CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters

Da-Han Wang, Cheng-Lin Liu, Jin-Lun Yu, Xiang-Dong Zhou
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation of Chinese Academy of Sciences (CASIA)
P.O. Box 2728, Beijing 100190, P.R. China
{dhwang, liucl}@nlpr.ia.ac.cn

Abstract

This paper describes a publicly available database, CASIA-OLHWDB1, for research on online handwritten Chinese character recognition. This database is the first of our series of online/offline handwritten characters and texts, collected using Anoto pen on paper. It contains unconstrained handwritten characters of 4,037 categories (3,866 Chinese characters and 171 symbols) produced by 420 persons, and 1,694,741 samples in total. It can be used for design and evaluation of character recognition algorithms and classifier design for handwritten text recognition systems. We have partitioned the samples into three grades and into training and test sets. Preliminary experiments on the database using a state-of-the-art recognizer justify the challenge of recognition.

1. Introduction

Automatic recognition of unconstrained handwriting remains a great challenge: the performance of layout analysis, word/character segmentation and recognition is still far behind the human recognition capability. For design and evaluation of handwriting recognition algorithms and systems, the availability of large-scale, unconstrained handwritten dataset is necessary.

In the past decades, a number of databases in different languages have been published and have significantly benefited the research community. Most of the databases are of offline data (images converted from paper documents). Among them are the CENPARMI digits [1], CEDAR English words and characters [2], English sentence database IAM [3], Japanese Kanji character databases ETL8B and ETL9B, Korean database PE92 [4], Indian database of ISI [5], Arabic databases [6], Chinese databases HCL2000 [7] and HIT-MW [8], and so on. Databases of online handwritten data (trajectory data of strokes) are not so popular as offline ones because the collection of online data relies on special devices such as digitizing tablet, tablet PC and PDA. A few efforts in the area are the UNIPEN project [9], the Japanese online handwriting databases Kuchibue and Nakayosi [10][11], and the very recent Chinese online handwriting database SCUT-COUCH2008 [12]. The French database IRONOFF contains both online and offline data, collected by attaching paper on digitizing tablet while writing [13].

The current databases of Chinese handwriting are either too neat in writing quality or not large enough. The offline databases HCL2000 and CASIA (a subset of a large dataset collected by the Institute of Automation of CAS around 1990), both containing isolated character images of 3,755 categories, have been reported test accuracies higher than 98% [14][15], while the accuracy in realistic applications is far lower. The database HIT-MW has only 853 page images containing 186,444 characters. The online database SCUT-COUCH2008 contains samples of Chinese words, isolated characters (6,763 categories) and Pinyin, with 1,392,900 character samples in total, written by only 50 persons.

To support research on recognition of unconstrained Chinese handwriting, we have collected online and offline samples (isolated characters and continuous texts) written by 760 persons using Anoto pen on paper. Currently, the online isolated characters have been annotated. The online texts and offline characters and texts will be annotated in near future. This paper describes a subset of the online character samples, called CASIA Online Handwriting Database 1 (CASIA-OLHWDB1). It contains character samples of 4,037 categories (3,866 Chinese characters and 171 symbols) written by 420 persons. To demonstrate that the database is challenging, we conducted experiments using a state-of-the-art recognizer and obtained test accuracies of 92.44% (4,037 categories) and 92.91% (3,866 categories).

Our database is freely public to the academic community. The licensing information can be found at http://www.nlpr.ia.ac.cn/databases/CASIA-OLHWDB1.htm

2. Data Collection

We requested 760 persons (mostly university students) to write isolated characters and continuous texts. Each person wrote 171 symbols, either 3,866 frequent Chinese characters (60% persons) or 3,319 non-frequent Chinese characters (40% persons), and five pages of continuous texts (containing 1,000-1,400 characters). From the online samples of frequent characters written by 456 persons, we selected 420 sets (a set refers to the samples of a person) of high percentage of eligibility (some samples are not valid either because of sloppy writing or failure of pen trajectory caption by Anoto Pen) for release.

2.1 Character Set

The 171 symbols (Fig. 1) include 10 digits, 52 English letters, and some symbols that are frequently used.

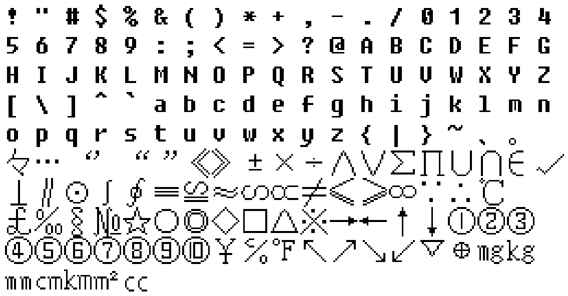


Figure 1. 171 symbols.

The set of Chinese characters include the 6,763 characters in GB2312-80 standard and other 422 frequently used characters. We reordered the 7,185 characters according to their frequencies in a Chinese text corpus and divided into 3,866 frequent characters and 3,319 less frequent ones.

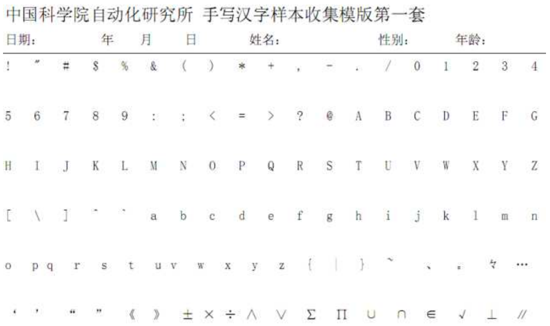
The 3,866 frequent characters (FC) are related to the level 1 set (L1, 3,755 characters) and level 2 set (L2, 3,008 characters) of GB2312-80 as follows. The FC has 3,740 characters in L1, 124 characters (Fig. 2) in L2, and two characters (啰瞭) outside GB2312-80. The L1 has 15 non-frequent characters lacking in the FC: 珐锟烩硷粳傈镊醛傈酞烃硒矽锗柞.

丐噩匕夭匾卦偎凫禀谒阱芙茉茴荞荠荸莺蒿薇
奕拗捺叽叨吆咧咪唠唧嗦嗤嘁嘀嘹幔岖徙猬馍
庵悴愕憔沐涮漩姊嬉缤缭缰玷璧杈桦榄楣榛榕
橄橘檐檩昙昵掰牍肴胧腌朦臊飒炫祠盹铐铛锉
秕秫鸠鹉鹦瘾癞衩裆蚣蚪蚓蚯蛉蜓蜈蜻蝠蝌蝙
蟆螃蟥蟋蟀笙笤筝箫簸翎翩麸跛跷踱蹂蹦雳霎
鲫鳄鳍鬓

Figure 2. 124 frequent characters in level 2 of GB2312-80.

2.2 Form Design

People tend to write sloppily when then get tired. So, the writing quality of thousands of characters changes gradually, though we did not pose any constraints of writing. To make different categories have approximately the same writing quality, we partitioned the 3,866 Chinese characters into six subsets and printed them in six forms, each form has a different order of six subsets, such that each subset located in six different positions in six forms. Each form begins with the 171 symbols, followed by six subsets of Chinese characters. The characters were printed on papers with dot pattern, and persons were required to write below the printed characters. Fig. 3 shows a part of the first page of a printed form. Each form has 15 pages of isolated characters and five pages of texts. Fig. 4 shows a part of a handwritten page captured by Anoto pen.



中国科学院自动化研究所 手写汉字样本收集模版第一套

日期： 年 月 日 姓名： 性别： 年龄：

! " # $ % & ( ) * + , - . / 0 1 2 3 4

5 6 7 8 9 : ; < = > ? @ A B C D E F G

H I J K L M N O P Q R S T U V W X Y Z

[\ ] ^ ` a b c d e f g h i j k l m n

o p q r s t u v w x y z { | } ~ 、 。 々 …

' ' " " 《 》 ± × ÷ ∧ ∨ Σ Π ∪ ∩ ∈ √ ⊥ //

Figure 3. Part of a printed form.

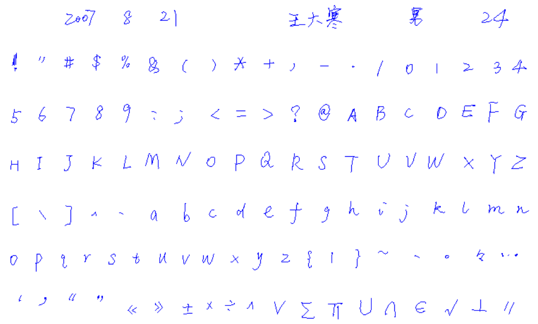


Figure 4. Part of a handwritten page.

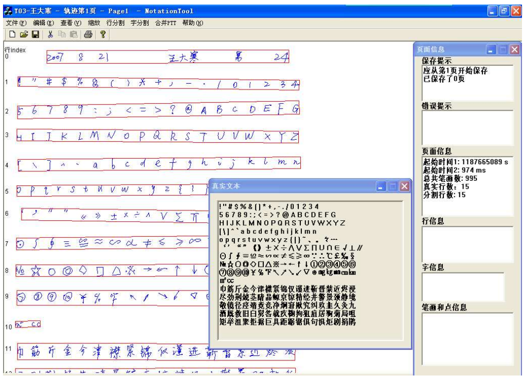
The writers were asked to write the characters in their most comfortable and familiar manner. No constraints were imposed to the quality of character shapes.

3. Dataset Labeling

In our data, each page is an online handwritten document with multiple lines and multiple characters in each line. We developed a software tool to first segment the text lines of each page, and then align the characters of each line with the text transcript (groundtruth). After alignment, each character sample is attached with a label (GB code). The format of our dataset is in binary.

3.1 Line Segmentation and Correction

Since the layout of the handwritten forms is neat, we used a simple technique (the pre-segmentation stage of the method of [16]) to group the text lines. Inevitably, there are some segmentation errors. The wrong number of segmented lines can be found automatically according to the transcript, and the mis-segmented lines can be corrected manually. Fig. 5 shows the result of line segmentation.



保存提示
应从第1页开始保存
已保存了0页

错误提示

页面信息
起始时间1: 1187665089 s
起始时间2: 974 ms
总共笔画数: 995
真实行数: 15
分割行数: 15

行信息

字信息

笔画和点信息

!"#$%&()*+,-./01234
56789:;<=>?@ABCDEFG
HIJKLMNOPQRSTUVWXYZ
[\]^`abcdefghijklmn
opqrstuvwxyz{|}~、。々—
‘’“”〔〕±×÷∧∨∑∏∪∩∈√⊥∥
⊙∫∮≡≌≈∽∝≠≤≥∞∵∴℃£‰§
№☆○●◇□△※→←↑↓①②③④⑤⑥
⑦⑧⑨⑩￥%℉↖↗↘↙▽㊣㎎㎏㎜㎝㎞
㎡㏄
巾筋斤金今津襟紧锦仅谨进靳晋禁近烬浸
尽劲荆兢茎睛晶鲸京惊精经井警景颈静境
敬镜径痉靖竟竞净炯窘揪究纠玖韭久灸九
酒厩救旧臼舅咎就疚鞠拘狙疽居驹菊局咀
矩举沮聚拒据巨具距踞锯俱句惧炬剧捐鹃

Figure 5. Line segmentation, the top line is page header. The text in small window is transcript.

3.2 Character Segmentation and Correction

Each text line (excluding the page header) is aligned with its transcript to segment the characters and attach labels to them. For most of the characters are well separated, we did not use a character recognizer for alignment. Instead, we merge the strokes into blocks according to off-stroke (pen lift) distances and merge consecutive blocks into characters according to between-block distances. When the number of segmented characters is different from the number in the transcript of the line, the user will be reminded to find and correct segmentation error.

Using mouse click, segmentation errors can be manually corrected by merging two characters, splitting one character to two, moving a stroke of one character to another, and breaking a stroke to two to separate two connected characters.

In labeling a text line, extra characters are deleted and mis-written characters are labeled as “abnormal” (can be deleted later). If some characters in the transcript are missing in the written text line, the text file of transcript should be modified to fit the written text line.

Fig. 6 shows some examples of character segmentation errors and correction.

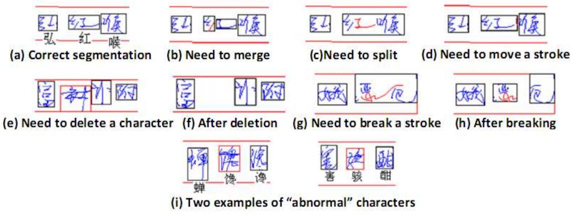

弘
红
喉
(a) Correct segmentation
(b) Need to merge
(c)Need to split
(d) Need to move a stroke
(e) Need to delete a character
(f) After deletion
(g) Need to break a stroke
(h) After breaking
蝉
馋
谗
害
骇
酣
(i) Two examples of “abnormal” characters

Figure 6. Examples of character segmentation errors and correction.

4. Dataset Statistics

From the sample sets of 456 persons (76 for each of six forms), we selected 420 sets with high percentage of sample eligibility (after deleting abnormal samples) for release.

4.1 Writers Distribution

All the writers come from the Institute of Automation of CAS and a national university in Beijing. Most of them are undergraduate or graduate students, originating from all areas of China.

Among the 420 writers, 333 (79.3%) are male, 85 are female, and the gender of 2 is unknown to us. Regarding the age, most of them are between 19 and 25. Table 1 gives more details of age distribution.

Table 1. Age distribution of 420 writers.

Age	#Persons	Percentage
18 or below	62	14.76%
19~25	314	74.76%
26~30	32	7.62%
31 or older	6	1.43%
Unknown	6	1.43%
Total	420	100%

4.2 Numbers of Abnormal Samples

Some sample sets of the 420 writers still have some abnormal or missing characters. After deleting the abnormal samples, the total number of valid samples is 1,694,741 (the ideal number is 420*4037=1,695,540), 785 samples are abnormal and 14 are missing. The distribution of abnormal samples in 420 sets is shown in Table 2. As we selected sets of high percentage of eligibility, the maximum number of abnormal/missing samples in a set is 14.

Table 2. Distribution of abnormal/missing samples.

Abnormal+missing	Number of sets
0	224
0~9	174
10~14	22

5. Preliminary Experiments

We have done some experiments on the database using a state-of-the-art recognizer [17]. The recognizer uses trajectory-based normalization and direction histogram feature extraction, and modified quadratic discriminant function (MQDF) classifier. The feature dimensionality is reduced from 512 to 160 by Fisher linear discriminant analysis (LDA). For trajectory normalization, we take linear normalization and moment normalization.

We first sorted the 420 sets of samples into different grades using a recognizer trained with all the sample data. After sorting the data into three grades, we divided the whole dataset into training and test subsets, with the same proportion in each grade. The recognizer is then re-trained on the training set and evaluated on the test set.

The 420 sets were evaluated using the recognizer (with linear normalization or moment normalization) trained with all the sample data: the resubstitution accuracies on each set are ordered in descending order. We found that the accuracies of recognizer with linear normalization are more consistent with the intuition of human perception than those with moment normalization. For example, a set of very regular samples (Fig. 7) is ranked 3rd position by linear normalization but 29th position by moment normalization. Linear normalization is more consistent to human intuition because it brings small shape deformation though its recognition accuracy is lower than that of moment normalization.

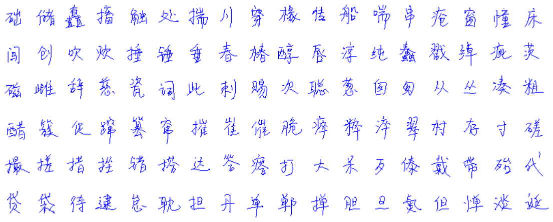


Figure 7. A sample set of regular writing.

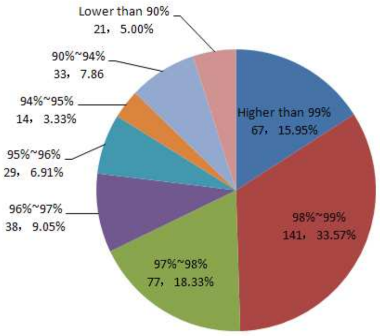


Figure 8. Distribution of resubstitution accuracies of 420 sets (recognizer with linear normalization).

Fig. 8 shows the distribution of resubstitution accuracies (by recognizer with linear normalization) of 420 sets. We grouped the sets with accuracies higher than 98% as grade 1 (G1), those with accuracies between 95% and 98% as grade 2 (G2), and those with accuracies lower than 95% as grade 3 (G3). For the following experiments, 5/6 of sets from each grade are used for training and the rest 1/6 for testing. Table 3 shows the distribution of three grades, and Table 4 shows the sample numbers of three grades.

Table 3. Distribution of three grades of 420 sets.

Grade	Resubstitut accuracy	#Sets	Percent	#Train	#Test
G1	⩾98%	208	49.52%	173	35
G2	95%~98%	144	34.29%	120	24
G3	<95%	68	16.19%	57	11
Total		420	100%	350	70

Table 4. Sample numbers of three grades.

Grade	#Train	#Test	Total
G1	698,091	141,251	839,342
G2	484,182	96,847	581,029
G3	229,977	44,393	274,370
Total	1,412,250	282,491	1,694,741

We then trained the recognizer on the training sample set and evaluate on the test set. We consider two category sets: 4,037 categories (symbol and Chinese) and 3,866 categories of Chinese characters. Table 5 and Table 6 show the test accuracies of 4,037 categories and 3,866 categories, respectively. Recognizers with linear normalization and moment normalization were used.

Table 5. Test accuracies of 4,037 categories.

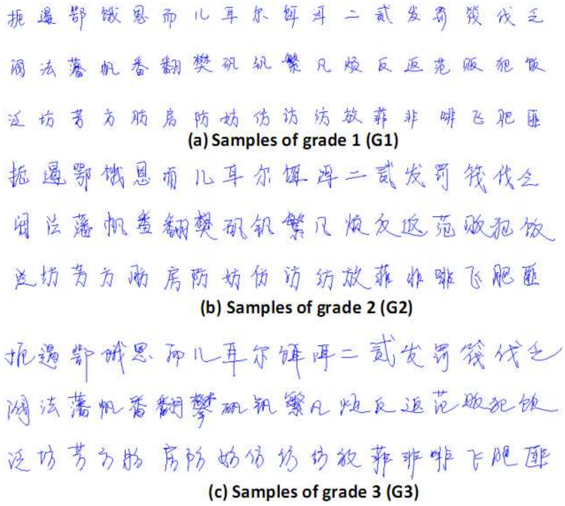
Grade	Linear norm	Moment norm
G1	93.18%	**96.30%**
G2	87.32%	**92.32%**
G3	69.86%	**80.41%**
Total	87.51%	**92.44%**

Table 6. Test accuracies (%) of 3,866 categories.

Grade	#Test	Linear norm	Moment norm
G1	135,267	93.73%	**96.86%**
G2	92,745	87.79%	**92.83%**
G3	42,512	70.03%	**80.10%**
Total	270,524	87.97%	**92.91%**

From Tables 5 and 6, it is evident that moment normalization yields significantly higher accuracies than linear normalization. The difference of accuracies between different grades of samples is also significant. Fig. 9 shows some samples of three grades.

It is noteworthy that even the highest accuracies, on average, 92.44% for 4,037 categories and 92.91% for 3,866 categories are much lower than the ones reported on other popular databases (e.g., 98.56% on HCL2000 [14], 97.84% and 98.24% on Japanese Kanji [17]). This confirms that our online database is challenging.


扼遏鄂饿恩而儿耳尔饵洱二贰发罚筏伐乏
阀法藩帆番翻樊矾钒繁凡烦反返范贩犯饭
泛坊芳方肪房防妨仿访纺放菲非啡飞肥匪
扼遏鄂饿恩而儿耳尔饵洱二贰发罚筏伐乏
阀法藩帆番翻樊矾钒繁凡烦反返范贩犯饭
泛坊芳方肪房防妨仿访纺放菲非啡飞肥匪
扼遏鄂饿恩而儿耳尔饵洱二贰发罚筏伐乏
阀法藩帆番翻樊矾钒繁凡烦反返范贩犯饭
泛坊芳方肪房防妨仿访纺放菲非啡飞肥匪

(a) Samples of grade 1 (G1)

(b) Samples of grade 2 (G2)

(c) Samples of grade 3 (G3)

Figure 9. Samples of three grades.

6. Conclusions and Discussions

We describe a large, publicly available database, CASIA-OLHWDB1, for research on online handwritten Chinese character recognition. The database contains 1,694,741 character samples written by 420 persons, in 4,037 categories (171 symbols and 3,866 Chinese characters). The samples are divided into three grades of quality and each grade is divided into approximately equal proportion of training and test subsets. Preliminary experiments using a state-of-the-art recognizer demonstrate the challenge of recognition. This leaves a big room for improvement and stimulates the community to seek for new recognition methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under grants no.60775004 and no.60825301. The authors thank all the members at the PAL group of NLPR for their collaboration in annotating the dataset.

References

[1] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam, Computer recognition of unconstrained handwritten numerals, *Proc. IEEE*, 80(7): 1162-1180, 1992.

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

[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] D.-H. Kim, Y.-S. Hwang, S.-T. Park, E.-J. Kim, P. S.-H, S.-Y. Bang, Handwritten Korean character image database PE92, *IEICE Trans. Information and Systems*, E79-D(7): 943-950, 1996.

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

[6] V. Margner, 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.

[7] J. Guo, Z. Lin, H. Zhang A new database model of off-line handwritten Chinese characters and its applications, ACTA *Electronica Sinica*, 28(5): 115-116, 2000.

[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] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, UNIPEN project of on-line data exchange and recognizer benchmarks, *Proc. 12th ICPR*, 1994, pp.29-33.

[10] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L. Higashigawa, K. Akiyama, On-line handwritten character pattern database sampled in a sequence of sentences without any writing instructions, *Proc. 4th ICDAR*, 1997, pp.376-381.

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

[12] Y. Li, L. Jin, X. Zhu, T. Long, SCUT-COUCH2008: A comprehensive online unconstrained Chinese handwriting dataset, *Proc. 11th ICFHR*, 2008, pp.165-170.

[13] 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.

[14] H. Liu, X. Ding, Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes, *Proc. 8th ICDAR*, 2005, pp.19–23.

[15] 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.

[16] X.-D. Zhou, D.-H. Wang, C.-L. Liu, Grouping text lines in online handwritten Japanese documents by combining temporal and spatial information, *Proc. 8th Int. Workshop on Document Analysis Systems (DAS)*, 2008, pp.61-68.

[17] C.-L. Liu, X.-D. Zhou, Online Japanese character recognition using trajectory-based normalization and direction feature extraction, *Proc. 10th IWFHR*, 2006, pp.217-222.