 2001, pp. 496-500.

[20] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, UNIPEN project of on-line data exchange and recognizer benchmarks, *Proc. 12th ICPR*, 1994, pp. 29-33.

[21] Y. Li, L. Jin, X. Zhu, T. Long, SCUT-COUCH2008: a comprehensive online unconstrained Chinese handwriting dataset, *Proc. 11th ICFHR*, 2008, pp. 165-170.

[22] C.-L. Liu, F. Yin, D.H. Wang, Q.-F. Wang, Online and offline handwritten Chinese character recognition: benchmarking on new databases, *Pattern Recognition*. 46 (1): 155-162, 2013.

[23] L.L. Ma, H.D. Liu, J. Wu, MRG-OHTC Database for online handwritten Tibetan character recognition , *Proc. 11th ICDAR*, Beijing, China, 2011, pp. 207-211.

[24] Creating and Supporting OpenType Fonts for the Mongolian Script. http://www.microsoft.com/typography/otfntdev/mongolot/

[25] L.-L. Ma, C.-L. Liu, On-line handwritten Chinese character recognition based on nested segmentation of radicals, *Proc of 2009 CCPR & First CJKPR*, Nanjing, China, 2009, pp. 929-933.

[26] F. Yin, Q.F. Wang, C.-L. Liu, A tool for ground-truthing text lines and characters in off-line handwritten Chinese documents, *Proc. 10th ICDAR*, Barcelona, Spain, 2009, pp. 436-440.

[27] I.J. Kim, X.H. Xie, Handwirtten Hangul recognition using deep convolutional neural networks, *Int. J. Document Analysis and Recognition*, 18: 1-13, 2015.

[28] Z.Y. Zhong, L.W. Jin, Z.C. Xie, High performance offline handwritten Chinese character recognition using goolenet and directional feature maps, *Proc. 13th ICDAR*, Nancy, French, 2015, pp. 846-850.

[29] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*. 86(11): 2278–2324, 1998.

[30] Y. Jia, Caffe: An open source convolutional architecture for fast feature embedding, http://caffe.berkeleyvision.org/, 2013.