2011 International Conference on Document Analysis and Recognition

An Empirical Evaluation on HIT-OR3C Database

Shusen Zhou, Qingcai Chen, Xiaolong Wang, Xinyi Guo, Hui Li
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, P.R. China
{*zhoushusen,qingcai.chen*}*@hitsz.edu.cn, wangxl@insun.hit.edu.cn, guoxy_mail@163.com, lihui0228@126.com*

***Abstract*—Recently, we have proposed a handwriting Chinese character database HIT-OR3C. Though it has been introduced in detail, to date, it has not been evaluated by any handwriting recognition method. To help the researchers use this database for algorithm evaluation, we propose the structure of HIT-OR3C database. Moreover, we evaluate the OR3C database with a series of experiments using state-of-the-art handwriting recognizer. These experiment results on the different subsets can be a benchmark for the researchers who will use the database. The low average recognition rate confirms that the HIT-OR3C database is challenging.**

***Keywords*-HIT-OR3C; online and offline handwriting recognition; Chinese character; Chinese document;**

I. Introduction

Machine recognition of handwriting has practical significance [1], as in reading handwriting notes in a smart phone or PDA, in handwriting document retrieval, in postal addresses on envelopes, in bank checks, in handwriting fields in forms, etc [2]. Online handwriting character recognition, based on the trajectories of pen tip movements, has attracted a renewed research interest for the booming of touch screen mobile devices. Up to now, numerous of handwriting input devices and interfaces have been invented to improve the precision of trajectory capturing and the comfort of writing [3]. Offline handwriting character recognition, based on the scanned images, is less accurate than online recognition [1]. However, it is an important method for some specific domains, like interpreting handwriting postal address on envelopes, reading amounts on bank checks, etc [1]. Though the high reported recognition precision on standard corpus, both online and offline recognition of Chinese handwriting characters are still big challenging problems for most of real applications. The efficiency and accuracy of existing handwriting character input software, are still far from satisfied to replace character input systems based on keyboard. Progress in document analysis has long been driven by sound experiments on carefully prepared test data [4]. The unconstrained character recognition remains one of the most challenging tasks [5]. One of the most critical bottlenecks for improving its recognition performance is the short of available large-scale unconstrained handwriting dataset [6].

The development trends of handwriting corpus include [7] that the scale of sampling grows from single characters to paragraphs, and the manner of handwriting styles are changed from regular to cursive and unconstrained. The HIT-OR3C (Harbin Institute of Technology Opening Recognition Corpus for Chinese Characters) [8], a Chinese handwriting character and document corpus that are inputted through handwriting pad, can be downloaded through IAPR-TC11 Website (http://www.iapr-tc11.org/mediawiki/index.php/Harbin_Institute_of_Technology_Opening_Recognition_Corpus_for_Chinese_Characters_(HIT-OR3C)) or Intelligence Computing Research Center Handwriting Group Website (http://www.haitianyuan.com/hw/hw_en.php). HIT-OR3C has several attractive characteristics. Firstly, it consists of 5 subsets that can be applied to evaluate some special algorithms designed for digit recognition, letter recognition etc. Secondly, it is the first published Chinese handwriting database that includes both characters and documents at the same time, and thus supports the character based training by previous four subsets and document level testing by the document subset. Thirdly, it is the first published online Chinese handwriting database written on the USB port based handwriting pad. Most of the computer users input the Chinese characters using handwriting pad, so the systems developed for handwriting input method can be evaluated by this database. Moreover, it also has pseudo-offline data based on HIT-OR3C online database by using digital ink techniques, so it can also be applied for offline handwriting recognition research.

In this paper, we describe the structure of HIT-OR3C database, and report the experiment results of different subsets using state-of-the-art recognizer. The rest of the article is as follows: Section II proposes the structure of HIT-OR3C database. Section III reports the experiment results on different subsets using state-of-the-art recognizer. The paper is closed with conclusion.

II. HIT-OR3C Structure

HIT-OR3C consists of 5 subsets listed in Table I. GB1 and GB2 are abbreviations of the Level 1 and Level 2 character sets of Chinese GB2312-80 standard respectively. Digit, Letter, GB1, and GB2 subsets are written by 122 persons, totally 832,650 character samples. Document subset includes 10 documents collected from the news of Sina (http://www.sina.com.cn). They had been written by 20 persons and each document was recorded 2 times. The document collection, consists totally 77,168 single characters that cover 2,442 characters, among which include 10 digits, 49 letters, 2,286 GB1 characters, and 97 GB2 characters.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.232

Table I
SUBSETS IN HIT-OR3C.

Subsets	Detail
Digit	10 numeric digits
Letter	52 English upper and lower case alphabets
GB1	3,755 characters in GB2312-80 Set1
GB2	3,008 characters in GB2312-80 Set2
Document	10 documents sampled from news reports

Table II
THE FORMAT OF HEADER INFORMATION IN THE VECTOR FILE.

Item	Length	Comment
N	4 B	Number of characters in vector file
$C_1, C_2, ..., C_N$	$2N$ B	Size of storage space for N characters

The samples of characters from 5 subsets can be seen in Fig. 1. These five lines are corresponding to Digit, Letter, GB1, GB2, and Document subsets separately.

In the previous 4 character subsets, characters written by one person are stored in two files, one for online version (vector file) and the other for offline version (image file). The different subsets are defined as index ranges within these files [Digit (1-10), Letter (11-62), GB1 (63-3817), GB2 (3818-6825)]. The document subset have been post-processed and split into individual characters. The characters for an article written by one person are stored sequentially in an image file and a vector file correspondingly. The format of the vector file is similar as the first four subsets. All the character images in OR3C are composed of 128×128 pixels, both online and offline version.

There are three types of files, i.e., vector, image and label file. Vector file stores the online information of handwriting characters. Image file stores the offline information

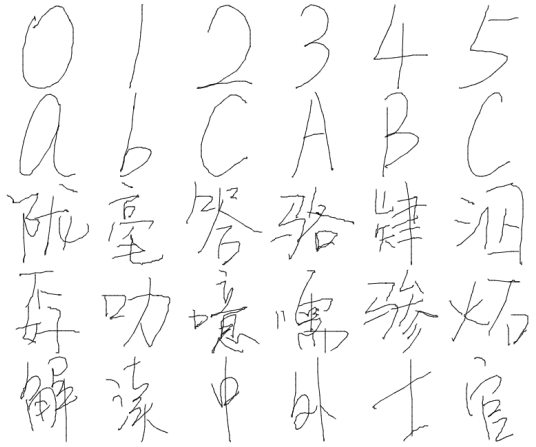

Figure 1. Samples of OR3C Dataset.

Table III
THE FORMAT OF HEADER INFORMATION FOR ONE CHARACTER IN THE VECTOR FILE.

Item	Length	Comment
M	1 B	Number of strokes in the character
$S_1, S_2, ..., S_M$	M B	Number of sampling point for M strokes

Table IV
THE FORMAT OF HEADER INFORMATION IN THE IMAGE FILE.

Item	Length	Comment
N	4 B	Number of characters in image file
H	1 B	The height of one character
W	1 B	The width of one character

of handwriting characters, which is generated from online information of characters by digital ink technology. Label file stores the label of handwriting characters corresponding to vector and image file. These files are defined in Table II, Table III, Table IV, and Table V. Table II defines the format of header information in the vector file. N is the number of characters, which use 4 Bytes in the vector file. The following $2N$ Bytes stores the size of storage space which used by these N characters. At last, the vector file use $\sum_{k=1}^{N} C_k$ Bytes to store the online information of N characters. Table III defines the format of header information for one character in the vector file. The previous 1 Byte stores the number of strokes for this character. The following M Bytes stores the number of sampling point for M strokes. The last $2\sum_{k=1}^{M} S_k$ Bytes stores the online information of M strokes. For the k^{th} stroke, we use S_k points to represent it. The x and y coordinates of each point were stored in the vector file, and 2 Bytes were used for each point. Table IV defines the format of header information in the image file. N is the number of characters, which is the same as the vector file. H and W are the height and width of one character, which means that there are $H \times W$ points for each character. The last $N \times H \times W$ Bytes are used to store the offline information of handwriting characters. Table V defines the format of label file. N is the number of labels, which is corresponding to the number of the characters in the vector and image files. L represents the size of storage space for each label, in OR3C datasets, $L = 2$. The last $N \times L$ Bytes stores the label of handwriting characters in the corresponding vector and image files.

We provide the toolkit of reading the OR3C dataset properly, which includes Matlab, C++ and JAVA source code. All the toolkits can be downloaded from IAPR-TC11 Website or Handwriting Group Website.

III. EXPERIMENTS

To evaluate the OR3C dataset, we conduct some experiments on the online and offline version of handwriting characters. For the previous 4 character datasets, there are

Table V
THE FORMAT OF LABEL FILE.

Item	Length	Comment
N	2 B	Number of labels in label file
L	1 B	Size of storage space for every label
$T_1, T_2, ..., T_N$	$N \times L$ B	N labels corresponding to N characters

Table VI
TEST ACCURACIES OF DIFFERENT CHARACTER SUBSETS IN ONLINE DATASET.

Dataset	1 candidate	5 candidates	10 candidates
Digit	95.45%	100%	100%
Letter	80.86%	99.13%	99.48%
GB1	87.48%	94.23%	94.93%
GB2	93.18%	97.58%	97.90%
Previous 3 subsets	86.64%	93.85%	94.65%

122 vector files and 122 image files to store the online and offline information of handwriting characters.The index number of these files are given from 1 to 122. In each file, the index ranges of these subsets are Digit (1-10), Letter (11-62), GB1 (63-3817), and GB2 (3818-6825). For the following character recognition experiments, the previous 100 files are used for training and the rest 22 files for testing. For the document dataset, every document is written by 2 persons, there are 20 vector files and 20 image files to store the online and offline information of handwriting characters. To imitate the real applications, we use previous character dataset to train the classifier, and use document dataset to test the classifier. In future, as more handwriting documents are collected, part of documents can be added to training data too.

A. Experiment with Online Dataset

For online version of OR3C dataset, we use a state-of-the-art recognizer [9] [10], the experimental setting is similar with [6]. First, we reduce the dimension of images to 64×64, normalize all the images with pseudo 2D moment normalization method. Second, extract the feature with direction feature extraction method, then reduce the feature dimensionality from 512 to 160 by Fisher linear discriminant analysis (LDA). Third, coarse classify the characters with K-mean method, then use modified quadratic discriminant function (MQDF) classifier for fine classification.

Table VI shows the test accuracies of different character subsets with 100 training files and 22 test files. In Table VI, accuracy for k candidates means that if the right character contained in the top k candidates returned by the classifier, then this recognition is counted as right. This is imitating the effect of Chinese Character Input Method, in which the user can choose the right result from the candidate. Through the table, we can find that the test accuracies of GB2 subset are better than GB1 subset. Because there are 3,755 categories for GB1 subset and 3,008 categories for GB2 subset, the number of category for GB1 subset is a little more than GB2 subset, so it is hard to distinguish more classes in GB1 subset. More importantly, there are more complex characters (i.e., characters with more strokes) in GB2, which carry more information and thus reach better recognition accuracy for GB2 subset.

Table VII
TEST ACCURACIES OF DIFFERENT DOCUMENT FILES IN ONLINE DATASET TRAINING WITH DIFFERENT CHARACTER SUBSETS.

Document	File	Previous 3 subsets	All subsets
1	1	68.94%	61.30%
1	2	91.02%	83.21%
2	1	91.37%	83.14%
2	2	80.64%	70.44%
3	1	87.31%	80.58%
3	2	86.80%	79.76%
4	1	74.45%	68.60%
4	2	51.00%	44.36%
5	1	89.43%	81.81%
5	2	73.27%	65.24%
6	1	84.98%	77.04%
6	2	88.22%	81.14%
7	1	68.77%	62.22%
7	2	65.96%	58.36%
8	1	89.98%	84.00%
8	2	82.07%	77.55%
9	1	86.49%	78.59%
9	2	88.86%	81.72%
10	1	87.38%	80.27%
10	2	92.27%	85.95%
mean		81.46%	74.26%

Beacause most of the characters in document subset are Digit, Letter, and GB1 characters, so the performance of classifier for these previous 3 subsets is very important. The last row of Table VI shows the average test accuracy of previous 3 subsets for 3,817 categories. It is shown that even with 10 candidates, there are still more than 5% characters can not be rightly recognized by the classifier. Table VI also shows that the test accuracies with 10 candidates is just a little better than the results with 5 candidates. It means that to improve the accuracy of the classifier by returning more than 5 candidates is useless.

It is possible to achieve a high recognition rate, above 98% on regular scripts. However, on fluent or fluent-regular scripts, it is difficult to achieve a recognition rate above 90% [3]. In our experiments, the average recognition rate is just 86.64% for previous 3 subsets with 3,817 categories, it is much lower than the ones reported on other popular online databases. For example, 98.24% on Japanese Kanji [9]. This confirms that the HIT-OR3C database is challenging.

The test accuracies of different document handwriting files training with different subsets are shown in Table VII. All the reported results are the test accuracies with 1 candidate. The first column of Table VII is the document number. There are 10 documents in document subset. For

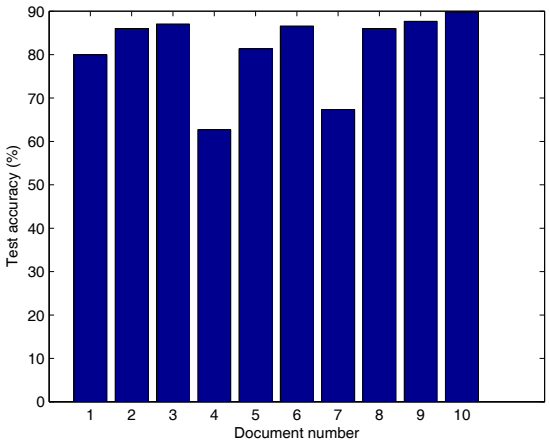

Figure 2. Test Accuracies of Documents with Training on Previous 3 Character Subsets.

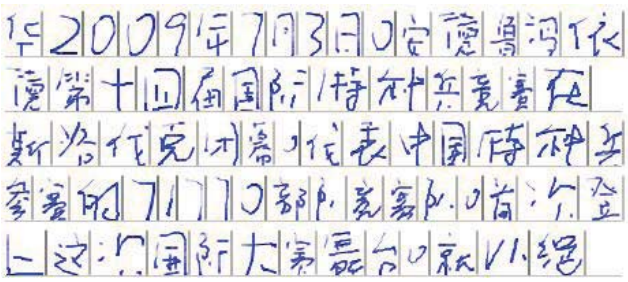

Figure 3. Samples of the Second File of Document 4.

each document, there are 2 files that are written by two different persons respectively. The second column is the file number of different documents. The rest 2 columns are the test accuracies with different training data. For the third column, the training data are the previous 3 character subsets, i.e., Digit, Letter, and GB1. For the fourth column, the training data are all the character subsets. Through the results of these two columns, we can find out that the performance of the classifier which is trained by just previous 3 subsets is better than the classifier which is trained by all the character subsets. Because there are only 97 chinese characters in GB2 subset appear in the document subset, comparing with 3,008 categories for GB2 subset, most of the categories are not be used. However, if we train the classifier with all character subsets for 6,825 categories, much more time should be used. Moreover, it is hard to distinguish so many classes. So we use previous 3 subsets to train the classifier firstly. When test the document files, all chinese characters that come from GB2 subset are counted as wrong classified characters.

Through Table VII, we can see that the test accuracies various from different files. For the test accuracies of training on previous 3 subsets, the worst result is 51% for the second file of document 4; the best result is 92.27% for the second file of document 10. It is caused by the various of handwriting quality of different files, and the various of recognition difficulty for different documents. The average test accuracies of different documents with the classifier training with previous 3 subsets are shown in Fig. 2. Through the figure, we can see that since the handwriting quality of the second file of document 4 is so worse, it is the hardest document to be recognized. The samples of this file are shown in Fig. 3.

Table VIII
TEST ACCURACIES OF DIFFERENT CHARACTER SUBSETS IN OFFLINE DATASET.

Dataset	1 candidate	5 candidates	10 candidates
Digit	91.81%	99.55%	100%
Letter	81.03%	99.91%	99.91%
GB1	77.71%	89.92%	91.55%
GB2	85.67%	94.91%	95.84%
Previous 3 subsets	77.53%	89.75%	91.33%

B. Experiment with Offline Dataset

For offline version of OR3C dataset, we just modify the online version recognizer slightly. First, we reduce the dimension of images to 64×64, normalize all the images with modified centroid-boundary alignment (MCBA) method [11]. Second, gradient direction feature is extracted, then the feature dimensionality is reduced from 512 to 160 by Fisher linear discriminant analysis (LDA). Third, the characters are coarse classified with K-mean method, and the modified quadratic discriminant function (MQDF) classifier is used for fine classification.

The test accuracies of different character subsets with 100 training files and 22 test files can be seen in Table VIII. Comparing with Table VI, we can see that the performance of the recognizer in online dataset is better than offline dataset. It is because that the stroke information for online characters is more accurate than offline characters.

As shown in Table VIII, the offline recognition rate is just 77.53% for previous 3 subsets with 3,817 categories. It is much lower than the ones reported on other popular offline databases. For example, 97.80% on HCL2000 [12]. Although the offline database is produced from online database, for the published version, we just connect the sample point of online version characters with unique stroke width. So the digital ink method should not affect the recognition rate of offline database, and the lower recognition rate confirms that our offline version of OR3C database is challenging.

The test accuracies of different document handwriting files training with different subsets are shown in Table IX. All the reported results are the test accuracies with 1 candidate. Similar with Table VII, the first column is the document number, the second column is the file number

Table IX
TEST ACCURACIES OF DIFFERENT DOCUMENT FILES IN OFFLINE DATASET TRAINING WITH DIFFERENT CHARACTER SUBSETS.

Document	File	Previous 3 subsets	All subsets
1	1	54.51%	50.13%
1	2	85.89%	82.79%
2	1	84.27%	82.05%
2	2	64.32%	60.20%
3	1	77.90%	75.55%
3	2	72.72%	68.75%
4	1	59.55%	56.70%
4	2	44.69%	41.29%
5	1	83.26%	80.52%
5	2	62.37%	57.60%
6	1	74.43%	72.12%
6	2	79.84%	77.16%
7	1	59.07%	53.26%
7	2	56.04%	51.86%
8	1	74.61%	72.14%
8	2	74.26%	68.56%
9	1	77.53%	75.31%
9	2	80.75%	78.66%
10	1	78.58%	74.94%
10	2	87.02%	85.15%
mean		71.58%	68.24%

of different documents, and the rest 2 columns are the test accuracies with different training data. For offline database, the performance of the classifier which trained by just previous 3 subsets is also better than the classifier trained by all the character subsets.

IV. CONCLUSION

The opening recognition corpus for Chinese characters (HIT-OR3C) is the first Chinese handwriting database that includes the single character dataset and document dataset at the same time. The document dataset can be used to research on the context dependent character recognition and to test recognizers under a more real situation. The single character dataset can overcome the data sparseness of seldom-occurred characters in the document dataset. Moreover, it is the first Chinese handwriting database which has both the online and offline version of the datasets, and can be used on both online and offline Chinese handwriting recognition research.

To help the researchers use HIT-OR3C more conveniently, we propose the structure of the database, describe the format of files which save the online and offline information of handwriting charactors. Moreover, we do several experiments on different subsets of the database with state-of-the-art handwriting recognizer, which can be a benchmark for the researchers who will use the database.

ACKNOWLEDGMENT

We would like to thank Xinggang Fu and all those volunteers for their contributions in HIT-OR3C. This work was supported in part by the National Natural Science Foundation of China (No. 60703015 and No. 60973076).

REFERENCES

[1] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[2] A. Almaksour and E. Anquetil, "Fast incremental learning strategy driven by confusion reject for online handwriting recognition," in *International Conference on Document Analysis and Recognition*, 2009, pp. 81–85.

[3] C. L. Liu, S. Jaeger, and M. Nakagawa, "Online recognition of chinese characters: The state-of-the-art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 198–213, 2004.

[4] G. Nagy, "Document systems analysis: Testing, testing, testing," Tech. Rep., 2010.

[5] Z. Huang, K. Ding, L. Jin, and X. Gao, "Writer adaptive online handwriting recognition using incremental linear discriminant analysis," in *International Conference on Document Analysis and Recognition*, 2009, pp. 91–95.

[6] D. H. Wang, C. L. Liu, J. Yu, and X. D. Zhou, "Casia-olhwdb1: A database of online handwritten chinese characters," in *International Conference on Document Analysis and Recognition*, 2009, pp. 1206–1210.

[7] Y. Y. Li, L. W. Jin, X. H. Zhu, and T. Long, "Scut-couch2008: A comprehensive online unconstrained chinese handwriting dataset," in *International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 165–170.

[8] S. Zhou, Q. Chen, and X. Wang, "Hit-or3c: An opening recognition corpus for chinese characters," in *International Workshop on Document Analysis Systems*, 2010, pp. 223–230.

[9] C. L. Liu and X. D. Zhou, "Online japanese character recognition using trajectory-based normalization and direction feature extraction," in *International Workshop on Frontiers in Handwriting Recognition*, 2006, pp. 217–222.

[10] C. L. Liu, "Handwritten chinese character recognition: effects of shape normalization and feature extraction," in *Arabic and Chinese handwriting recognition*, 2006, pp. 104–128.

[11] C. L. Liu and K. Marukawa, "Global shape normalization for handwritten chinese character recognition: A new method," in *International Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 300–305.

[12] H. Zhang, J. Guo, G. Chen, and C. Li, "Hcl2000- a large-scale handwritten chinese character database for handwritten character recognition," in *International Conference on Document Analysis and Recognition*, 2009, pp. 286–290.