SCUT-COUCH2008: A Comprehensive Online Unconstrained Chinese Handwriting Dataset

Yunyang Li, Lianwen Jin, Xinhua Zhu, Teng Long*

Department of Electronics and Information Engineering,
South China University of Technology
llyy26@163.com, *eelwjin@scut.edu.cn , vcpudding@gmail.com, tenglong@scut.edu.cn

Abstract

A comprehensive online unconstrained Chinese handwriting dataset, SCUT-COUCH2008, is built to facilitate the research of unconstrained online Chinese handwriting recognition. SCUT-COUCH2008 is a comprehensive dataset that consists of 8 subsets, namely SCUT-COUCH-Word, GB1, GB2, TradGB1, Pinyin, Letters, Digit and Symbol. Particularly, each complete set of SCUT-COUCH2008 contains samples of 8,888 frequently used Chinese words, 6,069 single Chinese characters, 2,010 Pinyin, and 184 other characters, totally 17,151 object units and 27,858 characters. The current version of SCUT-HLC2008 is collected with PDA's, including 50 complete sets of samples and 1,392,900 characters in total, written independently by 50 different persons. SCUT-COUCH2008 dataset is the first public available large vocabulary online Chinese handwriting dataset that contains frequently used Chinese words and Pinyin. It provides basic dateset for new research topics such as online handwritten Chinese word and Pinyin recognition.

Keywords: SCUT-COUCH, online Chinese handwritten dataset, handwritten Chinese word, handwritten Chinese Pinyin

1. Introduction

Online handwriting Chinese character recognition (OHCCR) is attracting more and more attention among researchers in recent years[1]. A comprehensive and unconstrained online handwritten Chinese dataset is gaining increasing importance. SCUT-COUCH2008[1] is proposed for the following reasons. First, it's going to provide researchers in the field of online handwritten Chinese recognition with a refined online Chinese dataset as training and testing samples. Second, it's also a standard dataset for comparing and evaluating performances of different algorithms.

During the last twenty years, numbers of handwriting datasets have been published in the literature. To name a few, there're English datasets CENPARMI[2] and CEDER[3]; French dataset IRONOFF[4]; Indian dataset ISI[5]; Japanese datasets Kuchibue and Nakayosi[6]. For offline handwritten Chinese datasets, there're ETL-8/ETL-9[7], IAAS-4M[8], HCL2000[9] and HIT-MW[10], giving powerful promotion to the rapid development of Chinese character recognition.

Briefly speaking, the development of handwriting datasets implicates some tendency. First, the collection of categories develops from a small range into a large one. Second, the scale of sampling grows from single characters to paragraphs. Third, the manner of handwriting styles is changed from regular to cursive and unconstrained. Although there are already exists many offline handwriting Chinese dataset, large scale public available online handwriting Chinese corpus are seldom reported in the field of OHCCR. Particularly, to our best knowledge, there is no any online handwriting Chinese word and Pinyin dataset publicly available yet.

Comparing with the datasets referenced above, our SCUT-COUCH2008 dataset has the following characteristics:

Firstly, SCUT-COUCH2008 is a comprehensive dataset composed of 8 subsets as shown in Table 1.

Table 1. 8 Subsets in SCUT-COUCH2008

Subsets	Detail
Word	8,888 frequently used Chinese words
GB1	3,755 characters in GB Set1
GB2	3,008 characters in GB Set2[2]
TradGB1	1,384 traditional characters in GB Set1
Pinyin	2,010 Pinyin
Letter	52 English upper- and lower- case alphabets
Digit	10 numeric digits
Symbol	122 frequently used symbols

Secondly, it's the first public available Chinese handwriting dataset that includes words as collecting

[1] SCUT is the abbreviation of South China University of Technology, and COUCH is the abbreviation of Comprehensive Online Unconstrained Chinese Handwriting.

[2] GB Set1 and GB Set2 are the abbreviations of the Level 1 and Level 2 characters of Chinese GB2312-80 standard (total 6763 characters), respectively.

objects. Currently available Chinese handwriting datasets are samples of either single characters or full-length paragraphs. The lack of dataset of Chinese words binds Chinese handwriting recognition to the level of single character recognition, leaving Chinese word recognition hardly addressed in the literature.

Thirdly, it's the first public available online handwriting dataset that covers Chinese Pinyin collections. Frequently enough in practical usage, people remember the pronunciations of some characters other than their shapes. A handwriting Pinyin input method will be much more convenient upon these situations. Similar solutions have been addressed recently[11]. An appropriate and unconstrained dataset of Pinyin is needed to train a robust Pinyin recognition engine.

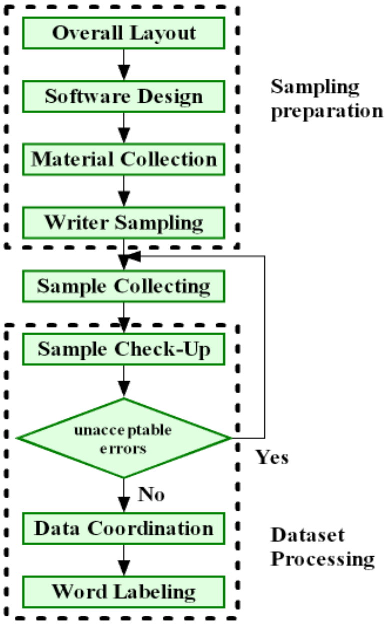


Figure 1. The flowchart of SCUT-COUCH2008.

The flowchart of building SCUT-COUCH2008 is shown in Figure 1. The rest of this paper is arranged as follows. The next section describes the sampling preparation. Then the dataset processing is introduced in section 3. Section 4 analyses the basic statistics of the dataset. Concluding remarks are given in section 5 finally.

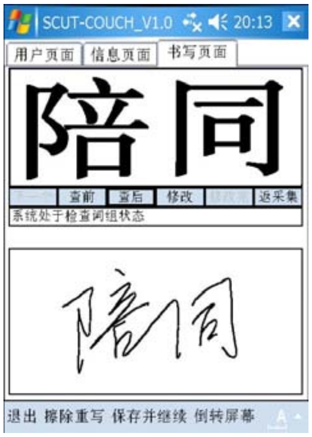
2. Sampling Preparation

As we deeply understand, building a comprehensive online handwritten Chinese dataset means great amount of work. Since, every stage in the collection of SCUT-COUCH2008 dataset, from the selection of material and sampling devices, the sampling of writers, to the establishment of sampling rules, is deliberately considered over. In this section, we will introduce how we accomplished full preparation on the pre-sampling stage.

2.1. Program Design

After analyzing and comparing currently available devices for sampling online handwritings, we finally decided to employ PDA's (Personal Digital Assistant), instead of trajectory recording pens or tablets, to be the sampling devices , for PDA's are more portable and universally used as handwriting input devices in daily life.

As illustrated in Figure 2, in our program, the input area is designed to be a box of static size, large enough to accommodate as many characters as possibly required. In addiction, noticing that most characters have their last strokes end in the bottom-right corners of the characters, we place the "Save and Continue" button on the bottom of the input box and thus makes it convenient for collectors to proceed to the next object quickly after finishing writing a sample.



系统处于检查词组状态

Figure 2. An illustration of layout.

2.2. Collection of Sampling Objects

There's always a specific strategy in sampling objects of a handwriting dataset. For example, HCL2000 took the 3755 frequently used Chinese characters of GB Set1 as sampling objects, while HIT-MW adopted various kinds of paragraphs in China Daily. We also have a strict principle in choosing objects of SCUT-COUCH2008.

2.2.1. Words Picking

The vocabulary of Chinese words is extremely huge. It would be impossible to collect all these words. Therefore, we pick word objects according to the frequencies of their usage. We employed the statistics on Chinese words' usage frequencies published by "Sogou Labs" [12]. The frequencies of appearance of 157,202 words in more than 100 million web pages are counted in the statistics. In our analysis, we found words frequently appear on the internet are also continually used in our daily life. Therefore, we intercepted the 8,888 words of highest usage frequencies to be word sampling objects, for the following reasons. First, we should not make the collection work too overloaded by adopting too large an amount of sample objects. Upon such consideration, 19,595 single Chinese characters are involved in the 8,888 frequently-used words, which is an appropriate amount. Second, we should not lose universal coverage in daily usage by adopting too small an amount of sample objects. As we can see in Figure 3, these 8,888 words cover 71.48% of the daily-used lexicon. Third, we picked 8,888 words because 8,888 is considered a lucky number in China and makes it easy for everyone to remember.

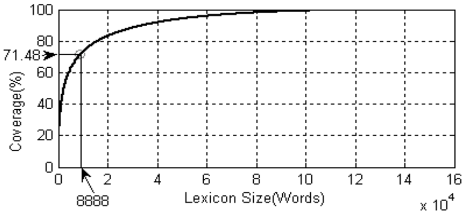


Figure 3. Words usage frequency distribution.

Particularly, the 8,888 words we chose involve 7,499 two-character-words, 984 three-character-words, 383 four-character-words, 20 five-character words, 1 six- and seven- words respectively (see Figure 4). These words are composed of totally 19,595 single characters, covering 2,078 characters among the 3,755 characters in GB Set1, with only 26 characters belonging to GB Set2, which also indicates that characters in GB Set1 are the most frequently used ones in daily life.

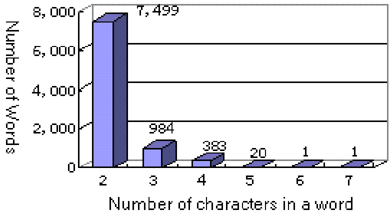


Figure 4. Statistics of lengths of 8,888 words

2.2.2. Single Characters Picking

To facilitate a full-range research on handwriting recognition, we must also cover all of the frequently used single characters in our dataset in addition to the 8,888 high-frequency Chinese words. Considering that single characters can be segmented from words (see Section 3.2), we only add 1,677 characters in GB Set1 that are missing in the 8,888 words and all of 3,008 characters in GB Set2. These single characters, adding to the sum of 6,763, are covering the whole set of GB2312-80.

Furthermore, noticing that traditional Chinese characters are also used occasionally, an appropriate range of traditional Chinese characters are also listed in our dataset. As 99.9% of daily used characters are covered within the GB Set1 [13], the set of traditional Chinese character samples is obtained by the transformation from 3,755 simplified Chinese characters in GB Set1 to their traditional forms. In particular, different traditional forms may be found in different lexicons in correspondence to a same simplified character (see Table 2). Finally, 1,384 traditional Chinese characters are collected as the merge of the sets found in different lexicons (including GBK and GB18030-2000).

Table 2. Different traditional characters correspond to the same simplified character

Simplified character	Traditional characters
干	幹乾
艳	豔艶艷
历	歷曆

2.2.3. Selection of Pinyin Material

In the pronunciation of Mandarin, every kind of spells may have as many as 5 different tones including the surd, such as shi(匙), shī(尸), shí(石), shǐ (史) and shì(试). Although, it's more often the case that one kind of spell has only 4 or less tones with correspondent Chinese characters. For example, no character exists relating to the pronunciation lié.

Through analyzing pronunciations of the lexicon of GB18030-2000, we came to a collection of Pinyin involving 415 kinds of spells (regardless of tones), 1,426 enunciations (considering tones). After excluding some spells that are rarely used (such as hng, hm, ng, etc), a lexicon of 402 kinds of spells is obtained. Then all 5 tones of every spell are included into our dataset (2,010 in total), regardless of whether or not they really have a correspondent Chinese character, out of the following concerns. First, this strategy results in five samples for every spell, and thus increases the stability in statistics of handwritten spells. Second, it facilitates the analysis on the differences among different tones of the same spell. Third, being a developing language, Chinese pronunciations may change along with the emergence of new characters.

2.2.4. Selection of Other Material

Aside from materials mentioned above, we also preserve the involvement of some regular materials, including 10 numeric digits, 52 upper- and lower- case alphabets and 122 frequently used symbols (see Figure 5).

！＂＃￥％＆＇（）＊＋，－．／：；＜＝＞？＠
［＼］＾＿｀｛｜｝￣ 《》 ＄￥÷ΠΣΦΨΩαβγδεζη
θλμξπρστφψω Ⅰ ⅡⅢⅣⅤⅥⅦⅧⅨⅩⅪⅫ ⅰ ⅱⅲⅳ
ⅴ ⅵⅶⅷⅸⅹ ←↑→↓↖↗↘↙∈√∞∫∮∴∵≈≌
≠≤≥ ⊕⊙±①②③④⑤⑥⑦⑧⑨⑩£、 ‘’ “”。

Figure 5. Symbols in SCUT-COUCH2008.

2.3. Writer Sampling

Generally, our strategy in selecting candidate writers is combining deliberation with randomness.

Deliberately, we selected our first group of writers from students in our school (South China University of

Technology), for the following reasons. On one hand, college students are enrolled from all over the country, so that their handwriting samples may be considered as samples written by users from different regions of the country. On the other hand, well-educated people, such as students and teachers in colleges, are most-likely potential users of handwriting recognition devices such as PDAs.

With the general range of candidate writers confirmed, we start a random selection of writers. As shown in the following tables (Table 3~5), the randomly employed writers indicate a balanced distribution on region, age and sex.

Table 3. Sampling percentage of three regions

Region	Proportion
South Region	48%
Middle Region	40%
North Region	12%

Table 4. Age distributions of writers

Items	Proportion
Below 18	4%
Between 19 and 35	88%
Between 36 and 50	6%
Older than 50	2%

Table 5. Gender distributions of writers

Items	Proportion
Male	56%
Female	44%

2.4. Policies of Pattern Collection

In order to collect patterns useful for developing powerful on-line handwriting recognizers in practical applications, the following policies were employed in the collecting procedure:

(1) So far all 50 complete sets of samples are collected individually by 50 writers, so as to ensure the stability in writing styles. In addition, among the 200 sets of samples in our plan, 30 sets of them will be collected by multiple writers cooperatively, which are useful in researches on writer-independent recognition.

(2) Writers are asked to input handwritings in the input box successively without knowing the content of the next object.

(3) Writers are aware that every object is an integral entity rather than a simple combination of multiple single characters, so as to preserve integrity of each collected sample and the dataset ’ s availability in researches that take them as entities.

(4) Writers are asked to perform handwritings in their most comfortable and familiar manner. No restrictions are imposed to writers on handwritings’ quality.

(5) Writers are ignorant about the dataset’s potential usage in Chinese character recognition, so as to avoid too regular handwritings from being produced deliberately and to ensure the unconstrained manner of collected samples.

(6) A continuous duration of each sampling is confined to 1.5 hours, not exceeding 2 hours at most, in order to prevent samples from being distorted due to exhaustion.

3. Dataset Processing

With 50 sets of data completely sampled by 50 writers individually, we started to check and pack up the collected data. A helpful word labeling has also been performed, which will be discussed in this section.

3.1. Error Correction

While checking the collected data, we found few samples with unacceptable errors. As Figure 6 shown, some samples involve sudden excursion on the trajectories, caused by occasional malfunctions in the hardware of touch screens.On the other hand, errors as shown in Figure 7 are not caused by hardware but subjective mistakes, which are also unbeneficial for analyzing the disciplines of Chinese handwriting.

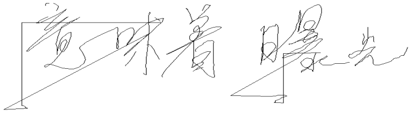


Figure 6. Unacceptable error caused by hardware.

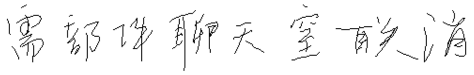


Figure 7. Writing mistake samples (the correct words should be 零部件, 聊天室 and 取消).

All error samples referenced above are required to be re-collected. Meanwhile we preserve these error samples in case of special research needs.

3.2. Dataset Labeling

To make the handwritten word samples available for segmentation-based researches, we labeled the 8,888 Chinese word samples in the collected data sets. Our method is to record the bounding box of every single character as well as the indices of strokes in each character. The labeling of word samples is meaningful for three reasons. First, it provides available data for conducting researches on Chinese character segmentation. Second, all single characters are separated and hence enrich the samples of isolated characters. Last but not the least, it offers information on both spatial and temporal relationships among character members of a Chinese word.

All the labeling of words is conducted on PC’s using a semi-automatic tool. Figure 8 shows some of the labeled word samples.

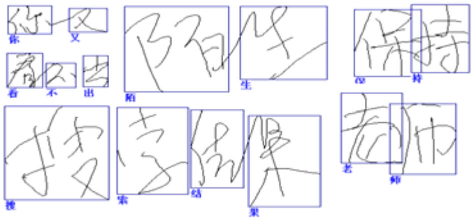


Figure 8. Word segmentation and labeling.

4. Data Analysis

The SCUT-COUCH2008 dataset is not only the first public available online unconstrained Chinese handwriting dataset, but also innovatively comprehends all different sample objects, which is useful for online Chinese handwriting recognition. Figure 9 shows the different kind of samples we collected. In this section, we are going to briefly introduce both statistical and individual characteristics of our dataset.

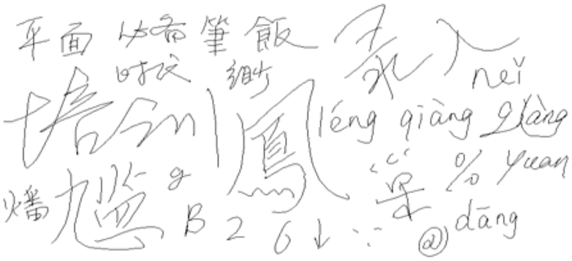


Figure 9. Some samples in our dataset.

4.1. Statistical Disciplines of Strokes

In this section, statistics of stroke number is analyzed in the range of 8,888 words in 50 sets of samples. As it's shown in Figure 10, we came to the following conclusions. First, many Chinese characters are written cursively, with their strokes connected. Second, some Chinese words are even cursively written in one stroke. Third, stroke number of every word varies from sample to sample, instead of being identical with the standard. Fourth, the ruleless number of strokes also reflects the original writing styles of different writers, giving proofs to the unconstrained characteristic of SCUT-COUTH dataset.

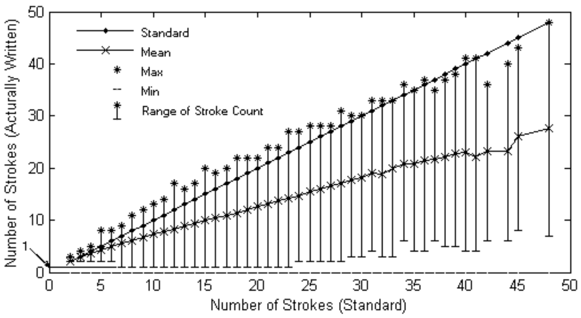


Figure 10. Statistics on words stroke numbers.

4.2. Handwriting Diversities in SCUT-COUCH

Every writer has his or her own writing style. Even if written by the same person, the shape of a character may differ from time to time. Moreover, as different types of sampling devices are used in the collection, data of rich diversities are presented in our collected data. The followings describe some special characteristics of our dateset.

Different styles written by the same person: different styles may be seen in handwritings of the same person, as figure 11shows. A and B, C and D are written by the same writers respectively.

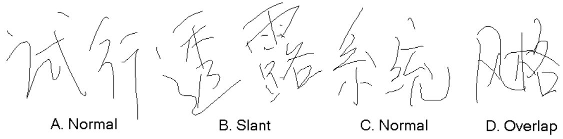


Figure 11. Different styles written by the same person.

Aliasing: Aliasing may sometimes appear on writing trajectories due to limited resolutions of touch screens. (See Figure 12)

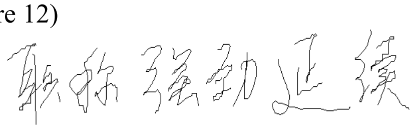


Figure 12. Examples of aliasing.

Disconnected stroke: It's the situation that a stroke is disconnected in its trajectory (see Figure 13), caused by a sudden failure in capturing part of the trajectory when the stroke is written too fast.

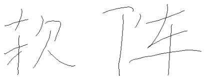


Figure 13. Disconnected stroke (characters written in the above figure:软 and 阵).

Redundant stroke: Some writers are used to make a point when finish writing a sample word. Also a redundant stroke may be produced by carelessly touching the touch screen (see Figure 14).

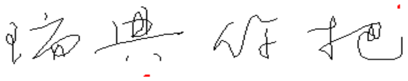


Figure 14. Redundant stroke.

Missing stroke: Some small strokes are missed out in fast writing (see Figure 15).

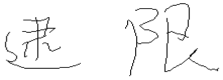


Figure 15. Missing strokes (characters written in the above figure: 逮 and 限).

Over-connected stroke: in the writing process, even some strokes in different characters are written connectively (see Figure 16).

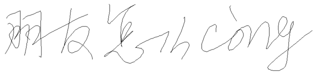


Figure 16. Over-connected strokes.

Mis-written traditional Chinese characters: When collecting samples of simplified Chinese characters, some writers mistakenly wrote them in the traditional forms following their habits (see Figure 17).

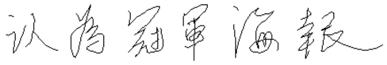


Figure 17. Mis-written traditional Chinese characters.

Rewritten stroke: Occasionally, when writers found some strokes to be bad-looking or disconnected, they would rewrite these strokes instead of erasing the whole unit (see Figure 18).

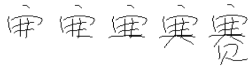


Figure 18. Rewritten stroke(the character '赛').

Styles variation of different writers: Being a flexible form of art itself, Chinese calligraphy may have massive styles for a same character. Figure 19 gives the examples of different styles of some characters.

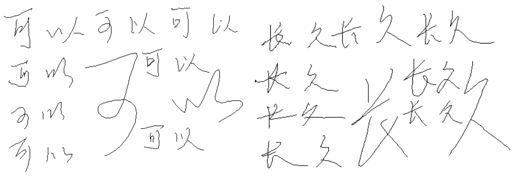


Figure 19. Varied styles of different writers.

5. Conclusion

The multi-type comprehensive online unconstrained Chinese handwriting database SCUT-COUCH2008, has important features that are missing in other datasets. It's the first public available online handwritten Chinese character dataset that involves 8,888 Chinese words and 2,010 Chinese Pinyin.

The collection of COUCH is conducted under the supervision of deliberately designed strategies. From the selection of material and sampling devices, the sampling of writers, to the establishment of sampling rules, are elaborately designed.

SCUT-COUCH was originated for the purpose of providing training and testing samples for researches of online handwritten Chinese word recognition. However, it promises much more substantial applications than it's initially designed. A number of novel research topics could be promoted under the assistance of our dataset, such as handwritten Pinyin input method, online handwritten Chinese word segmentation, segmentation-free recognition, etc.

Parts of the SCUT-COUCH2008 dataset and its latest detailed information are available at http://www.hcii-lab.net/data/SCUTCOUCH/. More information is also available upon request (Contact eelwjin@scut.edu.cn or scutcouch@gmail.com).

Acknowledgments

We would like to thank Kai Ding, ZhiBin Huang, Bin Zhang, Shengming Huang, Xuewen Liu, Guoqiang Deng, etc. for helping to supervise data collection. We would also like to thank all those warmly cooperative volunteers. This work is supported in part by the research fundings of NSFC (no. U0735004, 60772216) and GDNSF (no. 07118074).

6. References

[1] C.L. Liu, S. Jaeger, M. Nakagawa, "Online recognition of Chinese characters: the state-of-the-art", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume 26, Issue 2, pp.198 – 213, Feb 2004.

[2] C. Y. Suen, C. Nadal, R. Legault, T. A. Mai, L. Lam, "Computer recognition of unconstrained handwritten numerals", *Proceedings of the IEEE,* Vol. 80, No. 7, pp. 1162-1180, 1992.

[3] J. Hull, "A database for handwritten text recognition research", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 16, No. 5, pp. 550-554, 1994.

[4] C. Viard-Gaudin, P.M. Lallican, S. Knerr, P. Binter, "The IRESTE On/Off (IRONOFF) Dual Handwriting Database", *The 5th International Conference on Document Analysis and Recognition,* 1999. pp. 455-458.

[5] U. Bhattacharya, B.B. Chaudhuri, "Databases for research on recognition of handwritten characters of Indian scripts", *The 8th International Conference on Document Analysis and Recognition,* Seoul, 2005, pp. 789-793.

[6] K. Matsumoto, T. Fukushima, M. Nakagawa. "Collection and Analysis of On-Line Handwritten Japanese Character Patterns", *The 6th International Conference on Document Analysis and Recognition,* Seattle, 2001.

[7] S. Mori, K. Yamamoto, H. Yamada, T. Saito, "On a handprinted kyoiku-kanji character data base", *Bull. Electrotech. Lab*, 1979, 43(11-12), 752-733.

[8] Y.J. Liu, J.W. Tai, J. Liu, "An introduction to the 4 million handwriting Chinese character samples library", *Proceedings of the International Conference on Chinese Computing and Orient Language Processing,* Changsha, 1989, pp. 94-97.

[9] Y. Ge, Q. Huo, "A comparative study of several modeling approaches for large vocabulary offline recognition of handwritten Chinese characters", *The 16th International Conference on Pattern Recognition,* Quebec, 2002, 85-88.

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

[11] Y. Ge, F.J. Guo, L.X. Zhen, Q.S. Chen, "Online Chinese character recognition system with handwritten Pinyin input", *Proceedings of Document Analysis and Recognition,* 2005.

[12] http://www.sogou.com/labs/dl/w.html

[13] Y. S. Wu, X.Q. Ding, "principles, methods and implementation of handwritten Chinese character recognition", *Higher Education Press,* pp.30, 1993.