HCL2000—A Large-scale Handwritten Chinese Character Database for Handwritten Character Recognition

Honggang Zhang, Jun Guo, Guang Chen, Chunguang Li
Beijing University of Posts and Telecommunications
Pattern Recognition and Intelligent System Laboratory
{zhhg, guojun, chenguang, lichunguang}@bupt.edu.cn

Abstract

In this paper, we present a large scale off-line handwritten Chinese character database-HCL2000 which will be made public available for the research community. The database contains 3,755 frequently used simplified Chinese characters written by 1,000 different subjects. The writers' information is incorporated in the database to facilitate testing on grouping writers with different background such as age, occupation, gender, and education etc. We investigate some characteristics of writing styles from different groups of writers. We evaluate HCL2000 database using three different algorithms as a baseline. We decide to publish the database along with this paper and make it free for a research purpose.

1. Introduction

Offline handwritten Chinese character recognition has been studied for over 30 years, many systems have claimed high recognition accuracies but few have been applied to real world applications. What limited evaluations exist of recognition accuracy tends to be lack of public available large scale databases. Researchers used their own collected databases to evaluate system performances [1] [6] [12]. Thus the results often can't be compared directly. Due to the similarity between many Japanese and Chinese characters, Japanese character databases often have utility in the Chinese recognition setting. Data collection and evaluation play a very important role in the development of automatic object recognition technologies. For example, the face recognition community has benefited from the Facial Recognition Technology (FERET) database, which includes 2,413 still facial images, representing 856 individuals. The large datasets have spurred the development of new algorithms. The independent evaluations have provided an unbiased assessment of the state-of-the-art in the technology and have identified the most promising approaches. In addition, the evaluations have documented two orders of magnitude improvement in performance from the start of the FERET program through the Face Recognition Vendor Test (FRVT) in 2006. The similar efforts have been made in many other object recognition tasks. However, there is no large scale handwritten Chinese character database public available to date. The state-of-the-art offline handwritten Chinese character databases are ETL character database, KAIST database, JEITA, and IRTI. ETL character databases were collected at the Electrotechnical Laboratory under the cooperation with Japan Electronic Industry Development Association, universities, and other research organizations. These databases ETL1-ETL9 contain about 1.2 million hand-written and machine-printed character images that include Japanese, Chinese, Latin, and numeric characters for character recognition researches. Character images of the databases were obtained from scanning OCR (Optical Character Recognition) sheets or Kanji printed sheets with a scanner. All databases ETL1 - ETL9 are gray-valued image data. ETL8 and ETL9, binarized image data (ETL8B and ETL9B) are open to the public. ETL9 consists of 2,965 Chinese and 71 Hiragana, 200 samples per class written by 4,000 writers. ETL8 consists of 881 classes of handwritten Chinese and 75 classes of Hiragana, 160 samples per class written by 1,600 writers. There are 60 × 60, 64 × 63, 72 × 76 and 128 × 127 pixels versions of character images. The character image files consist of more than one record, which has a character image and ID information with a correct code. KAIST Hanja1 and Hanja2 were collected by the Korea Advanced Institute of Science and Technology. The Hanja1 database has 783 most frequently used classes. Each class contains 200 samples collected from 200 writers in experimental settings. The Hanja2 database has 1,309 samples collected from real documents. The number of samples in the Hanja2 database varies with the class. The image quality of Hanja1 is quite clean, while the Hanja2 database is very noisy. JEITA-HP was originally collected by Hewlett-Packard Japan and later released by JEITA (Japan Elec-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.15

IEEE computer society

tronics and Information Technology Association). It consists of two datasets: Dataset A (480 writers) and Dataset B (100 writers). Generally, Dataset B was written more neatly than Dataset A., The entire database consists of 3214 character classes (2,965 kanji, 82 hiragana, 10 numerals, 157 other characters (English alphabet, katakana and symbols)). The most frequent hiragana and numerals appear twice in each file. Each character pattern has a resolution of 64 × 64 pixels, which are encoded into 512 bytes. ITRI was collected by the Industrial Technology Research Institute (Taiwan, China). It contains 5,401 handwritten Chinese classes with 200 samples per class. In this paper, we present a large scale handwritten Chinese character database, HCL2000 (Handwritten Character Library 2000), to facilitate handwritten Chinese recognition research. The word ”2000” in HCL2000 indicates the database contains 2000 samples collected in the year of 2000. The collection of HCL2000 was supported by Chinese Government through the China 863 high-tech project. Unlike many existing handwritten Chinese character databases, HCL2000 has two characteristics: one is its large scale in number, the total image sample number is 3,755,000, the other is that it contains writer’s information, which can help researchers to study the styles of calligraphy of different writers and writer identification. The writers’ information includes the age, occupation, gender, education, and address etc. We describe the system models of HCL2000. We discuss the evaluation based on the HCL2000 and present the performances of three algorithms as a baseline. HCL2000 has been used for a few researchers to evaluate their systems in the past [9] [3] [15] [11] [10]. To promote handwritten Chinese character recognition, we decide to publish the database upon publishing this paper. We will make the database free for research communities for a research purpose.

2. The System Model

We use a system model to control the information within HCL2000 that contains not only the Chinese character image, but also information about writers. This model provides a mechanism to manage two databases. Fig.1 illustrates the system database model. HCL2000 includes two sub-databases, one is the handwritten Chinese character sample’ database, and the other is writers’ information database. In order to use the two databases easily, two management systems are built respectively. The user of the character sample database can query “who wrote these characters” by accessing the information of the writer while browsing Chinese character images, and the user also can query “how is the writing style of the writer like” by querying the Chinese character image database while browsing the information of the writer. By using the management system, a user can view all character samples of a writer (as shown in Fig.2(b)), or all the characters samples of some writers, or selected characters of a writer or selected characters of some writers (as shown in Fig2.(a)).

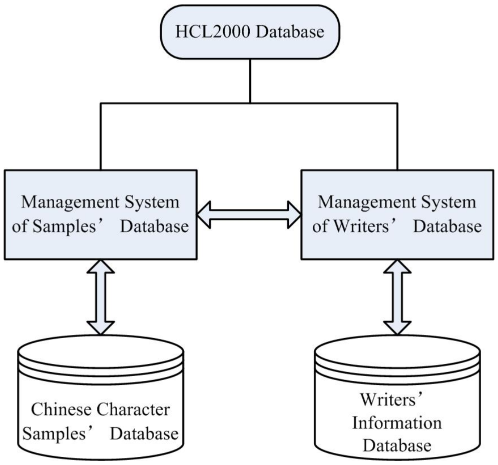
HCL2000 Database
Management System of Samples’ Database
Management System of Writers’ Database
Chinese Character Samples’ Database
Writers’ Information Database

Figure 1. System model of HCL2000

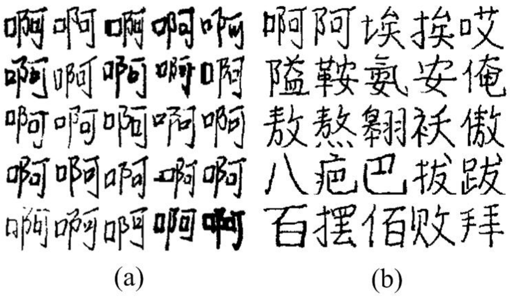

Figure 2. Some Samples from HCL2000, (a)same character of some writers, (b)some characters of a writer

2.1. The Handwritten Chinese Character Samples Database

The character samples are organized by different writers and stored in PID (Personal Identification) files, as shown in Fig.3(a), the samples written by the same writer are placed in the order of section code, as shown in Fig.3(b). The file format of a PID file is defined as the file header and character samples. There is a file header of 512 bytes in every file, which is used to contain the information including PID, scanning precision both in horizontal and vertical direction, the size of the whole file and so on. Each Chinese character

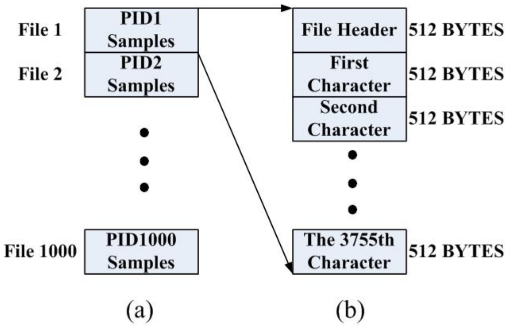

Figure 3. Organization of character samples

sample is described as 64×64 binary pixels, which is 512 bytes in size, shown as Fig.3 (b).

2.2. The Writers' Information Database

Besides PID about the writer, the writers' information includes the gender, age, occupation, education, tools of writing. All writers' information are stored in the file header by 512 bytes, the detailed definition of a file header is listed in Table 1.

Table 1. Content of the File Header

Byte Position	No. of Bytes	Type	Content of Record
0-2	2	Integer	Serial Number
2-3	1	Integer	Width of Character
3-4	1	Integer	Height of Character
4-14	10	ASCII	Name of the Writer
14-15	1	Integer	Gender
15-16	1	Integer	Age
16-17	1	Integer	Occupation
17-117	100	ASCII	Address
117-118	1	Integer	Education
118-511	393		Reserved

We assume the handwriting style varies from education, age, and occupation; therefore, different writers with different backgrounds were invited to write the samples according the proportion. Our research indicates that,writers with better education write character more formally, and writers with the same occupation have similar writing style. Table 2 shows the proportion of writers' educations, Table 3 shows the proportion of writers' ages and Table 4 is the proportion of writers' occupations in HCL2000.

Table 2. Proportion of Writers' Educations (%)

Above undergraduate	Technical College	Specialized Middle School	High School	Below Middle School
15.5	17	17.5	23.5	26.5

Table 3. Proportion of Writers' Ages (%)

Below Twenty	Twenties	Thirties	Forties	Fifties	Above Sixty
25.7	41.9	12.3	12.7	5	2.4

3. The Evaluation Based on Writers' Information

We have investigated the relationship between error rates with writers' information, and discovered some interesting results. We used 700 sets labeled from xx001 to xx700 for testing, and the rest sets labeled from hh001 to hh300 for training. The recognition process has two steps. First, an input image was preprocessed and roughly classified by the directional element feature (DEF), 30 candidates were selected to the classifier based on cosine transformation [4]. We first obtained error rates on the samples according to the writers' education, as shown in Table 5, the error rate decreases with the education degree increase, but increase on above undergraduate, because the sampling number is large and they are not serious to write the samples. Fig.4 is the samples from two kinds of education degrees, Fig.4(a) shows some examples written by subjects from technical colleges, Fig.4(b) are examples written by undergraduate students from universities. Table 6 illustrates the error rate among different ages, the error rates on forties is the lowest because the writers' writing style is stable and they are serious on this work. On the other hand, the young writers of twenties write the character more crabbedly, therefore, the error rate is the highest. Fig.5 show the samples from two different ages, Fig.5(a) are the samples from writers who are forties, Fig.5(b) are from writers who are twenties, as we can see, people's writing style of forties is more formal

Table 5. Error rate based on education (%)

Above undergraduate	Technical College	Specialized Middle School	High School	Below Middle School
6.9	4.6	5.1	5.9	6.1

Table 4. Proportion of Writers' Occupations (%)

Workers	Farmers	Students	Governors	Scientists	Sales	Teachers	Soldiers	Doctors	Others
17.1	2.7	25.2	11.6	4.6	4.5	5.5	13.2	2.1	13.5

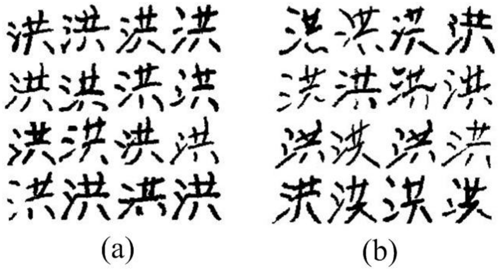
(a) (b)

Figure 4. Samples of a Chinese character written by subjects with different educations, (a) samples written by subjects from technical colleges, (b) samples written by subjects from universities.

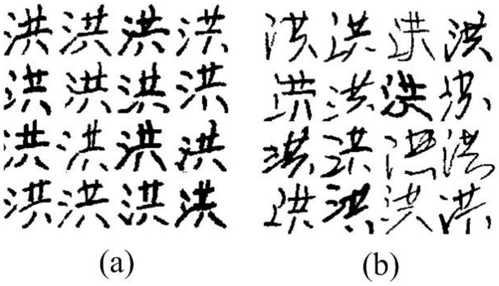
(a) (b)

Figure 5. Samples from subjects with different ages, (a) samples from subjects who are forties, (b) samples from subjects who are twenties

Table 6. Error rate based on ages(%)

Below Twenty	Twenties	Thirties	Forties	Fifties	Above Sixty
6.4	7.2	4.3	4.1	4.5	4.6

than others.

In China, one of the most difficult handwriting style is the doctor's script, because they write the prescription very crabbedly, even human being is hard to recognize them, the experiment results verify this point. Among all the occupations, the doctors' error rate is the highest. Some results are shown in Table 7, the error rate of scientists and teachers are low because their writing style are more formal.

4. Evaluation on HCL2000

To evaluate the HCL2000, we consider the similar Chinese characters reside on a nonlinear manifold structure and propose a cascade framework that combines global similarity with local discriminative cues to classify 3755 handwritten Chinese characters classes [16] [2] [13]. First, we find the similarity of different words using a nearest-neighbor (NN) classifier, and then followed by Linear Discriminant Analysis (LDA), Locality Preserving Projection (LPP) [5] and Marginal Fisher Analysis (MFA) [14],The recognition process has two steps. First, an input image is preprocessed and classified by the NN classifier to assign into a character group [8] [7]. Then we use the gradient feature for further discrimination. Gradient feature vectors are subsequently transformed to the LPP, LDA, and MFA representations with class label information respectively. For fair comparisons, both LPP sets k = 20 as the parameter for constructing the graph and use the simple-mind weight scheme for all the following experiments. For MFA, its parameter k_1 = 3 and k_2 = 20, the maximum number of similar characters in a group is 50. The result on HCL2000 is shown in Fig.6.

Table 7. Error rate based on occupations(%)

Workers	Students	Scientists	Teachers	Soldiers	Doctors
6.8	7.1	4.7	4.6	5.7	24.7

Based on the experimental results, MFA, LPP are better than LDA. LDA yields some meaningful projections since handwritten characters of the same class are mapped close to each other. However, LDA discovers only the Euclidean structure, and can not reveal the underlying nonlinear manifolds that handwritten characters lie on. Therefore its discriminating power is limited. LPP and MFA are supervised methods and preserve local neighborhood information. Their discriminating powers are better than LDA, since similar handwritten characters contain variations and

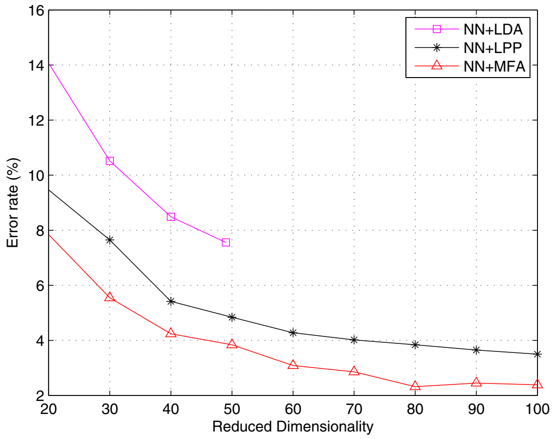

Figure 6. Experimental results on HCL2000 database.

significant overlapping among different classes.

5. Conclusions

We have presented a large scale handwritten Chinese character database, HCL2000, in this paper. To our knowledge, HCL2000 is the largest in number with the writers' information to date. It contains 3,755 frequently used simplified Chinese characters written by 1,000 different writers. The writers' information is incorporated in the database to facilitate testing on grouping writers with different background such as age, occupation, gender and education etc. We have studied some characteristics of writing styles from different groups of writers. We have evaluated HCL2000 database using three different algorithms as a baseline. We have decided to publish the database along with this paper and make it free for a research purpose.

6 Acknowledgments

This work was partially sponsored by the grants from NSFC(Natural Science Foundation of China) under Grant No.60675001, 111 Project(B08004) and 863 High-tech project under Grant No. 2007AA01Z417.

References

[1] S.-L. Chou and S.-S. Yu. Sorting qualities of handwritten chinese characters for setting up a research database. *Proceeding of International Conference on Document Analysis and Recognition (ICDAR'93)*, 1993.

[2] W. deng, J. Hu, J. Guo, and H.-G. Zhang. Comments on 'globally maximizing, locally minimizing: Unsupervised discriminant projection with application to face and palm biometrics'. *IEEE trans. Pattern Analysis and Machine Intelligence*, 30(8), 2008.

[3] Q. Fu, X.-Q. Ding, T. Li, and C. Liu. An effective and practical classifier fusion strategy for improving handwritten character recognition. *Proceeding of International Conference on Document Analysis and Recognition (ICDAR'07)*, 2007.

[4] J. Guo, N. Sun, Y. Kimura, H. Echigo, and R. Sato. Recognition of handwritten characters using pattern transformation method with cosine function. *IEICE Trans*, 76:853–842, 1993.

[5] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacian faces. *IEEE trans. Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[6] J.-J. Hull. A database for handwritten text recognition research. *IEEE trans. Pattern Analysis and Machine Intelligence*, 16(5), 1994.

[7] N. Kato, M. Suzuki, S. Omachi, H. Aso, and Y. Nemoto. A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance. *IEEE trans. Pattern Analysis and Machine Intelligence*, 21(3):258–262, 1999.

[8] C.-L. Liu. Normalization-cooperated gradient feature extraction for handwritten character recognition. *IEEE trans. Pattern Analysis and Machine Intelligence*, 29(8):1465–1469, 2007.

[9] H.-L. Liu and X.-Q. Ding. Improve handwritten character recognition performance by heteroscedastic linear discriminant analysis. *Proceeding of International Conerence on Pattern Recognition (ICPR'06)*, 2006.

[10] X. Liu, Y.-D. Jia, and M. Tan. Geometrical-statistical modeling of character structures for natural stroke extraction and matching. *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR06)*, 2006.

[11] T. Long and L.-W. Jin. Building compact mqdf classifier for large character set recognition by subspace distribution sharing. *Pattern Recognition*, 41(9):2916–2925, 2008.

[12] S. Srihari, X. Yang, and G. Ball. Offline chinese handwriting recognition: an assessment of current technology. *Frontiers of Computer Science in China*, (2), 2007.

[13] H. Suzuki, Y. Waizumi, N. Kato, and Y. Nemoto. Discrimination of similar characters with a nonlinear compound discrminant funtion. *Systems and Computers in Japan*, 38(11):704–715, 2007.

[14] S.-C. Yan, D. Xu, and B. Zhang. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE trans. Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[15] D. Yang and L.-W. Jin. Handwritten chinese character recognition using modified LDA and kernel FDA. *Proceeding of International Conference on Document Analysis and Recognition (ICDAR'07)*, 2007.

[16] H.-G. Zhang, W. Deng, J. Guo, and J. Yang. Handwritten chinese character recognition using local discriminant projection with prior information. *Proceeding of International Conerence on Pattern Recognition (ICPR'08)*, 2008.