HIT-MW Dataset for Offline Chinese Handwritten Text Recognition

Tonghua Su
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
tonghuasu@hit.edu.cn

Tianwen Zhang
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
twzhang@hit.edu.cn

Dejun Guan
Heilongjiang Mobile, Harbin, China
guandejun@hl.chinamobile.com

Abstract

A Chinese handwritten text dataset, HIT-MW, is presented to facilitate the offline Chinese handwritten text recognition. Texts for handcopying are sampled from China Daily corpus with a stratified random manner. To collect naturally written handwriting, forms are distributed by postal mail or middleman instead of face to face. The current version of HIT-MW includes 853 forms and 186,444 characters that are written by more than 780 participants under an unconstrained condition without preprinted character boxes. Its lexical coverage of 3,041 characters is about 99.33% measured on China Daily corpus with about 80 million characters. Handwritten texts of HIT-MW mainly written by college students follow a balanced distribution both in sex and in department. It can be used to conduct Chinese text-line segmentation, segmentation-free recognition, and to verify the effect of statistical language model in a real handwriting situation.

Keywords: Standardization, Data acquisition, Optical character recognition, Handwritten Chinese text

1. Introduction

Standard datasets play crucial roles in handwriting recognition research. On the one hand, they provide a large number of training and testing data, resulting in high model fit and reliable confidence in statistic. On the other, they offer a means by which evaluation among different recognition algorithms can be performed. More and more handwriting researchers begin to pay much attention to the dataset standardization and evaluate their work using standard datasets.

Dozens of handwriting datasets have been published since 1990s. In 1992, CENPARMI [1] and PE92 [2] were reported. The former consists of unconstrained handwritten postcodes sampled from real mail pieces. The latter is a Korean character dataset written by 1000 writers. Two years later, CEDAR [3] and CAMBRIDGE [4] were released. Similar to CENPARMI, CEDAR is also collected from real mail pieces. What's more, it includes a subset of handwritten city words extracting from mail addresses. CAMBRIDGE is the first handwritten English text dataset with a large vocabulary, which is written by a single writer in an unconstrained domain and used for writer-dependent handwriting recognition. Following that, the first version of IAM was put forward in 1998 [5], then the second version in 2002 [6], adapting some ideas from CAMBRIDGE. It is written by multiple writers and the texts for handcopying are progressively taken from the Lancaster-Oslo/Bergen (LOB) corpus. In 2000, a hand-printed Chinese character dataset named HCL2000 [7] and a handwritten Geek dataset named GRUHD [8] were published. Writers in HCL2000 are asked to write a comprehensive set of the First Level Chinese characters of GB2312-80 and the characters should be carefully written within a preprinted character box. GRUHD consists of two subsets. One includes hand-printed Greek characters and digits, the other an unconstrained Greek poem that can be used to conduct text-line segmentation experiments.

English handwriting recognition is one of the most thoroughly studied branches not only in recognition strategy but also in dataset standardization. There are three different recognition strategies: segmentation-based recognition, segmentation-free recognition, and holistic recognition [9]. When arranging the English handwritten datasets chronologically, we find that the handwritten unit has transmitted from digit or letter to city name, further to sentence and that application fields have expanded from small lexicon domains, such as bank check reading and address recognition, to large lexicon and general unconstrained domains [10, 11].

There are four Chinese handwriting datasets, namely, ETL-8/ETL-9 [12, 13], IAAS-4M [14] before 1990s, and recently HCL2000. All of them are hand-printed character-level datasets: Each participant is asked to write a large set of Chinese characters; each character had better be carefully written within a preprinted character box. However, ETL-8 and ETL-9 are seldom used in China because of the culture and writing style differences existing between China and Japan. IAAS-4M and HCL2000 are two popular datasets for general hand-printed Chinese character recognition used in China. Since both of them are character-level, the recognition stage must be performed after character segmentation. Just as Sayre's paradox [15] goes, segmentation is prone to error and difficult to make correction afterward. In fact, much of the error rate can be attributed to imperfect segmentation. Moreover, there are not enough data to support segmentation experiments, since the standard Chinese datasets include only characters. As a tradeoff, such experiments are conducted on Chinese mail addresses, though the number of them is limited. In

addition, there is no segmentation-free recognition of Chinese handwriting in general unconstrained manner yet. Indeed, a large handwritten Chinese text-level dataset is in great need.

Inspired by CAMBRIDGE and IAM, we put forward the first handwritten Chinese text dataset, HIT-MW[1]. Comparing to CAMBRIDGE and IAM, our dataset possesses at least three virtues. First, the handwriting is naturally written with no rulers that can be used to make the text-line straight by and large. This feature makes it suitable for conducting experiments on Chinese text-line segmentation. Second, the underlying texts for handcopying are sampled from China Daily corpus in a systematic way and the writers are carefully chosen to give a balanced distribution. Third, it is collected by mail or middleman instead of face to face, resulting in some real handwriting phenomena, such as miswriting and erasing. Besides text-line segmentation, the HIT-MW is fit to research segmentation-free recognition algorithms, to verify the effect of statistical language model in real handwriting situation, and to study the nerve mechanism of Chinese handcopying activity.

The flowchart of HIT-MW is shown in Figure 1. In addition, we present its basic statistics. The next section describes the sampling strategy. Then the handwriting collection and handwriting processing are discussed in section 3 and section 4, respectively. Section 5 analyses the basic statistics of the dataset. Finally, concluding remarks are given in section 6.

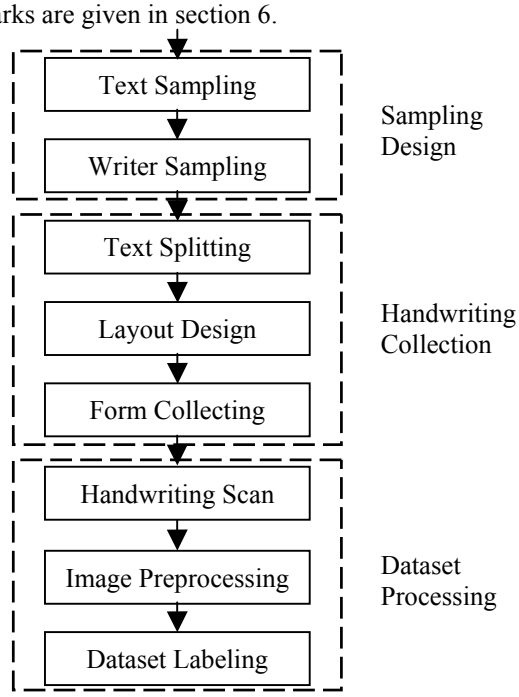

Figure 1. Flowchart of HIT-MW Dataset.

2. Sampling Schemes

Our dataset is to make a reasonable representation of the real Chinese handwriting, so it is crucial to carefully design sampling schemes. In this section, we describe two sampling schemes, dealing with objective writers and electronic data respectively.

2.1. Writer Sampling

We determine our potential users to be college students, government clerk graduated from higher school, and senior students in high school who are potential college students in the next year. There are three reasons. First, according to the handwriting theory, the handwriting goes into a stable and consistent state at 25 years old, and after that there is little change. Second, the college students are enrolled throughout the country, so the handwriting by them can be seen as samples from the whole country. This diminishes the sampling errors to some degree. Third, it is the well-educated people that are potential users of handwriting recognition, such as personal notes and manuscripts transcription.

Due to special users oriented, we need not sample the writers randomly. In fact, we divide the country into three regions, i.e., north region, middle region, and south region, and select one city handy from each region. Just using this simple sampling method, we obtain balanced writer samples (see section 5 for more details).

2.2. Text Sampling

We choose China Daily corpus as the data source of our dataset. In the natural language processing field, China Daily is used as Chinese written language corpus, since it covers a comprehensive topics such as politics, economics, science and technology, culture, et al. Using corpus as our data source instead of chaotic electronic texts demonstrates three advantages: Linguistic context is automatically built in; Dataset can be easily expanded with tremendous texts to sample from; More frequently a character occurs, more training samples it possess. As a result, our dataset can be collected in a progressive way and is helpful to conduct the linguistic post-processing after the recognition stage.

We sample texts with a stratified random manner. To reserve more data for future expansion, we only use texts of the China Daily 2004. We first divide texts into 12 groups according to month. Then we randomly draw 25 texts without replacement from each group. In this way, we obtain a compact and sound approximation to Chinese written language.

3. Handwritten Text Collection

As soon as the texts are extracted, it's time to start the collection process. Initially, we split each text into small and manageable segments. After several trials, we make them about 200 characters consisting of few complete sentences each. Next we format them into a clear and uniform layout. To design an informative layout, some considerations have been taken. Whenever all those have been done, we distribute forms to writers. Finally, we select forms according to special criteria.

[1] HIT is the abbreviation of Harbin Institute of Technology, and MW is the abbreviation of Multiple Writers.

3.1. Text Splitting

Texts previously sampled from corpus should be split into smaller segments. The number of characters in texts ranges from tens to thousands, which is inconvenient to distribute. In order to split each of them into a series of reasonable-size text segments, we consider the following two factors. First, it is wise to avoid breaking complete sentence, in which as much linguistic context as possible can be held. Some punctuation marks, the period, the exclamation mark, the question mark, and combination of them with quotation marks, serve as sentence end. Others, such as the semicolon, the dash, and ellipsis mark can also be selected as sentence end if necessary.

Second, segment should have a reasonable number of characters. If it is too short, the writer's style and handwriting variability are hardly obtained. In the opposite case, it makes tired the writer's hand-muscle and vision-muscle, which in turn mostly makes the handwriting illegible. Moreover, we will not collect the handwriting completely when big-size characters are presented.

Based on these two factors, we conduct simulated experiments several times. It seems that segments between 50 and 400 characters are acceptable. The further discussion is presented in the next subsection.

3.2. Layout Design

When we print text segments as forms, it is the layout that serves as an interface to writers. Obviously, how to make it friendly and informative is a nontrivial task. The design of layout follows three criteria. First, the layout is simple and clear. Each form is divided into three distinct blocks: guideline block, text block, and writing block. The horizontal lines are used to separate the adjacent blocks and the faces of the font to discern different information within block.

Second, we compress the writing guidelines to give more space reserved for handwriting. We make our commands concise by using short phrases and arrange them within five text lines with small font.

Third, we make use of implicit restrictions. In some cases, we want the writer to follow a special pattern, but it has difficulties to express in words. For example, we expect that the handwriting has a relatively small skew angle, but if we express it as a command, it will make the writer too careful to write naturally. Then we use horizontal lines both at top and bottom as references. It can help the writer know whether his handwriting is skew or not, and make some remedies reduce the skew adaptively.

After several recursions of feedback and modification, the final layout is illustrated in Figure 2 (The writing block showed here is scaled down vertically to make the graph smaller). Each form is identified by a 4-pair-digit code and each pair stands for certain meaning, e.g. 04070207 means that it is the seventh text segment of the second text sampled from July 2004.

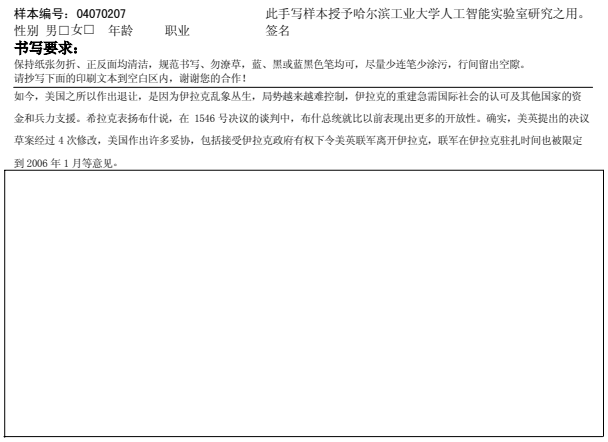
样本编号：04070207　　此手写样本授予哈尔滨工业大学人工智能实验室研究之用。
性别 男□女□　年龄　　职业　　签名
书写要求：
保持纸张勿折、正反面均清洁，规范书写、勿潦草，蓝、黑或蓝黑色笔均可，尽量少连笔少涂污，行间留出空隙。
请抄写下面的印刷文本到空白区内，谢谢您的合作！
如今，美国之所以作出退让，是因为伊拉克乱象丛生，局势越来越难控制，伊拉克的重建急需国际社会的认可及其他国家的资金和兵力支援。希拉克表扬布什说，在 1546 号决议的谈判中，布什总统就比以前表现出更多的开放性。确实，美英提出的决议草案经过 4 次修改，美国作出许多妥协，包括接受伊拉克政府有权下令美英联军离开伊拉克，联军在伊拉克驻扎时间也被限定到 2006 年 1 月等意见。

Figure 2. An Illustration of Layout.

3.3. Form Distributing and Selecting

Forms are distributed by mail or middleman instead of face to face. This makes the writers write naturally and impossible to tailor the handwriting for easy recognition, not exactly knowing what their handwriting will be used.

Once a pile of handwriting forms are collected, we accept the legible ones, and the illegible or lost ones are reprinted and distributed again. Handwriting is thought as legible, if it runs from left to right, its contents are what we have appointed (Few miswriting or/and erasing are allowed), and a majority of it can be read correctly by human.

4. Dataset Processing

Accepted handwriting is scanned into computer as digital image and then pixel-level processing is applied on it. The processing includes frame eliminating and binarization to give a clean and compact representation of the handwriting. Eventually, we transcribe the handwriting's ground truth that will serve as standard answers when calculating the recognition rate.

4.1. Handwriting Scanning

Each writing block of legible forms is digitized by Microtek ScanMaker 4180. The resolution is set to 300dpi. Images are saved as grayscale BMP files with no compression and named after their forms' code. The average storage space of each image is about 2.1M bytes.

4.2. Image Preprocessing

We perform image preprocessing on each scanned image. First, we eliminate the frame lines enclosing the writing block. We deal with them in an automatic way, and manually eliminate them once the lines are off standard positions. We pay special attention to preserving the smoothness of its strokes intersecting the frame lines.

Afterwards, we binarize handwriting image using Otsu algorithm [16]. The binary image is named after the grayscale image and a letter "b" is inserted as the prefix.

The black-white version of handwriting image named 04070207 is showed in Figure 3.

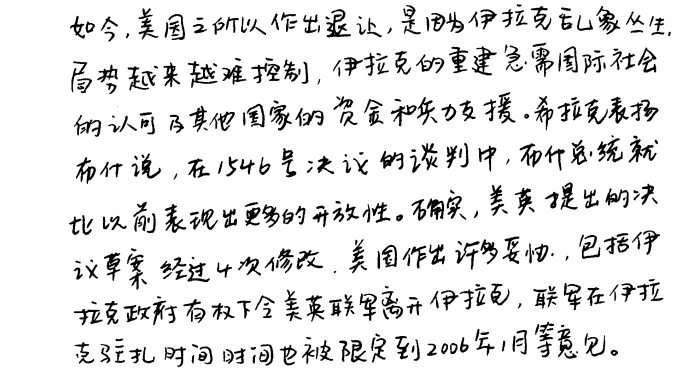

Figure 3. Binary Image of Handwriting Sample Named 04070207.

4.3. Dataset Labeling

The ground truth acts as the standard answers to the handwriting image. To evaluate the performance, the output of recognition engine is compared with the ground truth. That is to say, labeling the dataset to generate its ground truth is the preliminary stage for recognition system development.

Fortunately, we have already saved the electronic data of each text segment. Generating the ground truth file involves two different level alignments: a text-line level alignment and a character level alignment. The former makes text segment produce a new line where corresponds to the end of each handwriting text line. The latter crosses off the deleted characters from each segment, key in the inserted characters and modify the substituted characters. An example of labeled ground truth is illustrated in Figure 4.

如今，美国之所以作出退让，是因为伊拉克乱象丛生，
局势越来越难控制，伊拉克的重建急需国际社会
的认可及其他国家的资金和兵力支援。希拉克表扬
布什说，在 1546 号决议的谈判中，布什总统就
比以前表现出更多的开放性。确实，美英提出的决
议草案经过 4 次修改，美国作出许多妥协，包括伊
拉克政府有权下令美英联军离开伊拉克，联军在伊拉
克驻扎时间时间也被限定到 2006 年 1 月等意见。

Figure 4. The Ground Truth of Figure 3.

Note that, we don't label the ground truth character by character. This is determined by our research goal. Our recognition engine follows a segmentation-free strategy. Since the output of our engine is a row of characters, labeling each character's location is needless.

5. Dataset Statistics

The HIT-MW dataset is the first collection of Chinese handwritten texts in handwriting recognition domain. More than 780 participants produce their handwriting naturally. In this section we will present HIT-MW's features by a data-driven way. Due to space limitation, we only describe the fundamental statistics. Other features such as miswriting and erasing phenomena will be reported elsewhere.

We have collected 853 legible Chinese handwriting samples. There are 186,444 characters in total including letters, punctuations besides Chinese characters, and these characters lead to 8,664 text lines. By simple computation, we get following statistics: Each sample has 10.16 text lines; each text line has 21.51 characters; each sample includes 218.57 characters.

Moreover, the lexicon of the dataset has 3,041 entries. In other words, each character averagely occurs 61.31 times. To check its representative capability, we plot its coverage over China Daily corpus with 79,509,778 characters in Figure 5. Note that, the corpus has already excluded the data of China Daily 2004 to give objective coverage estimation. From the graph, we can see that a 1,800-character lexicon covers 97.60% of the corpus, and the full-size lexicon 99.33% of the corpus. This good coverage shows our sampling schemes work well.

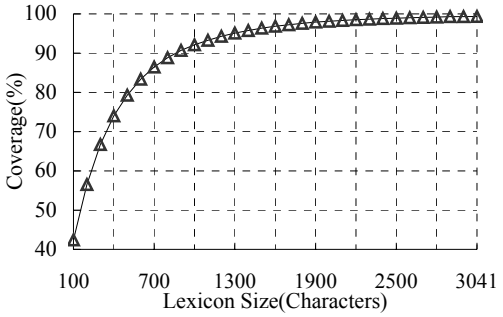

Figure 5. Lexicon Size of HIT-MW versus Coverage of China Daily Corpus.

Further, we calculate the writer's distribution. We mark the three sampled cities as City A, City B, and City C, respectively. From the view of city distribution in Figure 6, the sampled writers are mainly from City A with a proportion of 67 percent. Seen from Table 1, the department distribution of writers is near to that calculated from real data of college students of 2004[2]. Similarly, Table 2 shows that the sex distribution of our dataset has a good coincidence with that calculated from real educational statistics of 1998[3].

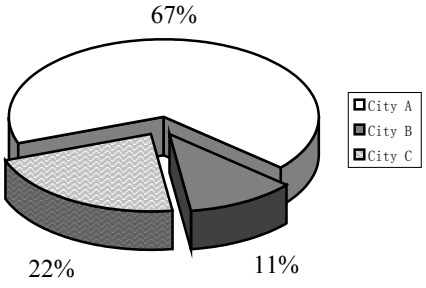

Figure 6. Sampling Percentage of Three Regions.

[2] The data are calculated from "China statistical yearbook 2005".

[3] The data are calculated from "Educational statistics yearbook of 1999".

Table 1. Writers from Science and Engineering Departments versus College Students of 2004 from that.

Items	Of 2004	Sampled
Percentage (%)	61.37	60.69

Table 2. Sex Distribution Comparison between Writers Sampled and Students of Year 1998.

Items		Percentage (%)
Boy Students Of 1998	High School	57.26
	Higher School	63.29
Boy Writers Sampled	High School	57.25
	Higher School	62.54

In summary, both the distribution of writer and the coverage of lexical entry show the effectiveness of the proposed sampling schemes.

6. Discussion and Conclusion

The handwritten Chinese text dataset discussed in this paper addresses several important aspects not covered by most other datasets. It is naturally written by multiple writers. As a result, there are real text lines and real handwriting phenomena, such as miswriting and erasing. In addition, not only the texts for handcopying are well sampled, but also the writers are carefully determined, resulting in high lexical coverage over China Daily corpus and a balanced writer distribution both in sex and in department.

The original purpose of the HIT-MW dataset is to facilitate the fundamental research of offline Chinese handwriting recognition from new perspective. There is no attempt to replace the Chinese character datasets already existing. On the contrary, they can be used to overcome HIT-MW's data sparseness derived from natural language.

Acknowledgement

We would like to thank Yiping Deng, Hui Xia, Ling Song, Di Zhang, Xuecai Yu, Haidan Xie, Yufeng Sun, Cuan Su and Guangjin Shao for their collaborations. We would also like to thank Haijing Wang and Yu Zhou for their valuable suggestions.

References

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

[2] 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 Transactions on Information and Systems*, Vol. E79-D, No. 7, pp. 943-950, 1996.

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

[4] A. W. Senior, A. J. Robinson, "An off-line cursive handwriting recognition system", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, pp. 309-321, 1998.

[5] U. V. Marti, H. Bunke, "A full English sentence database for off-line handwriting recognition ", *Proceedings of the 5th International Conference on Document Analysis and Recognition*, Bangalore, India, 1999, pp. 705-708.

[6] U. Marti, H. Bunke, "The IAM-database: an English sentence database for off-line handwriting recognition", *International Journal on Document Analysis and Recognition*, Vol. 5, No. 1, pp. 39-46, 2002.

[7] H. Zhang, J. Guo, "Introduction to HCL2000 database", *Proceedings of Sino-Japan Symposium on Intelligent Information Networks*, Beijing, 2000.

[8] E. Kavallieratou, N. Liolios, E. Koutsogeorgos, N. Fakotakis, G. Kokkinakis, "The GRUHD database of Greek unconstrained handwriting", *Proceedings of 6th International Conference on Document Analysis and Recognition*, Seattle, WA, USA, 2001, pp. 561-565.

[9] R. G. Casey, E. Lecolinet, "A survey of methods and strategies in character segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 7, pp. 690-706, 1996.

[10] A. Vinciarelli, S. Bengio, H. Bunke, "Offline recognition of unconstrained handwritten texts using HMMs and statistical language models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 6, pp. 709-720, 2004.

[11] M. Zimmermann, H. Bunke, "N-Gram Language Models for Offline Handwritten Text Recognition", *9th International Workshop on Frontiers in Handwriting Recognition*, Kokubunji,Tokyo, Japan, 2004, pp. 203-208.

[12] S. Mori, K. Yamamoto, H. Yamada, T. Saito, "On a handprinted Kyoiku-Kanji character data base", *Bull. Electrotech. Lab.*, Vol. 43, No. 11-12, pp. 752-773, 1979.

[13] T. Saito, H. Yamada, K. Yamamoto, "On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis", *IEICE Transactions*, Vol. J68-D, No. 4, pp. 757-764, 1985.

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

[15] K. Sayre, "Machine recognition of handwritten words: A project report", *Pattern Recognition*, Vol. 5, No. 3, pp. 213-228, 1973.

[16] N. Otsu, "A threshold selection method from gray-level histogram", *IEEE Transactions on System, Man, and Cybernetics*, Vol. SMC-9, No. 1, pp. 62-66, 1979.