himalayan linguistics

A free refereed web journal and archive devoted to the study of the languages of the Himalayas

Himalayan Linguistics

Online unconstrained handwritten Tibetan character recognition using statistical recognition

Long-Long Ma **Jian Wu**

Chinese Academy of Sciences

ABSTRACT

This paper describes a recognition system for online handwritten Tibetan characters using advanced techniques in character recognition. To eliminate noise points of handwriting trajectories, we introduce a de-noising approach by using dilation, erosion, and thinning operators of mathematical morphology. Selecting appropriate structuring elements, we can clear up large amounts of noise in the glyphs of the character. To enhance the recognition performance, we adopt a three-stage classification strategy, where the top rank output classes by the baseline classifier are re-classified by a similar character discrimination classifier. Experiments have been carried out on two databases MRG-OHTC and IIP-OHTC. Test results show the recognition algorithm employed is effective and can be applied to pen-based mobile devices.

KEYWORDS

online handwritten Tibetan character recognition, de-noising, pre-processing, three-stage classification

This is a contribution from *Himalayan Linguistics, Vol. 15(1): 31–40.*
ISSN 1544-7502

Himalayan Linguistics, Vol. 15(1). © Himalayan Linguistics 2016
ISSN 1544-7502

Online unconstrained handwritten Tibetan character recognition using statistical recognition

Long-Long Ma **Jian Wu**
Chinese Academy of Sciences

1 Introduction

Tibetan is a language with a long history of over 1,300 years. The Tibetan language is still used by more than six million people in China, especially in the Tibet Autonomous Region (Xizang), Yunnan and Qinghai provinces. Research on Tibetan characters, which will facilitate the engagement of Tibetans with modern technologies and enable the digitization of Tibetan documents, is very important both from theoretical and practical perspectives.

Due to the increase of new pen input devices and pen applications, online handwritten character recognition is gaining increasing interest. However, compared to the existing research work on CJK (Chinese, Japanese and Korean) and Arabic, online handwritten Tibetan character recognition (OHTCR) is a relatively unexplored field.

More research focusses on printed Tibetan characters. Ding (2007) designed a novel and effective recognition method for multi-font printed Tibetan OCR. Ngodrup (2010) proposed local self-adaptive binary algorithm and grid-based fuzzy stroke feature extraction to improve recognition accuracy. Kojima et al. (1996) used Euclidean distance with deferential weights to discriminate similar characters. There is comparatively less work on OHTCR. Liang (2009) and Wang (2002) combined HMM (hidden Markov model) based on stroke type with HMM based on the position relation between strokes to improve the recognition performance. Ma (2011) created an online handwritten Tibetan character database named MRG-OHTC and published the database freely for researchers.

This paper describes an online recognition system for handwritten Tibetan characters and reports our experimental results using an approach based on de-noising and a three-stage classification strategy. As for all handwritten recognition problem, handwritten Tibetan character recognition is difficult due to the wide variability of writing styles and the confusion between similar characters. The methods of online character recognition can be roughly grouped into two categories: statistical and structural (see Liu (2004)). Whereas structural matching is more relevant to human learning and perception, statistical methods are more computationally efficient. Taking advantage of learning from samples, statistical methods can give higher recognition accuracies.

We adopt a statistical classification scheme, wherein recognition accuracy depends on the techniques of pre-processing, feature extraction, and classifier design. We use a de-noising approach to eliminate noise points of trajectories (see Sun (2009)). Equidistance re-sampling, smoothing and

nonlinear shape normalization are used in a pre-processing step. The local stroke direction of a character pattern is decomposed into direction maps, which are blurred and sub-sampled to obtain feature values (see Hamanaka (1993) and Zhou (2009)).

The confusion between similar characters is one of the main reasons for lower recognition accuracy. The recently proposed LDA (linear discriminant analysis)-based compound distance (see Gao (2008)) and the critical region analysis based pair discrimination (see Leung (2010)) further improve recognition accuracy. A logistic regression (LR) classifier is used to discriminate confusing characters and costs only small storage for extra parameters, compared to those methods of Gao (2008) and Leung (2010). To discriminate between confusing characters, we use a three-stage classification strategy, similar to the strategy by Zhou (2010) for Japanese. Firstly, candidate classes are selected with a coarse classifier according to the Euclidean distance (ED) to class means. Secondly, fine classification with the modified quadratic discriminant function (MQDF) (see Kimura (1987)) re-orders the candidate classes. Thirdly, the top ranked candidate classes are re-classified by a similar character discrimination classifier. Our experiments on two databases, MRG-OHTC and IIP-OHTC, demonstrate that similar character discrimination classifier can improve the recognition accuracy by about 3%.

The rest of this paper is organized as follows. Section 2 describes the structural characteristic of Tibetan characters. Section 3 gives an overview of the recognition system. Section 4 describes a de-noising approach during pre-processing and section 5 introduces a three-stage classification strategy. Section 6 reports our experimental results and section 7 provides concluding remarks.

2 Structural characteristic of Tibetan characters

The Tibetan script distinguishes four vowels and 30 consonants, which may be called its 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 (see Ding (2007)), whose structure is shown in Fig.1. Each syllable may consist of up to four characters (those parts surrounded by red dash line boundary boxes in Fig.1). The combined characters, which consist of at least an essential consonant (EC, in Tibetan called *ming gzhi*) and may include a top vowel (TV), a consonant above the EC (CaEc, *mgo can*), a consonant below the EC (CbEc, *'dogs can*), and a bottom vowel (BV). Some consonants (in total 20) can serve as a character located to the left of the CC (CbCC, *sngon 'jug*), located or the immediate right of the CC (1-CaCC, *rjes 'jug*), or located two positions to the right of the CC (2-CaCC, *yang 'jug*). Fig.2 gives an example of a four character syllable. In this paper we only consider isolated Tibetan character (SC and CC) recognition.

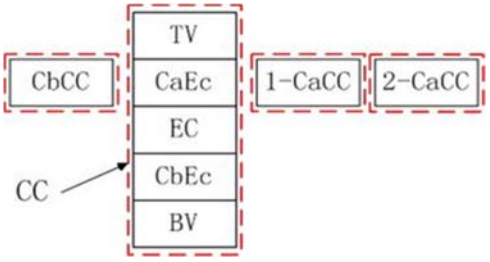

Figure 1. The syllable structure

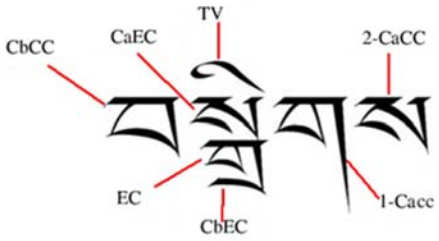

Figure 2. Example of a four-character syllable

3 System overview

The OHTCR system is shown diagrammatically in Fig. 3. The input pattern trajectory is composed of the coordinates of sampled pen-down points. To eliminate noise points and remove certain variations among character samples of the same class, we introduce a de-noising approach at the pre-processing stage. For feature extraction, we use a direction feature extraction method (see Hamanaka (1993)) where the stroke direction is the one in the original pattern, not in the normalized pattern. With the three-stage classification process, we introduce similar character discrimination to reduce the recognition error caused by confusion between similar characters.

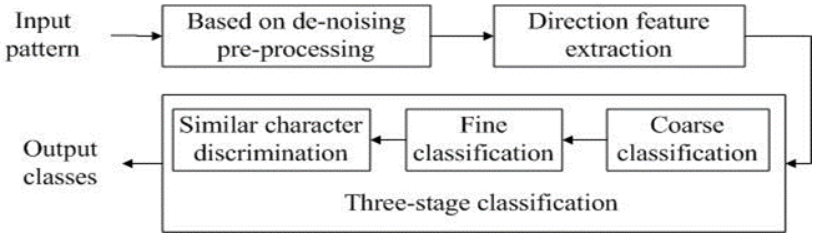

Figure 3. Overview of the

4 De-noising pre-processing

Pre-processing is used to regulate the pattern shape for reducing the class internal shape variation. Nonlinear shape normalization (NSN) is used to normalize shape variability (see Bai (2006)). However, the proportion of noise points in character point trajectories is magnified after NSN. Fig. 4 shows two examples. It is important to eliminate these noises in order to permit these characters to be recognized accurately.

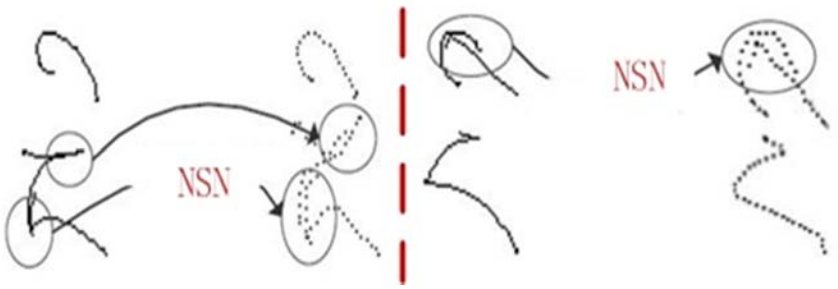

Figure 4. Variation of noises after NSN

To eliminate these noise points, we introduce a de-noising method before NSN. The pre-processing process is illustrated in Fig. 5, where the de-noising step employs the operators (dilation, erosion and thinning) of mathematical morphology. Linear size transformation is to ensure that the character samples of the same class have approximately the same size. Re-sampling is used to reduce distance variation between two adjacent online points. We use Gaussian smoothing (Gaussian blur) to reduce stroke variation in a small local region. We take the image composed of trajectory points as the binary image.

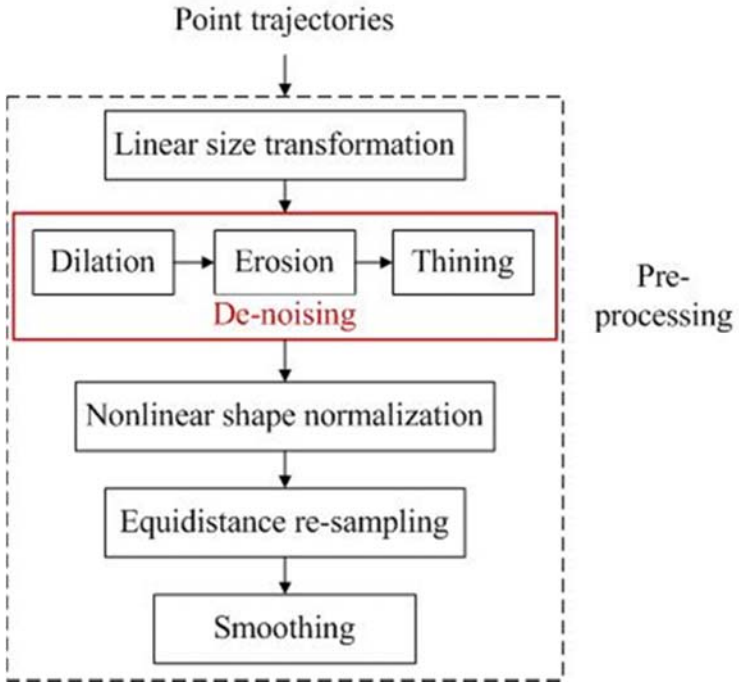

Figure 5. De-noising pre-processing flowchart

4.1 Dilation and erosion

Assuming that A is a region in a binary image I, and that B is a structure element, the dilation of A by B (written $A \oplus B$) is defined as

$$A \oplus B = \{p \mid \hat{B}_p \cap A \neq 0\} \qquad (1)$$

where $\hat{B}$ is the symmetric of B and B_P is the translation of B by the vector p. Under the same assumptions, the erosion of A by B (written $A \ominus B$) is defined as

$$A \ominus B = \{p \mid B_p \subseteq A\} \qquad (2)$$

Using the dilation operation, some useless or re-written strokes can be connected to a component. Unlike offline recognition, online handwriting records stroke direction and point time sequences. According to the point time information, we dilate the binary image. For the erosion operator, we use the same size structuring element (a 3×3 square, with the origin at its center) that is

used for the dilation operator. Fig. 6(a) and 6(b) respectively give the dilation and erosion results of an original pattern.

4.2 Thinning

After dilation and erosion operators, we get the images composed of the strokes with different pixel widths, which is shown in Fig 6(b). In order to regulate the transformed image, we use the thinning algorithm (see Zhang (1984)) to extract a stroke skeleton. Finally the binary image consists of strokes with one pixel width. Compared to the original pattern, we can see the noise points are removed from Fig. 6(c).

(a) Dilation (b) Erosion (c) Thinning (d) Origin pattern

Figure 6. An example of the de-noising method

Fig. 7 gives the transformed image after pre-processing with and without the de-noising method. We can see the image (c) in shape is more similar to the corresponding intended shape.

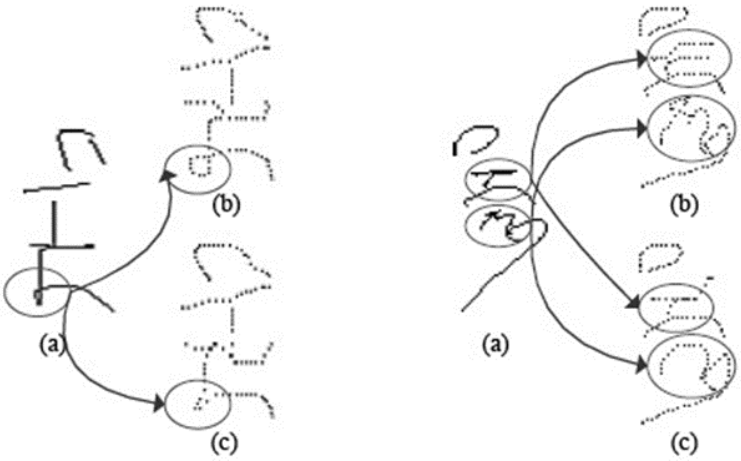

Figure 7. (a) Origin trajectories image (b) the image after pre-processing without de-noising (c) the image after pre-processing with de-noising

5 Three-stage classification

After pre-processing and feature extraction of the input pattern, the feature dimensionality is reduced by LDA. The coarse classifier gives some candidate classes according to the ED from the reduced vector, and the fine classifier (MQDF as the baseline classifier) reorders the candidate classes.

The similar character sets are built on the training dataset by 5-fold cross validation, i.e. rotationally using 4/5 for training the baseline classifier and the remaining 1/5 for validation. We use the selection criterion proposed by Gao (2008) to get the similar character sets.

Assuming the baseline classifier outputs a ranked candidate list. In our experiment, we select the top 5 outputs. If any two of the candidate list belong to one of the similar character sets, we use two-class LDA (see Gao (2008)) to discriminate between them. In LDA, the projection axis *w* for discriminating two classes is estimated to maximize the Fisher criterion. The optimal discriminant vector is represented as

$$\mathbf{w} = \mathbf{S}_w^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) = \boldsymbol{\Sigma}_{ij}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) \tag{3}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are the means of two classes. $\boldsymbol{\Sigma}_{ij}$ is the average covariance matrix of two classes and can be rewritten as

$$\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^{\mathrm{T}} = \sum_{m=1}^{d} \lambda_m \boldsymbol{\psi}_m \boldsymbol{\psi}_m^{\mathrm{T}} \tag{4}$$

where $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_d]$ and $\boldsymbol{\Lambda} = diag[\lambda_1, \lambda_2, \ldots, \lambda_d]$ ($\lambda_1 \geq \lambda_2, \cdots, \lambda_d$). So

$$\mathbf{w} = \boldsymbol{\Sigma}_{ij}^{-1}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) = \boldsymbol{\Psi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Psi}^{T}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) = \sum_{n=1}^{d} \frac{1}{\lambda_n} \boldsymbol{\psi}_n \boldsymbol{\psi}_n^{T}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right) \tag{5}$$

Finally the similar character discriminant function is formulated as

$$f(x) = f_{LDA}(x) = w^T x \tag{6}$$

For the top N output classes from the baseline classifier, we can get at most N×(N-1)/2 classification results after similar character discrimination. The final decision is determined by majority voting.

6 Experimental results

We evaluated the recognition performance on two databases of online handwritten Tibetan characters: MRG-OHTC and IIP-OHTC. The MRG-OHTC, collected by our research group, contains handwritten Tibetan samples of 910 character classes, 130 samples for each character class. We choose the first 105 samples from each class for training, and the remaining 25 samples from each class for testing. The IIP-OHTC database, collected by the Northwest University for Nationalities, contains 562 characters, 150 samples for each character class. We choose 120 samples per class for training and the remaining 30 samples per class for testing.

For each character pattern, we extract a 512-dimensional directional features (see Hamanaka (1993)). The 512-dimensional feature vector is projected onto a 140-dimensional subspace learned by global LDA. The 140-dimensional projected vector is then fed to the MQDF classifier (fine classification), with 40 principal eigenvectors for each class. Table I lists the recognition accuracy for the two databases using the baseline (MQDF) classifier.

From Table I we can see that the test accuracy is lower on the two databases, and there is a large accuracy difference between top1 and top2, and between top2 and top5. The inaccuracy is mainly attributable to the confusion between similar characters. We use a two-class LDA to further identify similar characters. The recognition accuracy improves about 3%. Obviously, it is very challenging to present new algorithms for higher accuracy.

	MRG-OHTC	IIP-OHTC
Top1	81.70%	77.01%
Top2	90.96%	88.99%
Top5	96.04%	95.79%

Table 1. Test Accuracy

Fig.6 shows the samples misrecognized by MQDF, but corrected by two-class LDA discrimination classifier. Fig.8 gives some examples with 5 candidate outputs, where the correct results are labeled in red. We can see these five candidates are very similar in shape. The misrecognized results cannot be corrected using the two-class LDA.

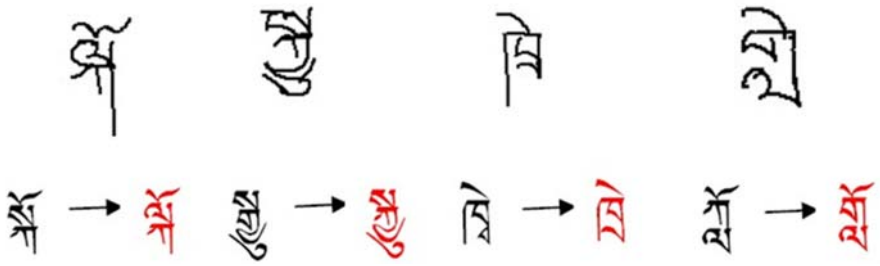

Figure 8. Examples of misrecognized characters corrected by similar character discrimination

Figure 9. Examples of the top 5 candidate outputs

Though the correct rate on two databases is lower, the accumulated recognition rate of the top 10 is higher than 97%. We apply the recognition algorithms to pen-based applications such as mobile phones. Fig.10 shows the interface of our recognition system, where the left sub-window displays the trajectories of handwriting, and the right sub-window gives the top ten recognition results. When the correct recognition result is selected, the character class is surrounded by red bounding box. The bottom sub-window gives the character string with correct outputs.

Figure 10. Interface of the recognition system

7 Conclusion

In this paper, we describe a recognition system for online handwritten Tibetan characters. At the pre-processing step, the de-noising method is used to eliminate the noise points of character trajectories. A three-stage classification strategy reduces the error from the confusion between similar characters. The experiments on MRG-OHTC and IIP-OHTC databases demonstrated the recognition algorithm can be applied as a real recognition system. To further improve the recognition system, we are considering more recognition algorithms and a strategy of combining multiple classifiers.

Acknowledgements

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). The authors would thank Professor Weilan Wang from Northwest University for Nationalities, for providing the IIP-OHTC database.

References

Bai, Zhen-long; and Huo, Qiang. 2006. "A study of nonlinear shape normalization for online handwritten Chinese character recognition: dot density vs. line density equalization". In: *Proceedings of the 18th international conference on pattern recognition (ICPR)*, pp. 921-924.

Ding, Xiaoqing; and Wang, Hua. 2007 "Multi-font printed Tibetan OCR". In: Chaudhuri, Bidyut B. (Ed.) *Advance in pattern recognition*, 73-98. London: Springer.

Gao, Tian-fu; and Liu, Cheng-Lin. 2008. "High accuracy handwritten Chinese character recognition using LDA-based compound distances". *Pattern Recognition* 41.11: 3442-3451.

Hamanaka, M.; Yamada, K.; and Tsukumo, J. 1993. "On-line Japanese character recognition experiments by an off-line method based on normalization-cooperated feature extraction", *Proceedings of the 3rd International Conference Document Analysis and Recognition (ICDAR)*, pp. 204-207.

Kitamura, Fumitaka; Takashina, Kenji; Tsuruoka, Shinji; and Miyake, Yasuji. 1987. "Modified quadratic discriminant functions and its application to Chinese character recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 9.1: 149-153.

Kojima, Masami, Yoshiyuki Kawazoe, Masayuki Kimura. 1996. "Character recognition of wooden blocked Tibetan similar manuscripts by using Euclidean distance with differential weight". *IPSJ SIGNotes Computer and Humanities*, pp.13-18. 小島 正美, 川添 良幸 , 木村 正行 1991 差分重み付ユークリッド距離法による木版刷チベット類似文字認識

Liang, Bi; Wang, Weilan; and Qian, Jianjun. 2009. "Application of Hidden Markov Model in on-line Recognition of handwritten Tibetan characters". *Journal of Microelectronics and Computer* 26.4: 98-101. 梁弼，王维兰，钱建军，基于HMM的分类器在联机手写藏文识别中的应用，微电子学与计算机，26(4): 98-101 (2009).

Liu, Cheng-Lin; Jaeger, Stefan; and Nakagawa, Masaki. 2004 "Online recognition of Chinese characters: the state-of-the-art". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2: 198-213.

Leung, K. C.; and C. H Leung, 2010 "Recognition of handwritten Chinese characters by critical region analysis", *Pattern Recognition*, 43(3): 949-961.

Ma, Longlong; Liu, Hui-dan; and Wu, Jian. 2011. "MRG-OHTC database for online handwritten Tibetan character recognition" In: *Proceedings of the 11th International Conference Document Analysis and Recognition (ICDAR)*, pp.207-211.

Ngodrup; and Zhao, Dongcai. 2010. "Study on printed Tibetan character recognition". In: *Proceedings of Artificial Intelligence and Computational Intelligence (AICI)*, pp. 280-285.

Sun, Yan; Liu, Hanmeng; Rui, Jianwu; and Wu, Jian. 2009. "De-noising approach for online handwriting character recognition based on mathematical morphology". *Journal of Computer Science* 36.10: 237-239. 孙嫣，刘瀚猛，芮建武，吴健，基于数学形态学的联机手写藏文字符识别去噪方法，计算机科学，36(10): 237-239 (2009) .

Sun, Yan. 2009 "The study on online handwritten Tibetan character recognition". MA Thesis, Chinese Academy of Sciences. 孙嫣，藏文联机手写识别若干算法研究，中国科学院研究生院硕士学位论文.

Wang, Weilan; Ding, Xiaoqing; Qi, Kunyu. 2002. "Study on similitude characters in Tibetan character recognition". *Journal of Chinese Information Processing* 16.4: 60-65. 王维兰，丁晓青，祁坤钰，藏文识别中相似字丁的区分研究，中文信息学报，16(4): 60-65 (2002).

Zhang, T. Y.; and Sun, C. Y. 1984 "A fast parallel algorithm for thinning digital pattern". *Communicaition of the Association for Computing Machinery (ACM)* 27.6: 236-239.

Zhou, Xiang-Dong; Wang, Da-Han; Nakagawa, Masaki; and Liu, Cheng-Lin. 2010 "Error reduction by confusing characters discrimination for online handwritten Japanese character recognition". In: *Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 495-450.

Zhou, Xiang-Dong; Liu, Cheng-Lin; and Nakagawa, Masaki 2009 "Online handwritten Japanese character string recognition using conditional random fields". In: *Proceedings of the 10th International Conference Document Analysis and Recognition (ICDAR),* pp. 521-525.

Long-Long Ma
longlong@iscas.ac.cn