SCUT-COUCH2009----A Comprehensive Online Unconstrained Chinese Handwriting Database and Benchmark Evaluation

Lianwen Jin[+], Yan Gao, Gang Liu, Yunyang Li, Kai Ding

School of Electronic and Information Engineering,
South China University of Technology
[+]Lianwen.jin@gmail.com

Abstract: A comprehensive online unconstrained Chinese handwriting dataset, SCUT-COUCH2009, is introduced in this paper. As a revision of SCUT-COUCH2008 [1], the SCUT-COUCH2009 database consists of more datasets with larger vocabularies and more writers. The database is built to facilitate the research of unconstrained online Chinese handwriting recognition. It is comprehensive in the sense that it consists of 11 datasets of different vocabularies, named GB1, GB2, TradGB1, Big5, Pinyin, Letters, Digit, Symbol, Word8888, Word17366 and Word44208. In particular, the SCUT-COUCH2009 database contains handwritten samples of 6,763 single Chinese characters in the GB2312-80 standard, 5,401 traditional Chinese characters of the Big5 standard, 1,384 traditional Chinese characters corresponding to the level 1 characters of the GB2312-80 standard, 8,888 frequently used Chinese words, 17,366 daily-used Chinese words, 44,208 complete words from the Fourth Edition of "The Contemporary Chinese Dictionary", 2,010 Pinyin and 184 daily used symbols. The samples were collected using PDAs (Personal Digit Assistant) and smart phones with touch screens, and were contributed by more than 190 persons. The total number of character samples is over 3.6 million. The SCUT-COUCH2009 database is the first publicly available large vocabulary online Chinese handwriting database containing multi-type character/word samples. We report some evaluation results on the database using state-of-the-art recognizers for benchmarking.

Keywords*: SCUT-COUCH, online handwritten character database, handwritten Chinese word samples, handwritten Pinyin dataset, benchmark evaluation*

1 Introduction

Online handwriting Chinese character recognition (OHCCR) has attracted more and more academic and industrial attention in recent years [2]. OHCCR is a key technology for the input method for many portable devices, such as Personal Digital Assistants (PDA), smart mobile phones, Tablet PCs, Electronic Learning Machines (ELM) etc. Since different people have different writing styles, a large scale online database of handwriting samples contributed by many writers is crucial for the development of a high performance online handwriting recognizer.

Since 1993, the UNIPEN project has organized the collection of online handwritten characters from several countries [3,4]. In recent decades, numbers of handwriting databases have been published in the literature. To name a few, for offline databases, there are the English databases such as CENPARMI[5] and CEDER[6], French database IRONOFF[7], Indian database ISI[8], Japanese/Kanji database ETL-8/ETL-9[10], Chinese databases of IAAS-4M[11], HCL2000[12] and HIT-MW[13]. For online databases, there are the online Japanese databases Kuchibue and Nakayosi[9], and recently the publicly available online Chinese handwritten database of CASIA-OLHWDB1[14] and SCUT-COUCH2008[1], resulting in powerful promotion and rapid development of handwritten character recognition. In particular, it has been noted that in recent years, there has been an increasing interest in building publically accessible ground-truth databases for different languages, such as the online Chinese character database [14], Arabic offline handwritten character database [15], handwritten offline Spanish text database [16], Farsi handwritten text database [17], etc.

In short, the development of handwriting databases shows definite trends. First, the collection of categories has developed from a small range into a large collection. Second, the scale of sampling has grown from single characters to words, text lines and paragraphs. Third, the manner of handwriting styles has changed from regular to cursive and unconstrained forms.

In the field of online handwritten Chinese character recognition, although many key players such as HanWang Technology in China, PenPower Technology in TaiWan, Microsoft and Motorola in the United States, have their own non-public handwritten database, no one is willing to share their database with other researchers in the field. Up to now, to the best of our knowledge, there are only two online handwriting Chinese character databases, SCUT-COUCH2008 [1] and CAISA-OLHWDB1 [14], which have both been made publicly available.

CAISA-OLHWDB1 is a database which is freely available to the academic community and was released in 2009[14]. It contains 4,037 characters (3,866 Chinese characters and 171 symbols) written by 420 people. Although it is contributed by

many writers, CAISA-OLHWDB1 does not yet include samples of words, traditional Chinese characters, complete GB2312-80 level2 Chinese characters and Pinyin.

SCUT-COUCH2009 is a revision of SCUT-COUCH 2008 which was first released in 2008 [1]. It has the following improvements: (1) Many more samples were collected and contributed by more than 190 different people. (2) The 5401 Big5 traditional character set was added. (3) All word phrases of the Fourth Edition of "The Contemporary Chinese Dictionary (Fourth Edition)", an official popular Chinese word dictionary, were added. (4) The number of datasets was increased from the original 8 to 11 in SCUT-COUCH2009.

The rest of this paper is organized as follows: Section 2 gives a brief description of the SCUT-COUCH2009 database. In Section 3, general consideration on how to design and build the SCUT-COUCH2009 database is presented. Ground-truth processing of the database is given in Section 4. The analysis of the database is presented in Section 5. Section 6 presents benchmark testing of the database using a state-of-the-art recognizer. Section 7 concludes this paper.

2 A brief description of the database

SCUT-COUCH2009 was built for the purpose of providing researchers in the field of online handwritten Chinese recognition a refined online Chinese database with training and testing samples, and also provides a standard publicly available database for comparing and evaluating the performance of different algorithms.

Compared with the databases mentioned above, the proposed SCUT-COUCH2009 database has the following special characteristics: Firstly, SCUT-COUCH2009 is a comprehensive online unconstrained database composed of 11 datasets as shown in Table 1. Secondly, it's the first publicly available Chinese handwriting database to include word samples. Some currently available online Chinese handwriting databases, such as HCL2000 and CAISA-OLHWDB1, provide only samples of single characters. The lack of a database of Chinese words confines Chinese handwriting recognition to the level of single character recognition, leaving Chinese word/text recognition barely addressed in the literature. Thirdly, it's the first publicly available online handwriting database that covers the Chinese Pinyin collections. More frequently in practice, people can remember the pronunciation of some characters rather than their shape. A handwriting Pinyin input method will be convenient in these situations. Similar solutions have been addressed recently [18]. An unconstrained handwritten Pinyin dataset is needed to train a robust Pinyin recognition engine.

.

Table 1 11 datasets in SCUT-COUCH2009

datasets	Number of writers	Number of Categories	Number of samples	Description
Word8888	130	8,888	1,155,440	8,888 frequently used Chinese words
Word44208	5	44,208	221,040	Including all phrases of "The Contemporary Chinese Dictionary", the Fourth Edition, except words which consist of more than 12 characters.
Word17366	10	17366	173,660	Daily used words picked from the PowerWord dictionary
GB1	188	3,755	705,940	3,755 characters in GB Set1[1]
GB2	195	3,008	586,560	3,008 characters in GB Set2[1]
TradGB1	130	1,384	179,920	1,384 traditional characters in GB Set1
Pinyin	130	2,010	261,300	2,010 Pinyin
Letter	195	52	10,140	52 English upper- and lower- case alphabets
Digit	195	10	1,950	10 numeric digits
Symbol	195	122	23,790	122 frequently used symbols
Big5	65	5,401	351,065	5,401 traditional Chinese character in the Big5 standard
Total Samples			3,670,805	

3 General consideration and preparation for building SCUT-COUCH 2009 database

We fully understand, building a comprehensive online handwritten Chinese database means a great amount of work. Since, every stage in the collection of the SCUT-COUCH2009 database, from the selection of participants, sampling devices, the sampling of writers, to the establishment of sampling rules, is carefully planned and considered. In this section, we will introduce the process we used to accomplish the complete preparation for the building of the SCUT-COUCH database.

3.1 Program Design

After analyzing and comparing currently available devices for sampling online handwriting, we finally decided to employ five PDAs (Personal Digital Assistant) and five smart mobile phones with touch screens, instead of using

[1]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.

trajectory recording pens or tablets as the sampling devices. This was primarily because smart phones and PDAs are more portable and universally used as handwriting input devices in daily life.

As illustrated in Fig. 1, the input area in our program is designed to be a box of static size, large enough to accommodate as many characters as may be required. In addition, noticing that most characters have their last strokes ending in the bottom-right corners of the characters, we place the "Save and Continue" button on the bottom of the input box and thus make it convenient for writers to proceed quickly to the next object after writing a sample.

Fig. 1 An illustration of layout of our data collection software

3.2 Corpus Preparation

There is always a specific strategy in sampling objects of a handwriting database. For example, HCL2000 took the 3755 frequently used Chinese characters of level 1 in the GB2312-80 standard as sampling objects, while HIT-MW adopted various kinds of paragraphs in China Daily. We also have a strict principle in the choice of sampling materials for SCUT-COUCH2009.

3.2.1 Word Picking

The vocabulary of Chinese words is extremely large. It would be impossible to collect all the words. Therefore, taking into account the three levels, defined as small, medium and large scale, we build the handwritten Chinese word database into three datasets, namely Word8888, Word17366, and Word44208 respectively.

The word corpus of the Word8888 dataset is picked according to the frequencies of usage. We employed the statistics on Chinese word usage frequencies published by the "Sogou Labs" [19]. The frequency of appearance of 157,202 words

in more than 100 million web pages is counted in the statistics. In our analysis, we found words that frequently appear on the internet are also continually used in our daily life. Therefore, we intercepted the 8,888 words with the highest usage frequencies to be word sampling objects, based on the following considerations. On the one hand, we should not make the collection work too burdensome by adopting too large a number of sample objects. Upon such consideration, 8,888 frequently-used words which involve 19,595 single Chinese characters were picked as an appropriate number. On the other hand, we should not lose universal coverage in daily usage by adopting too small a number of sample objects. As we can see in Fig. 2, these 8,888 words cover 71.48 % of the daily-used lexicon. We picked 8,888 words because 8,888 is considered a lucky number in China which makes it easy for everyone to remember.

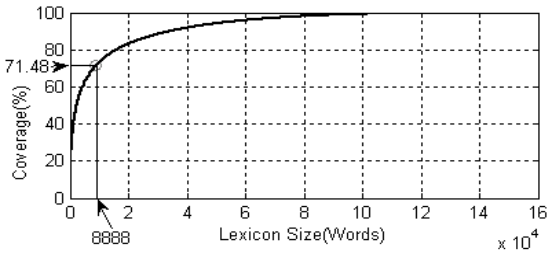


Fig. 2 Words usage frequency distribution

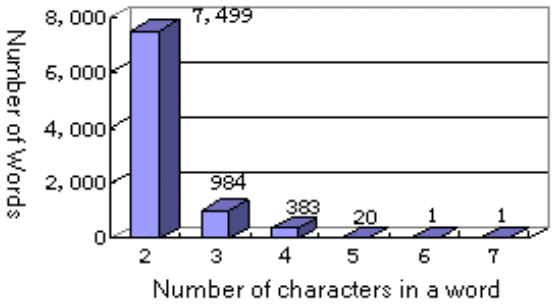


Fig. 3 Statistics of lengths of the 8,888 words

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

The corpus of the Word17366 dataset is picked from PowerWord, the most popular digital Chinese dictionary software developed by the KingSoft Company [20]. We manually picked 17,366 daily-used words as our sampling objects for building a middle scale handwritten word dataset, which involves 14,822 two-character words, 1,195 three-character words and 1,349 four-character words. As the corpus of the Word8888 dataset was picked from websites, we hope that picking the corpus of words from the most popular digital dictionary software provides an alternative representation.

To build a large vocabulary and relatively complete word database, we built the third word dataset, the Word44208 dataset, which includes all word phrases except 4 long-characters words which consisted of more than 12 characters from an official Chinese dictionary, “The Contemporary Chinese Dictionary” (Fourth Edition). This dataset contains 44,208

words in total. As shown in Fig. 4, it involves 34,039 two-character-words, 6,517 three-character-words, 2,722 four-character-words, 531 five-character-words, 252 six-character-words, 73 seven-character-words, 41 eight-character-words, 19 nine-character-words, 9 ten-character-words, 4 eleven-character-words, and 1 twelve-character-word. In total there are 103,840 single characters which cover 3,643 characters in level 1 of the GB2312-80 standard and 1,781 characters in level 2 of the GB2312-80 standard.

It is worth noting that the vocabularies of the three vocabularies overlap. In fact, there are 4,952 common words shared between the Word8888 and Word17366 datasets, 4,713 words shared by the Word8888 and Word44208 datasets, and 11,907 words shared between the Word17366 and Word44208 datasets.

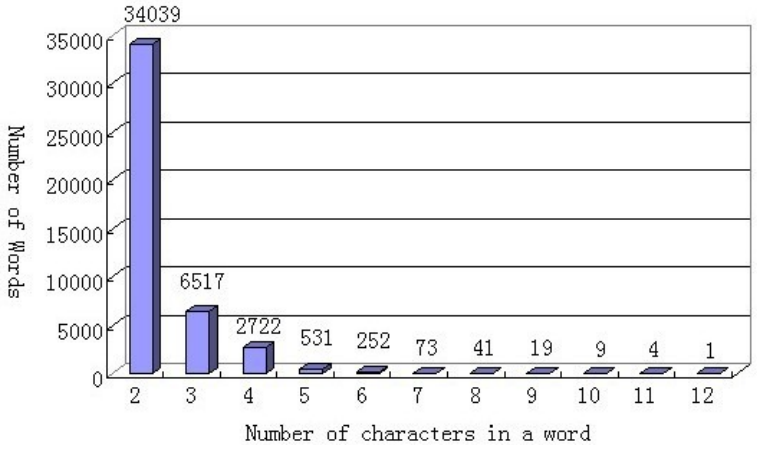


Fig. 4 Statistics of lengths of words in Word44208 dataset

There were several words which consisted of more than 12 characters when we collected the original handwritten samples from the contents of "The Contemporary Chinese Dictionary" (Fourth Edition) (in fact there are 1 twelve-character-word, 1 thirteen-character-word, 1 sixteen-character-word and 1 seventeen-character word). As it is hard to write these words in one line on a PDA, most of them are written in two lines. Fig. 5 shows some long words as handwritten samples. The bottom-left sample shown in Fig. 5 contains as many as 17 characters. However, these long-characters words provide us with a type of text sample rather than word sample. Therefore we do not keep them when building the dataset of the Word44208. Instead, we build a separate subset for them which is not released in the current version of the SCUT-COUCH2009 database, but is available from us upon request via email.

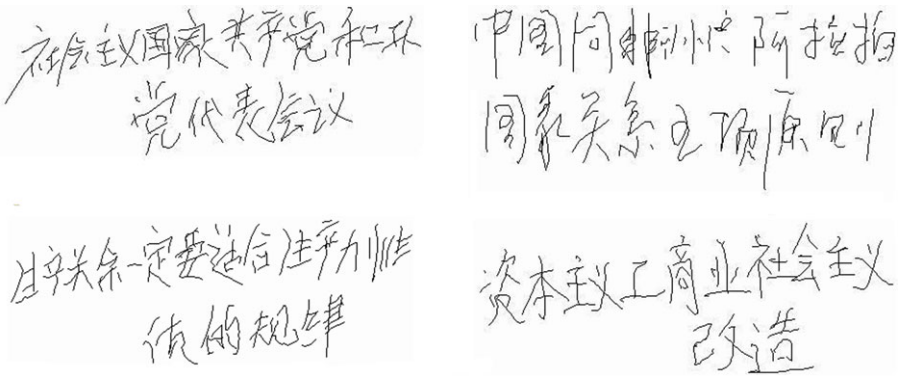


Fig. 5 Illustration of some long-characters word samples

3.2.2 Selecting Single Characters

We selected all 3,755 level 1 and all 3,038 level 2 simplified Chinese characters in the GB2312-80 standard as our single simplified body of characters, since the 6,763 characters in the GB2312-80 standard cover more than 99.9 % of daily usage in China.

Furthermore, noticing that traditional Chinese characters are also used occasionally, an appropriate range of traditional Chinese characters is also listed in our database. Two traditional Chinese character datasets, the Big5 and TradGB1 were designed. The Big5 contains 5,401 traditional Chinese characters, which are widely used in Taiwan and Hong Kong. In China, as about 99 % of daily used characters are covered within the GB2312-80 level 1, the set of TradGB1 traditional Chinese character samples is obtained by the transformation of the 3,755 simplified Chinese characters in the GB Set1 to their traditional forms. In particular, different traditional forms may be found in different lexicons for the same simplified character (see examples shown in Table 2). Finally, 1,384 traditional Chinese characters are collected as a merger 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	豔艷艷
历	歷曆

3.2.3 Selection of Pinyin Material

In the pronunciation of Mandarin, every kind of spelling 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 is more often the case that one kind of spelling has only 4 or less tones with the corresponding Chinese character. For example, no character exists relating to the pronunciation lié.

Through analyzing the pronunciations of the lexicon of GB2312-80, we acquired a collection of Pinyin involving 415 kinds of spelling (regardless of tone), and 1,426 enunciations (considering tones). After excluding some spellings that are rarely used (such as hng, hm, ng, etc), a lexicon of 402 types of spelling was obtained. All 5 tones of every spelling were included into our dataset (2,010 in total), regardless of whether or not they really have a corresponding Chinese character, due to the following concerns. First, this strategy results in five samples for every spelling, and thus increases the statistical stability of handwritten spellings. Second, it facilitates the analysis of the difference between different tones of the same spelling. Third, being a widely used and developing language, Chinese pronunciations may change along with the emergence of new characters.

3.2.4. Selection of other materials

Aside from materials mentioned above, we also preserved the involvement of some regular materials, including 10 numeric digits, 52 upper- and lower- case English alphabetical characters and 122 frequently used symbols (see Fig. 6).

Fig. 6 Symbols in SCUT-COUCH2009

3.3 Writer diversities

Generally, our strategy in selecting candidate writers combined deliberation with randomness. Deliberately, we selected two groups of writers from two cities in China, Guangzhou (a southern China city), and Hefei (a middle northern China city). A total of 194 different people contributed to the collection of the database. Most writers were students from universities, which were suitable for the following reason. 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.

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

Table 3 Sampling percentage of three regions

Region	Proportion
South Region	62 %
Middle Region	20 %
North Region	18 %

Table 4 Age distribution of writers

Age	Proportion
Below 18	3 %
Between 19 and 35	87 %
Between 36 and 50	4 %
Older than 50	6 %

Table 5 Gender distribution of writers

Gender	Proportion
Male	57 %
Female	43 %

3.4 Policies of Samples Collection

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

(1) All sets of samples were written by many different writers. Every writer could write several sets, but for one dataset they could only write one set to ensure the stability in writing styles.

(2) Writers were asked to input handwritings in the input box (for both single characters and words) without knowing the content of the next object.

(3) Writers were aware that every object was an integral entity rather than a simple combination of multiple single characters, to preserve the integrity of each collected sample and ensure the database's availability in research that takes them as entities.

(4) Writers were asked to perform handwriting in their most comfortable and familiar manner. No restrictions were imposed on the quality of the writers' handwriting.

(5) Writers were ignorant about the database's potential usage in Chinese character recognition, to simulate natural handwriting and ensure the unconstrained style of the samples collected.

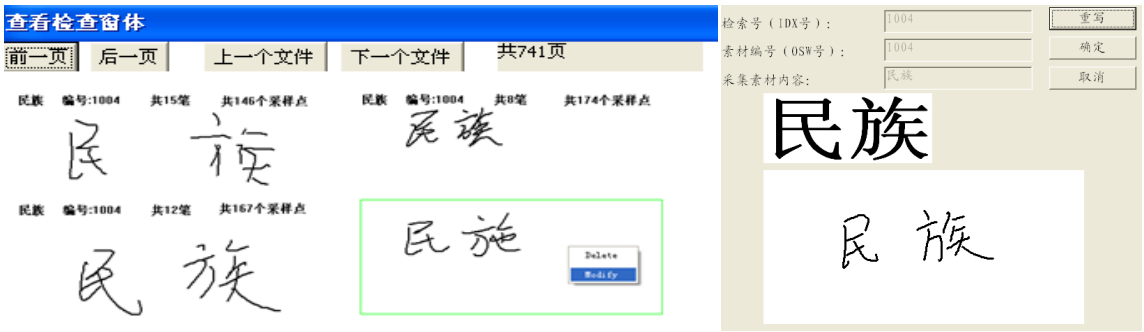
(6) A continuous duration of each sampling was confined to 1.5 hours in average, not exceeding 2 hours at most, to prevent samples from being distorted from exhaustion.

4 Ground-Truth Processing

With all sets of data completely sampled by more than 190 writers individually, we started to check and store the collected data. A helpful form of word labeling was also performed, which will be discussed in this section.

4.1 Error Correction

For the purpose of error correction, samples of the SCUT-COUCH2009 can be checked with a tool developed by us. For example, as shown in Fig. 7(a), there are four samples "民族" which are written by four people individually. But the fourth person wrongly writes "民族" as "民施". The checker can move the mouse to the fourth character and a green box will appear. By pressing the right button, the "delete" or "modify" option can be selected in the pop-up menu. If the "modify" option is chosen, a new layout will be launched as illustrated in Fig. 7(a). The sample can be re-written in the white area shown in Fig. 7(b).



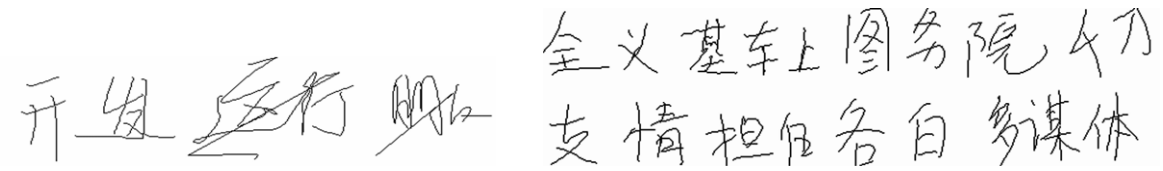
(a) (b)

Fig. 7 Illustration of error correction

The person was asked to re-write his/her error samples in the correction process. However in some cases when the number of error samples was few (usually no more than 20) and the original person was not available at that moment, we used our laboratory personnel to make such corrections.

While checking the collected data, we found few samples with unacceptable errors. As shown in Fig. 8(a), some samples involve sudden excursion on the trajectories, caused by occasional malfunctions in the hardware of the touch screens. The other errors were not caused by hardware but by subjective mistakes, which were also not beneficial for analyzing the disciplines of Chinese handwriting. Fig. 8(b) shows that some words were wrongly written as 全义, 基车

上，图务院，千刀，支情，担伍，各白 and 多谋体, while the correct words should have been 全文，基本上，国务院，千万，友情，担任，各自 and 多媒体. All the error samples mentioned above were required to be corrected. Meanwhile we preserved these error samples in case of special research needs.



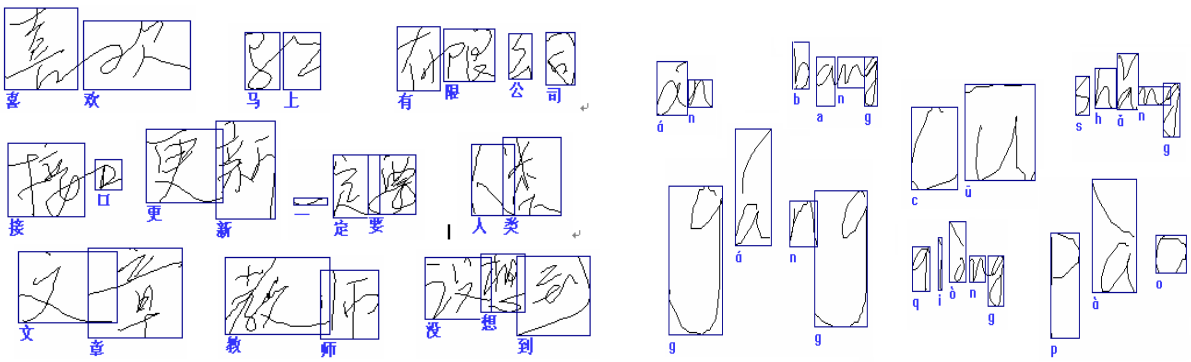
(a) (b)

Fig. 8 Some samples written mistakenly

4.2 Database Labeling

To make the handwritten word samples available for segmentation-based research, we segmented the 8,888 Chinese word samples into single characters and manually labeled each character. Our method was to record the bounding box of every single character as well as the indices of the strokes in each character. The labeling of word samples was meaningful for three reasons. First, it provided ground truth data for conducting research on Chinese character segmentation. Second, all single characters were separated and hence enriched the samples of isolated characters and also provided samples for special research purposes, such as writer adaptation using incremental samples. Last but not least, it offered information on both the spatial and temporal relationships between character members of a Chinese word.

All the labeling of words was conducted on a PC using a semi-automatic tool. Fig. 9(a) shows some of the labeled word samples. Similarly, we also labeled Chinese Pinyin samples in our data sets, enriching all kinds of samples of letters which were written by one person. Fig. 9(b) shows some of the labeled Pinyin samples.



(a) (b)

Fig. 9 Word & Pinyin segmentation and labeling

4.3 Fundamental Structure of the Database

All the handwritten character/word sample files of SCUT-COUCH2009 share the same storage data structure. Each file contains various handwriting samples, which are stored sequentially. Each handwritten sample consists of a number of strokes, where each stroke consists of writing points. The data structure of the handwritten sample contains general description information such as the corresponding GB code, total stroke number, total point number, etc., and detail data storage organization information, which is given in Table 6.

Table 6 Description of a data structure of SCUT-COUCH2009

Item	Type	Length(BYTE)	Description
WordLength	unsigned char	1	Number of characters in this sample
WordCode	array of unsigned char	WordLength×1	GB code of the sample
PointNum	unsigned short int	2	Total number in this sample
LineNum	unsigned short int	2	Total stroke number in this sample
GetTimePointNum	unsigned short int	2	Total number of points in this sample
GetTimePointIndex	array of unsigned char	GetTimePointNum×2	Index numbers of all points
ElapsedTime	array of DWORD	GetTimePointNum×4	Elapsed time interval between current point and previous point
StrokeData	Strokes data of the sample, the data structure is given below		
StrokePointNum	unsigned short int	2	number of points on the stroke
Points(x,y)	unsigned short int	2+2	x and y value for each point on the stroke

5 Data Analysis

The SCUT-COUCH2009 database is not only a publicly available online unconstrained Chinese handwriting database, but also innovatively contains multi-type sample objects, which is useful for online Chinese handwriting recognition. Fig. 10 shows some samples in the different datasets of the SCUT-COUCH2009. In this section, we briefly introduce both statistical information and some characteristics of our database.

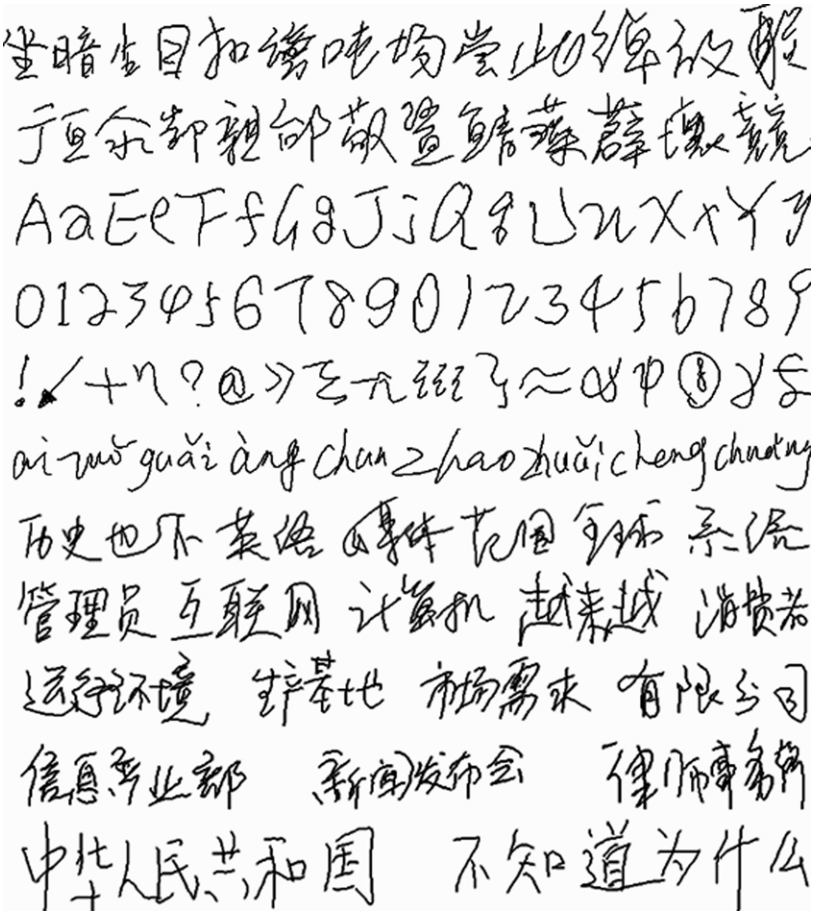


Fig. 10 samples in SCUT-COUCH2009

5.1 Statistical Discipline of Strokes

The statistics of the stroke number is analyzed for the range of 8,888 words in 130 sets of samples, which is shown in Fig. 11, From Fig. 11, we come to the following conclusions. First, many Chinese characters are written cursively, with two or more strokes connected, since the average stroke number of the handwritten samples is much less than that of standard characters. Second, some Chinese words are even cursively written in one stroke. Third, the stroke number of every word varies from sample to sample, instead of being identical with the standard. Fourth, the rule-less number of strokes also reflects the original writing styles of different writers, giving proof to the unconstrained characteristic of the SCUT-COUTH database.

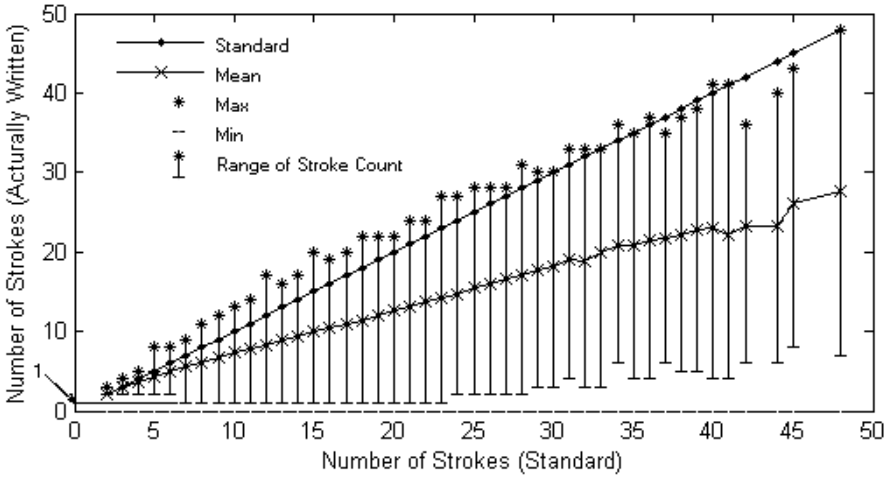


Fig. 11 Statistics on words stroke numbers.

5.2 Handwriting Diversities in SCUT-COUCH2009

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 were used in the collection, data of rich diversity were presented in our collected data. The following describes some special characteristics of our database.

- **Different styles written by the same person:** different styles may be seen in handwriting by the same person, as Fig. 12(a) shows. A, B, C and D are written by the same writers.
- **Aliasing:** Aliasing may sometimes appear on writing trajectories due to the limited resolution of touch screens. Fig. 12(b) shows examples of aliasing.
- **Disconnected stroke:** There can be a situation where a stroke is disconnected in its trajectory, Fig. 12(c) shows characters “号”, “我”, “软” and “阵” with disconnected strokes, caused by a sudden failure in capturing part of the trajectory when the stroke is written too fast.
- **Redundant stroke:** Some writers were used to making a point when finishing writing a sample word. In addition, a redundant stroke may be produced by carelessly touching the touch screen, see Fig. 12(d).
- **Missing stroke:** Some small strokes were missed out in fast writing, Fig. 12(e) shows characters “学”, “仪”, “逮” and “限”.
- **Over-connected stroke:** In the writing process, even some strokes in different characters were written connectively. Fig. 12(f) shows over-connected-stroke samples “主题”, “运行”, “朋友”, “ai” and “chà”.

- **Mis-written traditional Chinese characters:** When collecting samples of simplified Chinese characters, some writers mistakenly wrote them in the traditional forms following their habits. Fig. 12(g) shows samples as “认为”, “负责”, “冠军”, “首页” and “海报”. However, the number of such samples was very small.
- **Rewritten stroke:** Occasionally, when writers found some strokes to look bad or disconnected, they would rewrite them instead of erasing the whole character. Fig. 12(h) shows the writing process of the character “a” and “赛”.

Fig. 12 Handwriting Diversities in SCUT-COUCH2009

It is worth mentioning that in order to keep the handwriting diversities in SCUT-COUCH2009, most of the samples which were mis-written in traditional form or with rewritten strokes were not asked to be rewritten in the correction process. However, if a sample was written too badly to be recognized by a human, it was corrected.

- **Style variation of different writers:** Being a flexible form of art itself, Chinese calligraphy may have many styles for the same character. Fig. 13 gives examples of different styles of three characters.

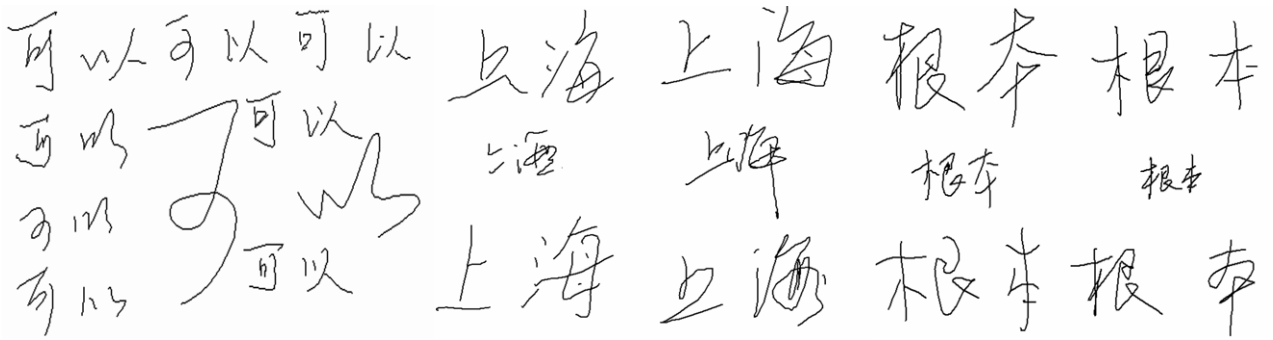


Fig. 13 Varied handwriting styles of different persons

6 Benchmark evaluations

The classifier we used is a state-of-the-art compact MQDF classifier proposed by Teng Long & Lianwen Jin [21]. After extraction of the 8-directional features [22], the dimensions of the original feature vector (512D) is reduced by LDA (Linear Discriminant Analysis). Then a two-level minimum distance coarse classifier was designed for the purpose of speeding up the recognition process. The first level coarse classifier uses the first 16 dominant LDA features for classification and generates about 300 candidates for the second level coarse classifier. The second level coarse classifier uses 160 dominant LDA features and generates 20 candidates for the MQDF classifier, which outputs the final recognition result. The MQDF parameters are compressed using a split vector quantization technique to form the final compact dictionary.

Experimental results on all 8 single character datasets and the Word8888 dataset of SCUT-COUCH2009 are given in Table 7. It is worthwhile noting that for the recognition of the Pinyin dataset and the Word8888 dataset, a holistic approach [23] was used here. For each dataset we randomly chose 80 % of the set as the training set and the remaining 20 % as the testing set.

Table 7 Recognition accuracy (%) of each dataset

dataset	Top1	Top2	Top5	Top10
Pinyin	81.73	92.06	97.22	98.50
Letter	77.66	94.87	98.32	99.06
Digit	98.21	99.49	100.00	100.00
Symbol	81.15	91.20	97.16	98.93
TradGB1	97.10	98.48	98.86	98.99
BIG5	95.94	98.24	99.08	99.28
GB1	95.27	97.99	99.01	99.26
GB2	95.62	97.48	98.19	98.38
Word8888	96.86	98.28	98.79	98.94

Experiments on the recognition of various combinations of different datasets were also conducted, with the results shown in Table 8. It is worthwhile noting that for “GB1+GB2+Pinyin” and “GB1+GB2+Word8888”, the same recognizer was used for both character and pinyin/word samples. This means that a Pinyin sample and a word sample were treated as a character in the same way. Another point which should be noticed is that for the case “GB1+GB2+Big5”, among the total 12,164 categories of characters, there are 3,328 categories that are the same between GB1+GB2 and Big5. After merging the same categories, the actual category number of the combination of datasets GB1, GB2 and Big5 was 8,836. “GB1+GB2+Big5+Symbol+Letter+Digit” was also treated in the same way.

Table 8 Recognition accuracy (%) of each dataset combination

dataset combination	number of categories	Top1	Top2	Top5	Top10
GB1+GB2	6763	94.63	97.87	99.03	99.29
GB1+GB2+TradGB1	8147	94.43	97.87	99.04	99.29
GB1+GB2+Pinyin	8773	91.34	96.07	98.30	98.83
GB1+GB2+Big5	8836	94.16	97.61	98.87	99.16
GB1+GB2+Word8888	15651	94.89	97.38	98.31	98.53
GB1+GB2+TradGB1+ Pinyin+Symbol+Letter+Digit	10341	91.01	95.97	98.21	98.72
GB1+GB2+Big5+Symbol+ Letter+Digit	9020	93.66	97.33	98.71	99.00

From Table 7, it can be seen that the recognition rates for the Pinyin dataset, Letter dataset, and Symbol dataset are not good enough. This may be due to the fact that there are many similar and confusable characters in these datasets, such as “x” and “X”, “o” and “O”, “w” and “W”, “ ‘‘ ” and “ ’’ ”, “ ’ ”, “ ’ ” and “ ‘ ” etc. For the simplified and traditional Chinese character datasets GB1, GB2 and Big5, the Top 1 recognition accuracies are relatively higher, but still less than 96 %, indicating there is still much room to study the problem of unconstrained online Chinese character recognition. From Table 8, we can see that for the combination of different datasets, the Top 1 recognition accuracies are even lower. This demonstrates that the problem of hybrid recognition of Chinese/Pinyin/Symbol/English/Digit/Word is still a challenging problem that needs a lot of research effort to address. It is also observed from Tables 7 and 8 that the Top 2 recognition rates are significantly higher than the Top 1 recognition rates. This may be due to the fact that there are many similar confusable character-pairs in the datasets, such as “人” and “入”, “日” and “曰”, “己”, “已” and “巳” in the GB1 dataset, “x” and “X”, “o” and “O”, “w” and “W” in the Letter dataset and so on. Therefore, the recognition accuracy is greatly increased naturally when taking into account of the Top 2 recognition candidates.

To test the recognition accuracy for Word17366 and Word44208, we used a segmentation based classification approach as mentioned in [24]. The results are given in Table 9. It can be seen from Table 9 that the accuracy is less than 90 % for most word datasets, indicating the problem of unconstrained handwritten Chinese word recognition is an even more challenging problem and is far from being solved.

Comparing the recognition results on Word8888 shown in Table 9 with the last row of Table 7, we find that using a holistic approach for word recognition achieves a much higher performance. This indicates that the holistic approach may be a good choice for the recognition of handwritten word samples. However, with the increase of categories of words (for

example, more than 17,000 for the Word17366 dataset and more than 44,000 for the Word44208 dataset), it would be very difficult and not be feasible to train a recognizer using the holistic approach, since the parameters of the MQDF classifier would become very large, and it may cause the singularity problem [25] when applying the LDA and MQDF due to the small number of word samples in the same class.

Table 9 Recognition Accuracy of Word

Word dataset	Recognition Accuracy (%)
Word8888	84.87
Word17366	86.81
Word44208	89.77

7 Conclusions and discussion

The multi-type online unconstrained Chinese handwriting database SCUT-COUCH2009, has important features that are not available in other databases. It is a comprehensive online unconstrained handwritten database composed of 11 datasets. It is the first publicly available online handwritten Chinese character database that involves a multi-type corpus of words such as Chinese words and Chinese Pinyins. The collection of the SCUT-COUCH2009 database was conducted under the supervision of carefully designed strategies. From the selection of material and sampling devices, the sampling of writers, to the establishment of sampling rules, all are elaborately designed.

The SCUT-COUCH2009 promises many more substantial applications than online isolated character recognition. A number of novel research topics could be promoted with the assistance of our database, such as handwritten Pinyin recognition, online handwritten Chinese word segmentation and recognition, hybrid handwritten Chinese/English/Digit/Word /Symbol/Pinyin recognition, incremental learning for writer adaptation [26], and so on.

We are still extending our database, and one study we are working on is to build an online handwriting sentence/text dataset, which is in the process of being collecting now and is expected to be completed by the end of this year. The corpus in text dataset comes from news of different topics from the China People's Daily, dated from 2009/01/01-2009/06/31, including international, domestic, economic, cultural, and sporting news, scientific education weekly, academic trends and so on. We are collecting the handwritten text data with two devices, one is a PC with touch screen, and the other is a ZPen digital pen [27]. This dataset will be released on the SCUT-COUCH website (http://www.hcii-lab.net/data/scutcouch/) when it is available.

The SCUT-COUCH2009 databases we collected are just a starting point for promoting research work in the area of online handwritten character recognition. To develop a high performance online handwriting recognition engine, a much

larger scale of databases contributed by hundreds or even thousands of people may be needed. Furthermore, the collection of handwriting character samples is never complete, since people's handwriting styles change as generations change, and also as the writing instruments change with the advance of hardware. With the progress of handwriting recognition technology, more and more challenging data will become necessary, such as realistic cursive character sentences/paragraph samples. Therefore, it would be fair to say that the task of online data collection is a journey without an end.

All the datasets of the SCUT-COUCH2009 database are freely accessible to researchers in the community and its latest information is available at http://www.hcii-lab.net/data/scutcouch/. More information is also available upon request (please email to lianwen.jin@gmail.com).

Acknowledgments

We would like to thank all anonymous reviewers for their valuable suggestions. We would like to thank ZhiBin Huang, Bin Zhang, Shengming Huang, Xuewen Liu, Guoqiang Deng, Hengzhi Zhang, Dapeng Tao, Hanyu Yan, Lisha Yang et. al. for helping to supervise data collection. We would also like to warmly thank all the cooperative volunteers. This work is supported in part by the National Science Foundation of China (NSFC) (grants no. U0735004, and no. 60772216), and the Guangdong National Science Foundation (grant no. 07118074).

References

1. Li, Y.Y., Jin, L.W., Zhu, X.H., Long, T.: SCUT-COUCH2008: A Comprehensive Online Unconstrained Chinese Handwriting Dataset. In: Proceedings of the 11th International Conference on Frontiers in Handwriting Recognition, ICFHR08, pp. 165-170 (2008).
2. Liu, C.L., Jaeger, S., Nakagawa, M.: Online recognition of Chinese characters: the state-of-the-art. IEEE Trans. Pattern Anal. Mach. Intell. 26(2), 198 – 213 (2004).
3. Jaeger, S.; Nakagawa, M.: Two on-line Japanese character databases in Unipen format. In: Proceedings of Sixth International Conference on Document Analysis and Recognition, ICDAR01, pp. 566 –570 (2001).
4. The UNIPEN Project, http://hwr.nici.kun.nl/unipen/.
5. Suen, C.Y., Nadal, C., Legault, R., Mai, T.A., Lam, L.: Computer recognition of unconstrained handwritten numerals. In: Proceedings of the IEEE, 80 (7), pp. 1162-1180 (1992).
6. Hull, J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. 16(5), 550-554 (1994).
7. Viard-Gaudin, C., Lallican, P.M., Knerr, S., Binter, P.: The IRESTE On/Off (IRONOFF) Dual Handwriting Database. In: Proceedings of the fifth International Conference on Document Analysis and Recognition, ICDAR99, pp. 455-458 (1999).
8. Bhattacharya, U., Chaudhuri, B.B.: Databases for research on recognition of handwritten characters of Indian scripts. In: Proceedings of the eighth International Conference on Document Analysis and Recognition, ICDAR05, pp. 789-793(2005).
9. Matsumoto, K., Fukushima, T., Nakagawa, M.: Collection and Analysis of On-Line Handwritten Japanese Character Patterns. In: Proceedings of the sixth International Conference on Document Analysis and Recognition, ICDAR01, pp. 496-500 (2001).
10. Mori, S., Yamamoto, K., Yamada, H., Saito, T.: On a hand printed kyoiku-kanji character database. Bull. Electrotech. Lab, 43(11-12), pp.752-733 (1979).
11. Liu, Y.J., Tai, J.W., Liu, J.: An introduction to the 4 million handwriting Chinese character samples library. In: Proceedings of the International Conference on Chinese Computing and Processing of Orient Language, ICCPOL89, pp. 94-97 (1989).

12. Ge, Y., Huo, Q.: A comparative study of several modeling approaches for large vocabulary offline recognition of handwritten Chinese characters. In: Proceedings of the 16th International Conference on Pattern Recognition, ICPR02, pp.85-88 (2002).
13. Su, T.H., Zhang, T.W., Guan, D.J.: Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. International Journal of Document Analysis and Recognition, IJDAR07, 10(1), 27-38 (2007).
14. Wang, D.H., Liu, C.L., Yu, J.L., Zhou, X.D.: CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters. ICDAR2009, In: Proceedings of the 6th International Conference on Document Analysis and Recognition, ICDAR09, pp. 1206-1210 (2009).
15. Alamri, H., Sadri, J., Nobile, N., Suen, C.Y.: A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition. In: Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR 08, pp. 664–669 (2008).
16. Perez, D., Tarazon, L., Serrano, N., Castro, F., Terrades, O.R., Juan, A.: The GERMANA Database. In: Proceedings of 10th International Conference on Document Analysis and Recognition, ICDAR09, pp.301-305 (2009).
17. Ziaratban, M., Faez, K., Bagheri, F.: FHT: An Unconstraint Farsi Handwritten Text Database. In: Proceedings of 10th International Conference on Document Analysis and Recognition, ICDAR09, pp.281-285 (2009).
18. Ge, Y., Guo, F.J., Zhen, L.X., Chen, Q.S.: Online Chinese character recognition system with handwritten Pinyin input. In: Proceedings of eighth International Conference on Document Analysis and Recognition, ICDAR05, pp.1265-1269 (2005).
19. Sogou Internet Word corpus, http://www.sogou.com/labs/dl/w.html.
20. PowerWord Website, http://cp.iciba.com/.
21. Long, T., Jin, L.W.: Building compact MQDF classifier for large character set recognition by subspace distribution sharing. Pattern Recognition. 41(9), pp. 2916-2925 (2008).
22. Bai, Z.L., Huo, Q.: A Study On the Use of 8-Directional Features For Online Handwritten Chinese Character Recognition. n: Proceedings of eighth International Conference on Document Analysis and Recognition, ICDAR05, pp. 232-236 (2005).
23. Ding, K., Jin, L.W., Gao, X.: A New Method for Rotation Free Method for Online Unconstrained Handwritten Chinese Word Recognition: A Holistic Approach. In: Proceedings of 10th International Conference on Document Analysis and Recognition, ICDAR09, pp.1131 – 1135 (2009).
24. Long, T., Jin, L.W.: A novel orientation free method for online unconstrained cursive handwritten Chinese word recognition. In: Proceedings of the 19th International Conference on Pattern Recognition, ICPR08, pp. 1-4 (2008).
25. Krzanowski, W.J., Jonathan P. et al.: Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. Applied Statistics, 44(2), pp.101–115 (2005).
26. Huang, Z.B., Ding, K., Jin, L.W., Gao, X.: Writer Adaptive Online Handwriting Recognition Using Incremental Linear Discriminant Analysis. In: Proceedings of 10th International Conference on Document Analysis and Recognition, ICDAR09, pp. 91-95 (2009).
27. ZPen, http://www.danedigital.com/6-Zpen/