MRG-OHTC Database for Online Handwritten Tibetan Character Recognition

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

Abstract*—A handwritten Tibetan database, MRG-OHTC, is presented to facilitate the research of online handwritten Tibetan character recognition. The database contains 910 Tibetan character classes written by 130 persons from Tibetan ethnic minority. These characters are selected from basic set and extension set A of Tibetan coded character set. The current version of this database is collected using electronic pen on digital tablet. We investigate some characteristic of writing style from different writers. We evaluate MRG-OHTC database using existing algorithms as a baseline. Experimental results reveal a big challenge to higher recognition performance. To our knowledge, MRG-OHTC is the first publicly available database for handwritten Tibetan research. It provides a basic database to compare empirically different algorithms for handwritten Tibetan character recognition.

Keywords-MRG-OHTC; online handwritten Tibetan character recognition; Tibetan script; evaluation

I. INTRODUCTION

Publicly available databases are important for the research community in handwritten recognition. They provide standard datasets for comparing and evaluating performances of different algorithms. Only the results obtained from standard datasets can be reliable. Consequently, standard datasets can strongly improve the handwritten recognition researches.

A number of handwritten databases have been published in the literature since 1990s. For offline handwritten databases, there're English databases CENPARMI [1]、CEDAR [2] and IAM [3], Indian database ISI [4], Japanese Kanji character databases ETL8B and ETL9B, Korean database PE92 [5], Chinese databases HCL2000 [6] and HIT-MW [7], Farsi database FHT [8], Arabic [9][10] and so on. In the field of online handwriting, several databases exist. Some widely used databases are Japanese databases Kuchibue and Nakayosi [11][12], UNIPEN project [13], SCUT-COUCH2008 [14], CASIA-OLHWDB1.0 and CASIA-OLHWDB1.1 [15]. To our knowledge, there is not any database of handwritten Tibetan available for the scientific community.

Tibetan language is still used by more than six million people in China at present [18]. The Tibetan character set, which records Tibetan language and Tibetan culture, is very special and different in comparison with other character sets in the world, such as Chinese and so on. Therefore, research on Tibetan character, which will enable easier modernization of Tibetan culture and digitization of Tibetan document, is very important in theoretical value as well as in extensive application perspective.

However, the development of Tibetan character recognition is slower than that of some other languages. One of the reasons is the intrinsic features of the Tibetan characters. Great similarities among different characters, wide varieties in writing style and different shapes for a character are important factors which cause Tibetan character recognition to be very difficult. Another substantial reason is lack of standard databases. For recent works, researchers in on-line Tibetan character recognition used their own collected databases to evaluate system performances [16][17]. Thus, the results often can't be compared directly. To compare different algorithms and select the best ones, these algorithms must be tested on the same database. Indeed, an online handwritten Tibetan character database is very necessary.

Inspired by the databases referenced above, we establish an online handwritten Tibetan character database, MRG-OHTC (MRG is the abbreviation of Multi-tech Research Group, and OHTC is the abbreviation of online handwritten Tibetan character). The database contains character samples of 910 classes. Online samples are written by 130 persons using electronic pen on digital tablet. Point trajectories and Unicode of the sample are kept after writing a sample. So there is no need of database annotation. Some errors caused by data labeling as in [15] can be avoided. Combining the properties of Tibetan characters, we analyzed briefly the collected samples from shape variation and character confusion. MRG-OHTC is evaluated using existing algorithms as a baseline. To promote online handwritten Tibetan character recognition, we decide to publish the database upon publishing this paper.

II. PROPERTIES OF TIBETAN CHARACTERS AND SCRIPTS

Tibetan script consists of 4 vowels and 30 consonants, which are called basic elements. There are two kinds of characters used in handwritten Tibetan characters, that is, single characters (SC) and combined characters (CC).

Syllables are basic spelling units [18], whose structure is shown in Fig.1. Each syllable consists of at most 4 characters

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.50

(those parts surrounded by red dash line bounding boxes in Fig.1). Some characters are called as CC, which are made up of EC (essential consonant), TV (the top vowel), CaEc (the consonant above the EC), CbEc (the consonant below the EC) and BV (the bottom vowel). Some consonants (SC, in total 20) can be as a valid character and they locate at the left (CbCC, the character before CC) or right (1-CaCC and 2-CaCC, the first and second characters after CC) of the CC. Fig.2 gives an example of 4-character syllable. We can see the research on character recognition is helpful to syllable recognition.

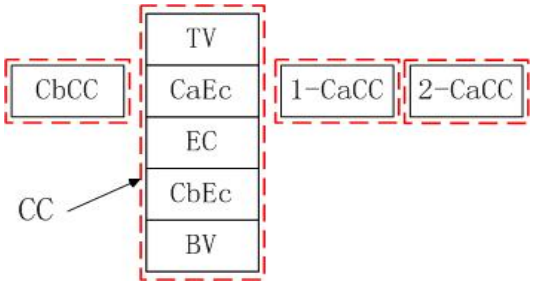


Figure 1. The syllable structure.

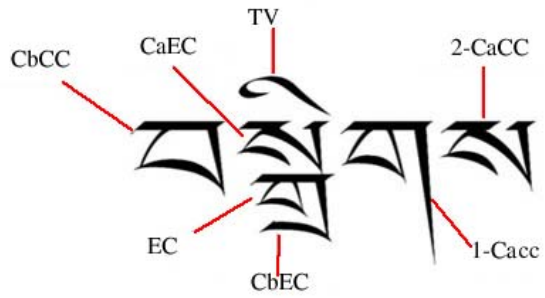


Figure 2. Example of a 4-character syllable.

In terms of the character constitution regulation, basic element is fairly easy for recognition due to smaller number of classes (in total 34). However, for handwritten Tibetan character, two adjacent basic elements in a character may touch tightly or overlap seriously in vertical direction. It is very difficult to decompose a character into basic elements using computer algorithms. So we select characters as basic recognition units.

Combining the above description, several main features of Tibetan character are as follows:

1) Superimposition in vertical direciton: Tibetan character is presented as a vertical combination of consonant and vowel.

2) Baseline feature: There is a flat line (baseline) existing in Tibetan character.

3) Transformation feature: The character displays in different forms when it is in different positions.

III. DATA COLLECTION

Before data collection, some key points need be considered over to make the databases as much representative as possible. We assume the handwritten style varies from education, age, sex, occupation and mood of the writer, etc. Enough attention is paid to include samples from the above style variation. The writers are allowed to write freely on the writing zone. In this section, we will introduce necessary works for sampling preparation.

A. Program Design

Online handwritten data are acquired during writing process and consist of the discrete pen trajectory information. We use electronic pen on digital tablet to collect online samples. Computer programs help to accomplish the sampling process.

As shown in Fig.3 in our program, left part of the interactive interface gives the Unicode and shape of a sampled character, and right part shows the collected sample. While writing on the tablet, point trajectories of the character are shown at the right part of the interface. We place the "Ok and Next" button on the left-top part of the interface and thus makes it convenient for the writers to proceed to the next character class quickly after writing a sample. At the same time, the point trajectories and Unicode of the sample are kept.

Writers are asked to write the samples in the writing zone successively without knowing the content of the next sample. No constraints are imposed to the quality of character shapes.

Compared to sampling strategy in [15], the advantage of our sampling process is no need of data annotation. So some errors of data annotation can be avoided.

Figure 3. An illustration of layout.

B. Character Picking

Now Tibetan coded character set includes basic set (GB/T 16959-1997), extension set A (GB/T 20542-2006) and extension set B (GB/T 22238-2008). The vocabulary of Tibetan characters is extremely huge. It would be impossible to collect all these characters. We pick character classes according to Tibetan character frequency tables from frequency dictionary of modern Tibetan.

To cover all of the frequently used characters in our database, we choose 910 classes according to the frequencies of their usage from basic set and extension set A. Devanagari characters aren't selected because they are seldom used. The range of selected character encoding is as follows: 0x0F00-0x0F0B, 0x0F0D-0x0F14, 0x0F1A-0x0F34, 0x0F3A-0x0F3D, 0x0F40-0x0F47, 0x0F49-0x0F6A, 0xF300-0xF600, 0xF602-0xF605, 0xF610, 0xF62D, 0xF65F, 0xF660, 0xF692-0xF693, 0xF6D7-0xF6D8, 0xF6DC-0xF6DE, 0xF6FB, 0xF6FC, 0xF71D, 0xF720, 0xF733, 0xF748, 0xF762-0xF763, 0xF766-0xF767, 0xF780, 0xF78E, 0xF797, 0xF79A-0xF79B, 0xF7C1, 0xF7DC, 0xF7FA, 0xF80D, 0xF85B, 0xF86F, 0xF89D, 0xF89F, 0xF8BD, 0xF8CA, 0xF8CC, 0xF8EB-0xF8EF, 0xF8F6, 0xF8FF.

C. Writer Sampling

Considering the particularity of Tibetan character, all the writers are from Tibetan ethnic minority. Firstly, we selected the writers from students in higher school (Minzu University of China and Tibet University). College students are enrolled from different Provinces (mainly Qinghai, Sichuan, Gansu and Yunnan). Secondly, we selected some government clerks from Education Technology of Tibet Autonomous Region. Most of these clerks are older than thirty.

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

TABLE I. GENDER DISTRIBUTIONS OF ALL THE WRITERS

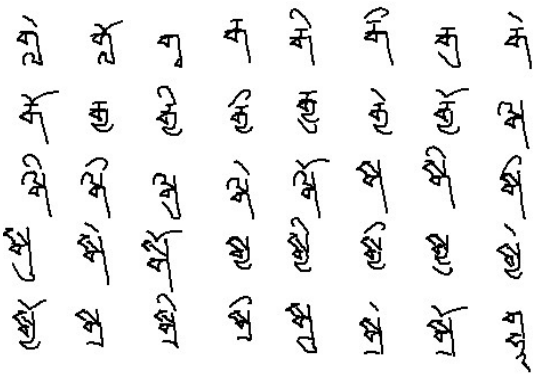
Items	Percentage
Male	42%
Female	58%
Total	100%

TABLE II. AGE DISTRIBUTIONS OF ALL THE WRITERS

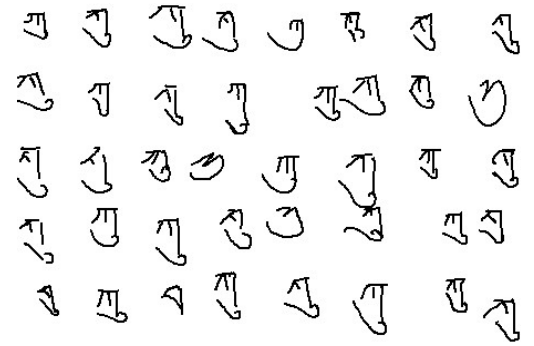
Items	Percentage
Below 20	13%
Between 21 and 25	62%
Between 26 and 30	11%
Older than 30	14%
Total	100%

IV. DATA ANALYSIS

MRG-OHTC is also an unconstrained database. We allow the writers to write these samples in their own writing style. Fig.4 shows some collected samples. In this section, we briefly introduce individual characteristic of our database in shape variation and character confusion.



(a) Samples of different classes from a writer



(b) Samples of one class from different writers

Figure 4. Part of sample sets.

A. Shape Variation

The shape of a character varies from different writers. Even if the character is written by the same person, different styles may be seen. Especially in cursive styles, some character shapes totally differ from the standard shape. So it is difficult to recognize them, even for humans. Fig.5 gives some examples of different writing styles.

B. Character Confusion

There are many similar characters in Tibetan scripts. The confusion of similar characters is a difficult problem for character recognition. Fig.6 shows some similar characters with their Unicode and character shapes. From recognition perspective, similar characters are classified into four categories [20], as follows:

1) Characters with different vowel ི, ེ and the same remains.

2) Superimposed characters composed of similar EC, that is, པ and བ; ས and ལ; ང and ད; ར and ང.

3) Characters with similar TV and CbEc.

4) Characters with other simliar characteristic:

▯, ▯ and ཞ˘, པ˘ etc.

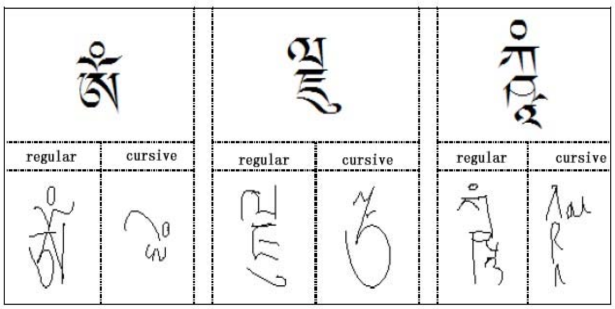


Figure 5. Examples of large shape variation in two different writing styles.

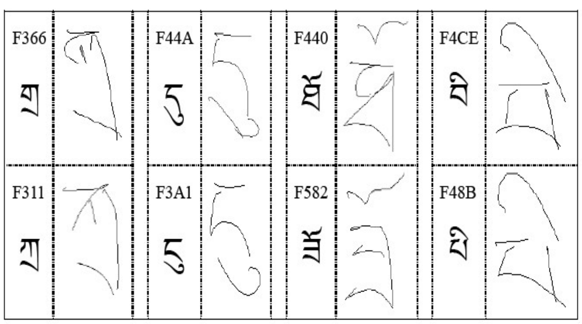


Figure 6. Examples of similar characers.

V. EVALUATION ON MRG-OHTC

To evaluate the MRG-OHTC database, we have done some experiments using existing recognition algorithms. We used 100 samples per class for training classifiers, and the remaining 30 samples per class for evaluating the recognition performance.

For character pre-processing and feature extraction, we adopt the same methods as in [19]. De-noising approach based on mathematical morphology is applied to the preprocessing step. 8-directional features are extracted from each online point of nonlinear normalized online character. The resulting 512-dimensional feature vector is projected onto a 160-dimensional subspace learned by global LDA (Fisher linear discriminant analysis). The 160-dimensional projected vector is then fed to classification. We used a modified quadratic discriminant function (MQDF) classifier [21] with different principal eigenvectors per class. Table III gives the recognition results with different principal eigenvectors.

TABLE III. TEST ACCURACY

K	10	20	30	40	50
Top1	81.43%	81.45%	81.56%	81.70%	81.04%
Top2	90.79%	90.72%	90.87%	90.96%	90.54%
Top5	96.07%	96.07%	95.94%	96.04%	95.86%

From table III we can see the test accuracy is lower than 82%, and there is a big accuracy difference between top1 and top2, between top2 and top5. This mainly attributes to similar character confusion. We use the LDA-based compound distance in [22] to further identify similar characters. The highest accuracy is about 85%. Obviously, it is very challenging to present new algorithms for higher accuracy. Fig.7 shows the samples misrecognized by MQDF, but corrected by LDA-based compound distance. Fig.8 gives some examples with 5 candidate outputs, where correct results are labeled using red colors. We can see these five candidates are very similar in shape. The misrecognized results can't be corrected using the method in [22].

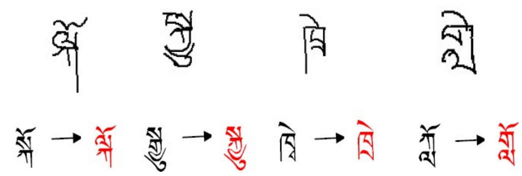


Figure 7. Examples of misrecognized characers corrected by similar character discrimination.

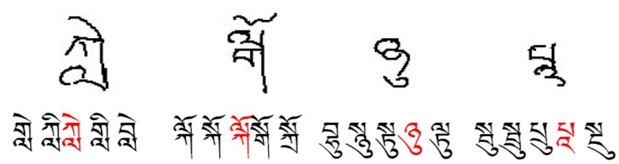


Figure 8. Examples of top 5 candidate outputs.

VI. CONCLUSION

MRG-OHTC, a database of online handwritten Tibetan character was introduced in this paper. To our knowledge, MRG-OHTC is the first publicly available database by now. It contains 910 Tibetan character classes written by 130 different writers. We have evaluated the database using existing algorithms. Evaluation on MRG-OHTC database

provides a benchmark for further research and demonstrates the big challenge of higher recognition accuracy.

However, the scale of MRG-OHTC database is not large enough. We are still collecting more Tibetan character samples. Our purpose of collection is not only for research need, but also promises more substantial application. MRG-OHTC is available for academic researches by contacting with the authors.

ACKNOWLEDGMENT

We would like to thank all the members at MRG group of NFS for helping to supervise data collection. We also would like to thank Yan Sun, Fanqiang Meng and Hanmeng Liu for some preliminary experiments. This work is supported by the CAS Action Plan for the Development of Western China (No.KGCX2-YW-512) and National Science & Technology Major Project (No.2010ZX01036-001-002, 2010ZX01037-001-002).

REFERENCES

[1] 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.

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

[3] 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.

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

[5] 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.

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

[7] 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.

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

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

[10] F. Slimane, R. Ingold, S. Kanoun, A. Alimi, J. Hennebert, A new Arabic printed text image database and evaluation protocols, *Proc. 11th ICDAR*, 946-950, 2009.

[11] 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.

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

[13] 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.

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

[15] C.L. Liu, F. Yin, D.H. Wang, Q.F. Wang, Online and offline handwritten Chinese character recognition: benchmarking on new databases. *Proc 2th CJKPR*, 2010.

[16] B. Liang, W.L. Wang, J. J. Qian, Application of Hidden Markov Model in on-line Recognition of handwritten Tibetan characters (in chinese), *Journal of Microelectronics& Computer,* 26(4): 98-101, 2009.

[17] Y. Sun, H.M. Liu, J.W. Rui, J. Wu, De-noising approach for online handwriting character recognition based on mathematical morphology (in chinese), *Journal of Computer Science,* 36(10): 237-239, 2009.

[18] X.Q. Ding, H. Wang, Multi-font printed Tibetan OCR, *Advance in Pattern Recognition*, pp.73-98, 2007.

[19] Y. Sun, The study on online handwritten Tibetan character recognition (in chinese), *Master Thesis*, 2009.

[20] W.L. Wang, X.Q. Ding, K.Y. Qi, Study on simlitude characters in Tibetan character recognition (in chinese), *Journal of Chinese Information Processing*, 16(4): 60-65, 2002.

[21] F. Kimura, K. Takashina, S.Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and its application to Chinese character recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 9(1):149-153, 1987.

[22] T.F. Gao, C.L. Liu, High accuracy handwritten Chinese character recognition using LDA-based compound distances, *Pattern Recognition*, 41(11): 3442-3451, 2008.