Online Chinese Character Recognition System with Handwritten Pinyin Input

Yong GE, Feng-Jun GUO, Li-Xin ZHEN, Qing-Shan CHEN
Motorola Labs, China Research Center, Shanghai, P. R. China
Tel: (8621) 52925800 x 2292, Fax: (8621) 52925782
Email: Yong.GE@motorola.com

Abstract

We have developed a novel online Chinese handwriting recognition system that can recognize a Chinese character either by its handwritten script or by its handwritten Pinyin syllable. The new system is particularly useful when the user forgets how to write the desired character or when the desired character is too complex to be written conveniently. To assure the accuracy and robustness, several classifiers with different characteristics are integrated. The experimental results show that we have achieved an accuracy of 92.5% for 6763-class freely-written Chinese characters and 87.1% for 412-class unconstrained-style Pinyin syllables.

1. Introduction

With the handhold devices being popular, the need of an easier way to enter text messages becomes ever more foremost. When a touch screen is available, online handwriting recognition provides a most natural text input interface, especially for the oriental ideographic character set such as Chinese. Online recognition of Chinese handwritten characters is known as a difficult pattern recognition problem. One difficulty is due to the large vocabulary. For example, a standard GB2312-80 defines 6763 simplified Chinese characters that are used in mainland China, while another standard Big5 defines 13053 traditional characters that are used in Taiwan. Another difficulty is due to the complex shapes of the characters. For example, the stroke number of a character may come up to more than 30, e.g. 齇, 龘.

These difficulties also bring troubles to the end user of a handwriting recognition product. Firstly, remembering all those characters is a tough task even for an educated person. Second, composing a very complex Chinese character is a time-consuming procedure. A handhold device user is usually reluctant to spend much time in text input.

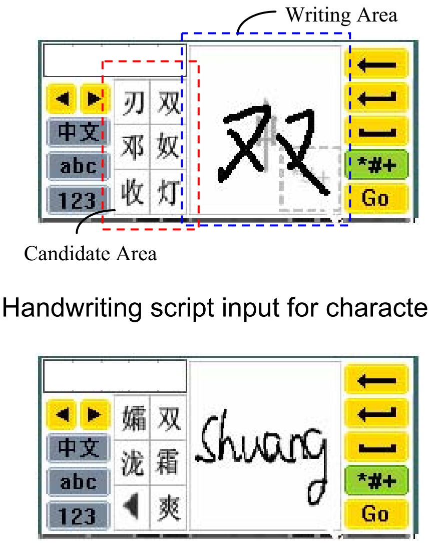
Our company has delivered to the market several high-end PDA phones with online Chinese handwriting recognition as their major text input method. The usability survey from the market shows that most handwriting users will switch to other text input methods such as the Pinyin based input methods when they forget how to write a character or when they feel the character is too complex. Meanwhile they admit that the switch between the two kinds of input methods is quite clumsy because this behavior interrupts the stream of consciousness.

The abovementioned problems drive us to develop a new handwriting recognition system that can take the advantage of the Pinyin based input methods. In mainland China, a Pinyin syllable is actually a systematically arranged letter string that indicates the character's pronunciation. While each character has (at least) one Pinyin syllable, a Pinyin syllable normally maps to several Chinese characters For example, the Chinese character "双" has its Pinyin syllable as "shuang". However, the Pinyin syllable "shuang" maps to several Chinese characters "双", "霜", "爽", "孀", "泷". A Pinyin based input method is such a kind of text input method that utilizes the mapping from a Pinyin syllable to Chinese characters.

In this paper, we have developed a novel recognition system that can recognize a Chinese character by either its handwritten script or its handwritten Pinyin syllable. Figure 1 illustrates our proposed recognition system. If a user writes a Chinese character, as shown in figure 1(a), the recognizer will give the top six recognition results in the candidate area. If the user forgets how to write the desired character or feels that the desired character is too complex, he or she could just write the Pinyin syllable of that character directly in the writing area, as shown in figure 1(b), and then the system would recognize the Pinyin syllable and translate it into Chinese characters. The translated characters are then displayed in a specified order. In this way, the users can input a Chinese character either by the handwritten script or by the handwritten Pinyin syllable.

1520-5263/05 $20.00 © 2005 IEEE

This paper is organized as follows: section 2 generally introduces the architecture of the underlying recognition system; section 3 describes the details of each algorithm module; section 4 presents the experimental results and the conclusion is enclosed in section 5.



(a) Handwriting script input for character “双”

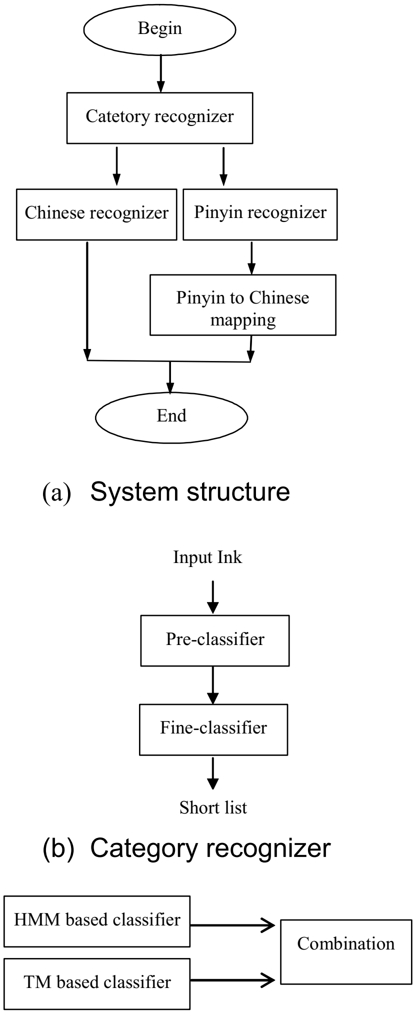
(b) Handwriting Pinyin input for character “双”

Figure 1. Mixed Chinese character and Pinyin syllable recognition system

2. System Architecture

Figure 2(a) shows the structure of the system in this paper. After the system receives the input ink, a category recognizer is applied to determine whether the input ink is a character or a Pinyin syllable. If the input is determined as a character, the system invokes the Chinese character recognizer to give the final recognition result. If the received input is determined as a Pinyin syllable, the system invokes the Pinyin syllable recognizer and then maps the first Pinyin candidate to the Chinese characters. The mapped Chinese characters are then sorted in the order of character frequency to form the final recognition result.

The category recognizer plays an important role in our system. Besides classifying the input ink to Chinese character or Pinyin syllable, this recognizer gives a short candidate list once the category is determined. As shown in figure 2(b), this recognizer is composed of two classifiers: a pre-classifier followed by a fine-classifier. The pre-classifier uses a fast algorithm (detailed in section 3.1) to output a candidate set which contains up to 1000 candidates. Then the fine-classifier uses an accurate algorithm to re-order the candidate set. The category of the input ink is determined by the category of the first candidate of the re-ordered set. Finally, we only keep the top N candidates of the same category, where N=100 for Chinese character and N=10 for Pinyin syllable. This fine-classifier is actually a template matching based classifier, which will be described in section 3.2.



(a) System structure

(b) Category recognizer

(c) Chinese/Pinyin recognizer

Figure 2. System architecture

As shown in figure 2(c), both the Chinese character and the Pinyin syllable recognizer are composed of two classifiers. The first classifier, a hidden Markov model (HMM) based classifier, is used to capture the temporal information. The second one, a template matching (TM) based classifier, is used to capture the shape information. The outputs of the two classifiers are combined to give the final decision. The combination method is suggested in [1] [2].

3. Classifiers

3. 1. Pre-classifier

In the pre-classifier, the input ink is firstly converted to a binary image. The image is then nonlinearly normalized to compensate the shape distortion [3]. From the nonlinearly normalized image, we extract a 16-dimensional regional stroke count (RSC) feature vector as described in [4]. In detail, let $\{f(x,y)|_{x,y=0}^{W-1,H-1}\}$ be the nonlinearly normalized image, where W and H are the image width and height respectively, $f(x,y)$ takes the value of either 1 or 0. The 16-dimensional RSC feature vector is actually the combination of the 8-dimensional horizontal RSC feature vector (HRSC) and the 8-dimensional vertical RSC feature vector (VRSC). To extract the HRSC feature vector, the image is uniformly separated into 8 horizontal rectangle regions. From *i*-th horizontal rectangle, we compute *i*-th horizontal RSC (HRSC) feature Γ_{HRSC}^{i} by:

$$\Gamma_{HRSC}^{i} = \sum_{y=0}^{H-1} \sum_{x=(i-1)W/8}^{iW/8-1} f(x,y)\bar{f}(x+1,y), \quad i = 1,2,...8$$

where $\bar{f}(x,y)$ is the logical negation of $f(x,y)$.

To extract the VRSC feature vector, the image is uniformly separated in to 8 vertical rectangle regions. From *i*-th vertical rectangle, *i*-th vertical RSC (VRSC) feature Γ_{VRSC}^{i} is computed by:

$$\Gamma_{VRSC}^{i} = \sum_{x=0}^{W-1} \sum_{y=(i-1)H/8}^{iH/8-1} f(x,y)\bar{f}(x,y+1), \quad i = 1,2,...8$$

For each character class, we compute the class center by averaging the RSC feature vectors of all the samples in the training set. After that, we cluster all the class centers into G groups by using LBG [5] clustering algorithm. Each group has a bucket consisting of dozens of character classes with their centers belonging to that group, as shown in figure 3. The group center is calculated as the mean of the class centers that belong to that group.

In the recognition step, we calculate the city block distance between the input RSC feature vector with each group center and sort the results in ascending order. Then we copy to the candidate set the content of the bucket in turn until the number of candidates reaches a preset value. A similar implementation could be found in [6].

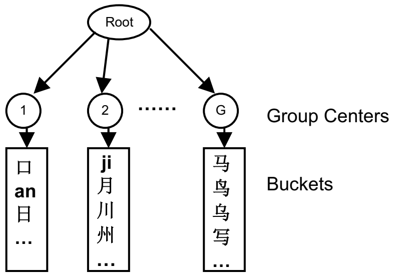


Figure 3. The pre-classifier

3. 2. Template Matching Based Classifier

We choose to use the *directional* features in this classifier. To extract these features, the nonlinearly normalized image is uniformly divided into $m \times n$ overlapped grids. An 8-dimensional feature vector is extracted from each grid, to characterize the image's local directions. For the whole image, we can derive an $m \times n \times 8$ dimensional directional feature vector. Readers are referred to [7] for the details.

In the training step, we firstly use LBG algorithm to generate several initial templates for each class. Each template is an $m \times n \times 8$ dimensional vector. For Pinyin syllable, we generate only one template per class. For Chinese character, the template number per class varies from 1 to *K* depending on the writing variability calculated from the training set and the accessing frequency calculated from another large language corpus. Furthermore, a minimum classify-cation error (MCE) training algorithm is applied to improve the recognition accuracy [8]. In the pattern classification step, we compute the city block distance between the input directional feature vector with each template and sort the distances in ascending order. The shortest N distances and their correspondent templates form the classifier's output.

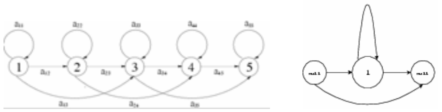
The TM based classifier is used in several recognizers in our system. The category recognizer and the Chinese recognizer share one common TM classifier. In this shared classifier, we use 7×7 grids to extract the directional features. The Pinyin recognizer uses another TM based classifier. In that classifier, 14×7 grids are applied, which improves the accuracy of the Pinyin recognizer about 8% compared with the 7×7 grids.

3. 3. HMM Based Classifier

The HMM based classifiers are used in both the Chinese recognizer and the Pinyin recognizer. In the training step, we create a Gaussian-mixture density, left-to-right HMM for each modeling unit. Figure 4(a) shows an HMM example in our system. In the pattern classification step, the Viterbi algorithm [9] is applied to find the N-best matching scores as well as the N-best decoding sequences. Since the Chinese characters are ideographic symbols while Pinyin syllables are essentially letter strings, the detailed HMM implement-tations for the two categories are quite different. We will detail the differences in the following sections.

3.3.1. HMM for Chinese Character

In the Chinese recognizer, the input ink is firstly smoothed and normalized to a standard size. The preprocessed ink is then divided into a sequence of line segments. This segmentation is based on the local maximum of directional angle change [10]. A two-dimensional feature vector $o = (dx, dy)$ is extracted for each segment, where *dx* and *dy* are the coordinate differences between the starting and the ending points of the segment. The segment connecting the two consecutive strokes are called *imaginary* segment, from which we extract the same feature vectors.



(a) (b)

Figure 4. HMM Topology and ligature model

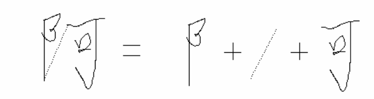


Figure 5. Character "阿" is composed of "阝" + λ^l + "可", "阝" and "可" are radicals, λ^l is a ligature

Among many possibilities, we choose *radical* as the basic modeling unit. A radical in Chinese is a stable pattern that could be shared by many characters. We create one HMM for each radical. The state number for each model is equal to the average segment number that is obtained from a bootstrap radical database. A special HMM is also created to model the ligatures connecting the two consecutive radicals, as shown in figure 4(b). Different from the radical models, the ligature model has only one emitting state, and could be skipped during the Viterbi decoding. By using this modeling method, a whole Chinese character could be represented by a sequence of radical models that are separated by the ligature models. For example, the character "阿" is represented by two radical models "阝" and "可" with a ligature model λ^l, as shown in figure 5.

3.3.1. HMM for Pinyin Syllable

In the Pinyin recognizer, the preprocessed ink is transformed into a sequence of fundamental short segments based on features such as curvature extrema [11]. Normally the segment is shorter than its counterpart used in the Chinese recognizer because the strokes in Pinyin syllable are more curving. These short segments are parameterized as 9-dimensional vectors of size-independent features (e.g. start and end angle, relative to the x-axis), and the resulting sequence of non-overlapping segment vectors is used as the input feature.

In the modeling step, we choose allograph as the basic modeling unit [12]. Allograph represents the different ways of writing a letter. The allograph models are generated manually or through clustering techniques. Similar to the radical HMM, the state number of each allograph HMM is equal to the average segment number that is obtained from a bootstrap Pinyin database. Since a letter might have several allographs, a Pinyin syllable is represented by a finite state network (FSN) of allographs. Figure 6 shows an FSN example for Pinyin "ang".

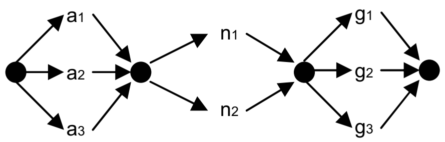


Figure 6. FSN representation of Pinyin "ang"

4. Experimental Result

To evaluate the proposed approach, a series of experiments are conducted in the handwriting database collected by our lab. The database covers 6763-class Chinese characters and 412-class tone-free Pinyin syllables. Each Chinese character class consists of more than 500 freely-written samples, of which 400 samples are used for training and the left for testing. Each Pinyin syllable class consists of more than 200 unconstraint-style samples, of which 100 samples are used for training and the left for testing.

Table 1 shows the hit rates of the category recognizer. For the Chinese character, about 99.5% of the testing samples are classified to the right category, while for the Pinyin syllable, only 94.9% of the testing samples are classified correctly. In fact we can balance the two hit rates if we adjust the model complexity of the two categories in the fine-classifier. That is, if we increase the templates for each Pinyin syllable and reduce the templates for each Chinese character, the hit rate for Pinyin syllable will increase while for Chinese character it will drop. We consider that in practice, the user would probably write much more Chinese characters than Pinyin syllables, so we must guarantee a higher hit rate for the Chinese characters but compromise at the hit rate for the Pinyin syllable.

Table 1. The hit rates of the category recognizer in the testing set

	Pinyin syllable	Chinese characters
Hit rate	94.9%	99.5%

Table 2 shows the accuracies of the two individual recognizers. The accuracy of the Chinese character recognizer is 92.5% and the accuracy of the Pinyin syllable is 87.1%. Although we cannot compare our result with other studies due to the lack of a standard database, a usability survey conducted by the product group shows that the end users are quite satisfied with the performance of our system.

Table 2. The accuracies of the Pinyin recognizer and the Chinese recognizer in the testing set

	Pinyin syllables	Chinese characters
Accuracy	87.1%	92.5%

5. Conclusion and Future Work

In this paper, we presented a novel online Chinese recognition system with the capability of handwritten Pinyin input. Currently our system can only recognize 412-class Pinyin syllables. In future, we will extend our system to support more Pinyin based input methods, such as tonal Pinyin input method, Pinyin phrase input method, etc.

From another perspective, this paper presented a system with the capability of recognizing both English and Chinese handwriting, although the English vocabulary is now limited within 412-class Pinyin syllables. By applying a more flexible and powerful category recognizer, we can implement a real online Chinese/English bilingual handwriting recognition system in future.

Reference

[1] O. Venek, S. Jaeger, M. Nakagawa, “A New Warping Technique for Normalizing Likelihood of Multiple Classifiers and Its Effectiveness in Combined On-line/Off-line Japanese Character Recognition”, *IWFHR 2002*, pp. 177-182.

[2] J. Zhen, X. Ding, Y. Wu, “Dynamic Combination of Multi-classifier Based on Minimum Cost Criterion. *Chinese J. Computers*, 2 (1999) pp. 182-187.

[3] S.W. Lee, J.S Park, “Nonlinear Shape Normalization Methods for The Recognition of Large-Set Handwritten Characters”, *Pattern Recognition*, Vol.27, No.7, pp.895-902.

[4] Y. Y. Tang, L. T. Tu, S. W. Lee, W. W. Lin, I. S. Shyu, “Offline Recognition of Chinese Handwriting by Multifeature and Multilevel Classification”, *IEEE Trans. PAMI*, 1998,5(20), pp. 556-561.

[5] R. Duda, P. Hart, D. Stork, *Pattern Classification and Scene Analysis*, John Wiley & Sons Inc. 2001.

[6] Z. Feng, Q. Huo, “Confidence Guided Progressive Search and Fast Match Techniques for High Performance Chinese/English OCR”, *ICASSP 2002*, Vol. 2, pp. 89-92.

[7] N. Kato, S. Omachi, H. Aso, Y. Nemoto, “A Handwritten Character Recognition System Using Directional Element Feature and Asymmetric Mahalanobis Distance”, *IEEE PAMI* 1999, 21(3), pp. 258-262.

[8] Q. Huo, Y. Ge, “High Performance Chinese OCR Based on Gabor Features, Discriminative Feature Extraction and Model Training”, *ICASSP 2001*, pp. 1517-1520.

[9] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of IEEE*, Vol. 77, No. 2, pp.257-286, 1989.

[10] K. Parthasarathy, “Method and Apparatus for Character Recognition of Handwritten Input”, *United States Patent*, Appl. No. 919,875.

[11] G. Seni, T. Anastasakos, “Non-Cumulative Character Scoring in a Forward Search for Online Handwriting Recognition”, *ICASSP 2000*, pp. 3450-3453.

[12] A. Biem, “Minimum Classification Error Training for Online Handwritten Word Recognition”, *IWFHR 2002*, pp. 61-65.