CASIA Online and Offline Chinese Handwriting Databases

Conference Paper · October 2011

DOI: 10.1109/ICDAR.2011.17 · Source: IEEE Xplore

4 authors, including:

Cheng-Lin Liu
Chinese Academy of Sciences
252 PUBLICATIONS **6,194** CITATIONS
SEE PROFILE

Fei Yin
China University of Petroleum
71 PUBLICATIONS **895** CITATIONS
SEE PROFILE

Da-Han Wang
Xiamen University of Technology
19 PUBLICATIONS **452** CITATIONS
SEE PROFILE

2011 International Conference on Document Analysis and Recognition

CASIA Online and Offline Chinese Handwriting Databases

Cheng-Lin Liu, Fei Yin, Da-Han Wang, Qiu-Feng Wang
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation of Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, P.R. China
Email: {liucl, fyin, dhwang, wangqf}@nlpr.ia.ac.cn

***Abstract*—This paper introduces a pair of online and offline Chinese handwriting databases, containing samples of isolated characters and handwritten texts. The samples were produced by 1,020 writers using Anoto pen on papers for obtaining both online trajectory data and offline images. Both the online samples and offline samples are divided into six datasets, three for isolated characters (DB1.0–1.2) and three for handwritten texts (DB2.0–2.2). The (either online or offline) datasets of isolated characters contain about 3.9 million samples of 7,356 classes (7,185 Chinese characters and 171 symbols), and the datasets of handwritten texts contain about 5,090 pages and 1.35 million character samples. Each dataset is segmented and annotated at character level, and is partitioned into standard training and test subsets. The online and offline databases can be used for the research of various handwritten document analysis tasks.**

***Keywords*-Chinese handwriting databases; online; offline; isolated characters; handwritten texts**

I. Introduction

Despite the intensive research in handwriting recognition for over 40 years, the recognition of unconstrained handwriting remains a challenge: the performance of text line segmentation, word/character segmentation and recognition is still far behind the human recognition capability. For the design and evaluation of handwriting recognition algorithms and systems, the availability of large-scale, unconstrained handwriting dataset is very important. Large sample datasets are critically demanded for Chinese handwriting recognition because of the large number of character classes.

In the past, public datasets have significantly benefited the research. Among the offline sample datasets are the CENPARMI digits, CEDAR English words and characters [1], NIST handprinted forms and characters database [2], IAM English sentence database [3], Japanese Kanji character databases ETL8B and ETL9B, Indian database of ISI [4], Arabic databases [5], Farsi databases [6], Chinese databases HCL2000 [7] and HIT-MW [8], and so on. Databases of online handwritten data include the UNIPEN project, the Japanese online handwriting databases Kuchibue and Nakayosi [9], and the very recent Chinese online handwriting databases SCUT-COUCH2009 [10] and HIT-OR3C [11]. The French database IRONOFF contains both online and offline data, collected by attaching paper on digitizing tablet during writing [12].

A general trend of handwriting recognition is the transition from isolated character recognition to script recognition and from constrained writing to unconstrained writing. The existing Chinese handwriting datasets do not satisfy this trend: they are either too neat in writing quality or not large enough. The offline databases HCL2000 and old CASIA, both containing isolated character images of 3,755 categories, have been reported test accuracies higher than 98% [13]. The handwritten text database HIT-MW has only 853 page images containing 186,444 characters. The online database SCUT-COUCH2009 [10] consists of 11 datasets of isolated characters, Chinese Pinyin and words, but all the samples were produced by only 195 writers. The online database HIT-OR3S [11] contains isolated characters of 6,825 categories produced by 120 writers and handwritten texts of 10 articles produced by 20 writers.

This paper introduces a pair of online and offline Chinese handwriting databases built by the Institute of Automation of Chinese Academy of Sciences (CASIA). The handwritten samples were produced by 1,020 writers using Anoto pen on papers and include both isolated characters and handwritten texts (continuous scripts). A portion of online handwritten characters, in the dataset called CASIA-OLHWDB1 (now called as CASIA-OLHWDB1.0), have been released at ICDAR 2009 [14]. The databases include six datasets of online data and six datasets of offline data, in each case, three for isolated characters (DB1.0–1.2) and three for handwritten texts (DB2.0–2.2). All the data has been segmented and annotated at character level, and each dataset is partitioned into standard training and test subsets. All these datasets are free for academic research[1].

II. Data Collection Settings

A. Character Sets

We requested each writer to write a set of isolated characters (in given form with considerable spacing) and five pages of continuous texts (given texts without format constraints). The isolated characters cover the most frequently used characters in daily life, and the texts are mostly from news.

The total number of Chinese characters is very large, e.g., the standard set GB18030-2000 contains 27,533 characters, which are not yet exhausted. We estimate that the number of daily used characters is about 5,000, which is almost the maximum that ordinary educated people can recognize. For our handwriting data collection, we compiled a character set based on the standard sets GB2312-80 and Modern Chinese Character List of Common Use (Common Set in brief) [15]. The GB2312-80 contains 6,763 Chinese characters, including 3,755 in level-1 set and 3,008 in level-2 set.

[1]Application forms for using the CASIA databases can be found at http://www.nlpr.ia.ac.cn/databases/handwriting/Home.html

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.17

The Common Set contains 7,000 Chinese characters. Both the two sets have an appreciable number of characters that are unknown to ordinary people. We nevertheless collected the union of the two sets, containing 7,170 characters, for possible recognition of practical documents. We further added 15 Chinese characters that we met in our experience. We also collected a set of 171 symbols, including 52 English letters, 10 digits, and some frequently used punctuation marks, mathematics and physical symbols. The total number of character classes is thus $7,185 + 171 = 7,356$.

For collecting handwritten texts, we asked each writer to hand-copy five texts. We compiled three sets of texts (referred to as versions V1–V3), mostly downloaded from news Web pages except there are five texts of ancient Chinese poems in both V1 and V2. Each set contains 50 texts, each containing 150–370 characters. The three sets were used in different stages of handwriting data collection. The texts in each set were further divided into 10 subsets (referred to as templates T1–T10), each containing five texts to be written by one writer.

B. Data Collection

We collected handwriting data in three stages using three sets (versions) of templates. Each set has 10 templates to be written by 10 writers. A template has 13–15 pages of isolated characters and five pages of texts. For a template set, the isolated characters are divided into three groups: symbols, frequent Chinese and low frequency Chinese. The symbols are always on the first page, followed by Chinese characters. The first six templates of a set print the same group of frequent Chinese characters in six different orders by rotating six equal parts, and the last four templates print the low frequency Chinese characters in four difference orders. Rotation guarantees that each character is written equally in different time intervals for balanced writing quality. In addition, each template has five pages of different texts.

The three sets (versions) of templates are summarized in Table I. V1 and V3 have the same set of isolated characters. The number of isolated Chinese characters in V1 and V3 is actually 7,184, not 7,185, because the templates of V1 were designed earliest. The templates of V3 inherited the isolated character set of V1 and updated the texts. The frequent Chinese character set of V1 and V3 is actually the level-1 set of GB2312-80, which was commonly taken as a standard set of Chinese character recognition research.

Table I
SUMMARY OF TEMPLATES. IN EACH ROW, PAGES/SYMBOLS/CHINESE ARE FOR ONE TEMPLATE, WHILE TEXTS/CHARS ARE THE TOTAL NUMBERS.

Version	Template	#pages	#symbols	#Chinese	#texts/chars
V1	T1–T6	20	171	3,755	30/7,464
	T7–T10	19	171	3,429	20/4,918
V2	T1–T6	20	171	3,866	30/7,802
	T7–T10	18	171	3,319	20/5,196
V3	T1–T6	20	171	3,755	30/9,039
	T7–T10	19	171	3,429	20/6,016

For handwriting data collection using Anoto pen, all the template pages were printed on papers with dot pattern. On the printed template pages, each isolated character was written in the space below the pre-printed character, and each text was written on a separate page with the template text printed in the upper part. During writing, the online (temporal) data were recorded by the Anoto pen and later transmitted to computers. For offline data collection, the handwritten pages were scanned (in resolution of 300DPT) to obtain color images, which were segmented and labeled using annotation tools. Fig. 1 shows two scanned pages of isolated characters and handwritten text, respectively.

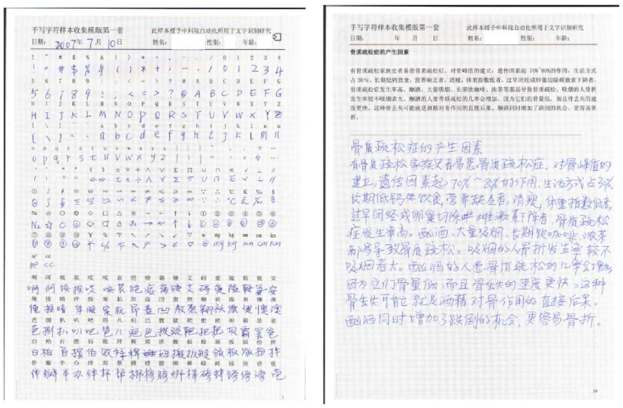

Figure 1. Scanned pages of isolated characters and handwritten text.

III. DATA ANNOTATION

We have different processing steps for segmenting and annotating online and offline data, and different steps for isolated character data and handwritten text data. For preparing annotation, the transcript characters (in GB codes) of each page (either online or offline, either isolated character or text) are ordered in the same layout as the handwritten page. For each page, the transcript (stored in a text file) has the same number of lines as the handwritten page, and the corresponding lines in the transcript and the handwriting have the same number of characters.

A. Annotation of Offline Data

From a scanned handwritten page, we need to first separate the handwritten characters from the background dots (pre-printed Anoto dot pattern) and the printed characters. We used two linear discriminant analysis (LDA) classifiers for pixel classification to separate the characters from background dots and separating handwriting from printed characters, respectively. Afterwards, the pixels of handwritten characters are stored in a gray-scale image.

For pages of isolated characters, since the handwritten characters are approximately evenly placed with large gaps, we simply segmented the lines according to horizontal projection and segmented the characters according to between-character gaps. If the number of lines or the number of characters in each line is inconsistent with the transcript, the human operator will be reminded to correct the segmentation errors.

For pages of handwritten texts, we used the annotation tool developed by our group [16]. The handwritten document image is first segmented into text lines using a connected

component clustering-based algorithm, with mis-segmented text lines corrected by human. The image of each line is matched with the transcript character string to group the connected components into characters aligned with the transcript string. Mis-segmentation and mis-labels were corrected by human operators. Fig. 2 shows an example of character segmentation and labeling by transcript mapping.

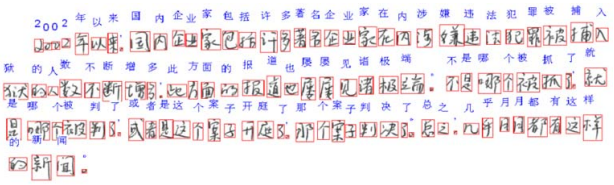

Figure 2. An example of text line transcript mapping. Each segmented character image is attached a label above it.

B. Annotation of Online Data

The segmentation and labeling procedure of online data is similar to that for offline data. The main difference is that, in addition to the spatial information, the temporal information of pen lift between adjacent strokes is also used for text line and character segmentation.

C. Data Format

We stored all the annotated data in writer-specific files. The online isolated character samples of each writer are stored in a file named xxx.POT (xxx is the writer index), and the offline character samples are stored in a file named xxx.GNT. A POT file stores multiple online character samples sequentially. Each sample has a record for total number of bytes, the class label (4-byte GB code), number of strokes, and sequence of (x, y) coordinates of stroke points with (-1,0) denoting pen lift. A GNT file stores multiple gray-scale character images sequentially. Each image has a record for total number of bytes, the class label (2-byte GB code), the width and height, and the bitmap (one byte per pixel). The gray-scale image has background pixels uniformly labeled as 255, and can be easily converted to binary image.

The handwritten text data are stored in files one per page, named after writer index-page number. The online text data of a page is stored in a file named xxx-Pyy.PTTS (yy is the page number), and the offline text data of a page is stored in a file named xxx-Pyy.DGR. A file stores the number of text lines and then the records of lines sequentially. The record of an online text line includes the number of characters, the number of strokes and the sequence of strokes (sequences of (x, y) coordinates), and then the sequence of character records. The record of each character includes the class label in 2-byte GB code, the number of strokes and the index numbers of corresponding strokes. The record of an offline text line includes the number of characters and the sequence of character records, each including the class label in 2-byte GB code, the position and the gray-scale image of the character.

IV. Statistics of Databases

From 2007 to 2010, we collected the online and offline handwriting data of 1,280 writers, among which 760 used the templates of V2, 520 used the templates of V1 and V3 (see Table I). We divided the isolated character samples of V2 into two datasets: (regardless of online/offline) DB1.0 (templates T1–T6) and DB1.2 (templates T7–T10). The online version of DB1.0, CASIA-OLHWDB1.0 in full name, is the one that we released in 2009 [14]. The 420 writers in DB1.0 were selected from the original 456 writers according to the percentage of legal characters [14]. Similarly, for DB1.2 we selected 300 writers of high percentage of legal characters from the original 304 writers of V2 templates T7–T10. We formed datasets DB1.1 and DB1.3 from the data of V1 and V3. The writers of DB1.1 include the 66 writers of V1 templates T1–T6 and selected 234 writers from the original 246 writers of V3 templates T1–T6. The writers of DB1.3 include the 44 writers of V1 templates T7–T10 and selected 156 writers from the original 164 writers of V3 templates T7–T10. Note that V1 and V3 have the same templates of isolated characters but different texts.

Having formed the isolated character datasets DB1.0–1.3, we formed the handwritten text datasets DB2.0–2.3 using the text data of the same writers of DB1.0–1.3. Specifically, the DB2.0 contains the handwritten text data of 420 writers of V2 templates T1–T6, DB2.2 contains the handwritten text data of 300 writers of V2 templates T7–T10, DB2.1 contains the handwritten text data of 66 writers of V1 templates T1–T6 and 234 writers of V3 templates T1–T6, and DB2.3 contains the handwritten text data of 44 writers of V1 templates T7–T10 and 156 writers of V3 templates T7–T10. Note that DB2.1 and DB2.3 have more different texts than DB2.0 and DB2.2 because they are formed of two sets of templates V1 and V3.

In total, we obtained eight datasets for either online or offline handwriting: isolated character datasets DB1.0–1.3 and handwritten text datasets DB2.0–2.3. Totally 1,220 writers contributed the data: 420 writers for DB1.0 and DB2.0, 300 writers for DB1.1 and DB2.1, 300 writers for DB1.2 and DB2.2, and 200 writers for DB1.3 and DB2.3. The set of Chinese characters in DB1.0 (3,866 classes) and that of DB1.1 (3,755 classes, level-1 set of GB2312-80) both consist of high frequency Chinese characters, and they have a large overlap of 3,740 characters. The set of Chinese characters in DB1.2 (3,319 classes) is a disjoint set of DB1.0.

Since the isolated character datasets DB1.0–1.2 are sufficient for character recognition of very large category set, we release DB1.0–1.2 and their corresponding handwritten text datasets DB2.0–2.2 for academic research. The handwritten sample data in these datasets was contributed by 1,020 writers. We keep the DB1.3 and DB2.3, containing handwriting data of 200 writers, for private and industrial purposes.

The released datasets of online data are named as OLHWDB1.0–1.2 (isolated characters) and OLHWDB2.0–2.2 (handwritten texts). The released datasets of offline data are named as HWDB1.0–1.2 (isolated characters) and HWDB2.0–2.2 (handwritten texts). The numbers of writers, character samples (including symbols and Chinese character samples/classes) of the isolated character datasets are summarized in Table II. We can see that for either online or offline data, the three datasets have about 3.9 million character samples and the number of Chinese character classes is as large as 7,185. Including the 171 symbol

classes, the total number of classes is 7,356.

Table II
STATISTICS OF ISOLATED CHARACTER DATASETS.

Dataset	#writers	#character samples		
		total	symbol	Chinese/#class
OLHWDB1.0	420	1,694,741	71,806	1,622,935/3,866
OLHWDB1.1	300	1,174,364	51,232	1,123,132/3,755
OLHWDB1.2	300	1,042,912	51,181	991,731/3,319
Total	1,020	3,912,017	174,219	3,737,798/7,185
HWDB1.0	420	1,680,258	71,122	1,609,136/3,866
HWDB1.1	300	1,172,907	51,158	1,121,749/3,755
HWDB1.2	300	1,041,970	50,981	990,989/3,319
Total	1,020	3,895,135	173,261	3,721,874/7,185

Table III
STATISTICS OF HANDWRITTEN TEXT DATASETS.

Dataset	#writers	#pages	#lines	#character/#class
OLHWDB2.0	420	2,098	20,573	540,009/1,214
OLHWDB2.1	300	1,500	17,282	429,083/2,256
OLHWDB2.2	299	1,494	14,365	379,812/1,303
Total	1,019	5,092	52,221	1,348,904/2,655
HWDB2.0	419	2,092	20,495	538,868/1,222
HWDB2.1	300	1,500	17,292	429,553/2,310
HWDB2.2	300	1,499	14,443	380,993/1,331
Total	1,019	5,091	52,230	1,349,414/2,703

In the handwritten text datasets, each writer originally contributed five pages of texts. Due to the failure of online trajectory data capturing, light intensity of strokes in scanned image or our mistakes in managing the data, some pages of online or offline data of a few writers were lost. So, OLHWDB2.2 contains the online data of 299 (not 300) writers, HWDB2.0 contains the offline data of 419 (not 420) writers, and the average number of pages retained per writer is also smaller than five. The numbers of writers, pages, lines and segmented character samples/classes are summarized in Table III. We can see that for either online or offline data, the three handwritten text datasets contain nearly 1.35 million segmented characters and the number of character classes involved is over 2,600. DB2.1 involves more character classes than DB2.0 and DB2.2 because it was formed from two sets of templates V1 and V3 and has 60 different texts, while DB2.0 and DB2.2 have 30 and 20 different texts, respectively. Though the online and offline data were produced by the same writers using the same templates, the number of character classes differs considerably (2,655 vs 2,703) between online and offline data. This is because the online and offline datasets have different missing pages, and even from the same handwritten text page, the online and offline data may have different missing characters and different mis-labeling. However, the differing classes have very few samples.

V. RECOMMENDATIONS OF USAGE

The CASIA Chinese handwriting databases have some favorable merits: (1) They have both online and offline data produced concurrently by the same group of writers; (2) They have both isolated character data and handwritten text data; (3) The data samples are stored writer by writer; (4) The offline samples are recorded in gray-scale images. We concede that the diversity of texts is not large, especially, the datasets DB2.0 and DB2.2, involving a large number of 720 writers, cover only 50 different texts. The dataset DB2.1 covers 60 different texts but involves a smaller number of 300 writers. Nevertheless, the diversity of texts does not influence the validity of handwritten text recognition performance because the language model is usually estimated from a general text corpus of huge size (at least 10 million characters, say).

For using the databases for research, we recommend standard partitioning into training and test sets, and propose some research scenarios.

A. Data Partitioning

We previously partitioned the OLHWDB1.0 dataset into a training subset of 350 writers and a test subset of 70 writers [14]. For uniform ratio in different datasets, we later reset the ratio of training writers and test writers as 4:1. The same partitioning of writers is taken between online and offline datasets and between isolated characters and handwritten texts datasets. In the following, we detail the partitioning of DB1.0–1.2, while DB2.0–2.2 have the same partitioning.

Following our previous file naming of OLHWDB1.0, the writers were numbered in decreasing order of re-substitution accuracies and divided into three grades. From each grade, 1/5 of writers are randomly selected for testing and the remaining 4/5 writers are used for training. We provide the users the lists of writer index numbers for training data and test data. The offline dataset HWDB1.0 follows the same partitioning.

For the datasets DB1.1 and DB1.2, we did not sort the writers according to the writing quality or recognition accuracy, but directly selected a ratio of writers randomly from each template (T1–T6 of V1 and V3, T7–T10 of V2) such that different templates have the same ratio of training to test writers. Since the isolated character templates T1–T6 are the same in V1 and V3, we did not differentiate V1 and V3, so that the text templates do not necessarily have the same ratio of training and test writers in the partitioning in DB2.1. For the datasets DB1.1 and DB1.2, DB2.1 and DB2.2, both the training writers and test writers are indexed as consecutive numbers.

B. Research Scenarios

Based on the CASIA online and offline handwriting databases, some typical research tasks of handwritten document analysis can be performed. Our recommendations are as follows.

1) Handwritten document segmentation: Our databases contain handwritten text pages produced by a large number of writers without instructions of format. There are 5,092 online text pages and 5,091 offline text pages in total. All the pages have ground-truths of text lines, and are convenient for training and evaluating text line segmentation algorithms.

2) Handwritten character recognition: For either online or offline handwritten character recognition, our databases contain large number of samples of alphanumeric characters and symbols (171 classes, nearly 1,020 sample per class), and particularly, Chinese characters of large category set. If the research focus is isolated Chinese character recognition,

we recommend to use the datasets DB1.0 and DB1.1. The Chinese character set of DB1.1 (3,755 classes as in level-1 set of GB2312-80) is commonly used in Chinese character recognition research. The DB1.0 has 3,740 Chinese character classes overlapping with the DB1.1. Merging the samples of 3,755 classes in DB1.0 and DB1.1 enables classifier training with large sample set.

3) Text line recognition: Character string recognition is the central task of handwritten text recognition. In Chinese handwriting, this amounts to text line recognition because a text line cannot be segmented into words prior to recognition. A feasible approach for character string recognition of large character set is integrated segmentation and recognition based on over segmentation. The isolated character datasets DB1.0–1.2 can be used to train a character classifier of large category set (7,356 classes). The geometric context model can be estimated from the training documents in DB2.0–2.2, and the language model is generally estimated from a corpus of pure texts. Finally, the performance of text line recognition is evaluated on the test documents in DB2.0–2.2. Handwritten text recognition can also be performed at page (or paragraph) level by combining the contextual information of multiple lines.

4) Handwritten document retrieval: In addition to text line segmentation and text line recognition research, the databases can also be used for document retrieval, including keyword spotting, handwritten text categorization and content-based retrieval. The researcher can utilize the text line segmentation ground-truth in the databases to focus on the retrieval task. As for text line recognition, document retrieval algorithms are recommended to be trained on the training documents in DB2.0–2.2, and evaluated on the test documents in DB2.0–2.2.

5) Writer adaptation: A merit of our databases is that all the samples are stored in writer-specific files. This ensures that all the samples in the same file are from the same writer and we know the writer index of each sample. This facilitates the research and evaluation of writer adaptation (including supervised adaptation and unsupervised adaptation) and style consistent field recognition.

6) Writer identification: In our databases, each writer has five handwritten text pages. We can perform experiments to judge whether two pages are from the same writer or not (writer verification) or classify a test page to a nearest reference page of known writer (writer identification) using either text-dependent or text-independent features.

VI. Conclusion

This paper introduces a pair of online and offline Chinese handwriting databases, containing both online and offline data produced concurrently by the same group of writers using Anoto pen. The databases are large in the sense that the number of writers is over 1,000 and either the online or offline isolated character datasets contain about 3.9 million samples of 7,356 classes, and the number of character samples in the online/offline handwritten text datasets is about 1.35 million. All the data has been segmented and labeled at character level and partitioned into standard training and test subsets. The databases can be used for research tasks of handwritten document segmentation, character recognition, text line recognition, document retrieval, writer adaptation and writer identification.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) under grants no.60825301 and no.60933010. The authors would like to thank Liang Xu for the algorithm for separating handwritten characters from images, and thank the members of the PAL group at the NLPR for checking through the annotated data.

References

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

[2] P.J. Grother, NIST Special Database 19: Handprinted Forms and Characters Database, 1995, http://www.nist.gov/srd/upload/nistsd19.pdf

[3] U.-V. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, *Int. J. Document Analysis and Recognition*, 5(1): 39-46, 2002.

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

[5] V. Märgner, H. El Abed, Databases and competitions: strategies to improve Arabic recognition, In: *Arabic and Chinese Handwriting Recognition*, S. Jaeger and D. Doermann (Eds.), LNCS Vol.4768, Springer, 2008, pp.82-103.

[6] F. Solimanpour, J. Sadri, C. Y. Suen, Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language, *Proc. 10th IWFHR*, 2006, pp. 3-7.

[7] H. Zhang, J. Guo, G. Chen, C. Li, HCL2000—A large-scale handwritten Chinese character database for handwritten character recognition, *Proc. 10th ICDAR*, 2009, pp.286-290.

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

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

[10] L. Jin, Y. Gao, G. Liu, Y. Li, K. Ding, SCUT-COUCH2009—A comprehensive online unconstrained Chinese handwriting database and benchmark evaluation, *Int. J. Document Analysis and Recognition*, to appear, 2011.

[11] S. Zhou, Q. Chen, X. Wang, HIT-OR3C: an opening recognition corpus for Chinese characters, *Proc. 9th IAPR Int. Workshop on DAS*, 2010, pp.223-230.

[12] C. Viard-Gaudin, P.M. Lallican, S. Knerr, P. Binter, The IRESTE On/Off (IRONOFF) dual handwriting database, *Proc. 5th ICDAR*, 1999. pp. 455-458.

[13] C.-L. Liu, Handwritten Chinese character recognition: effects of shape normalization and feature extraction, In: *Arabic and Chinese Handwriting Recognition*, S. Jaeger and D. Doermann (Eds.), LNCS Vol.4768, Springer, 2008, pp.104-128.

[14] D.-H. Wang, C.-L. Liu, J.-L. Yu, X.-D. Zhou, CASIA-OLHWDB1: A database of online handwritten Chinese characters, *Proc. 10th ICDAR*, 2009, pp.1206-1210.

[15] Modern Chinese Character List of Common Use (XianDai HanYu TongYong ZiBiao), http://wenke.hep.edu.cn/gfhz/html/tyzb4.asp

[16] F. Yin, Q.-F. Wang, C.-L. Liu, A tool for ground-truthing text lines and characters in off-line handwritten Chinese documents, *Proc. 10th ICDAR*, 2009, pp.951-955.