2012 10th IAPR International Workshop on Document Analysis Systems

An Empirical Evaluation on Online Chinese Handwriting Databases

Shusen Zhou, Qingcai Chen, Xiaolong Wang, Zou Chen and Suqin Ao
Shenzhen Graduate School, Harbin Institute of Technology, P.R. China
{zhoushusen,qingcai.chen}@hitsz.edu.cn, wangxl@insun.hit.edu.cn, chenzoudgh@163.com, aosuqin@163.com

***Abstract*—Several online Chinese handwriting databases have been proposed recently. Though they have been introduced in detail, to date, no one has ever evaluated these databases with experimental comparison. To help the researchers use the corresponding database properly for algorithm evaluation and real application, we compare the property of the handwriting characters in these databases, and evaluate them with the same experimental setup and handwriting recognizer. Moreover, we analyze the connection between the property and the corresponding recognition accuracy for the handwriting characters in different databases. These empirical evaluation results can help the researchers choose the right database for different algorithms and applications.**

***Keywords*-SCUT-COUCH2009; CASIA-OLHWDB1; HIT-OR3C; Online handwriting recognition; Chinese handwriting database;**

I. Introduction

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. Though the high reported recognition precision on standard corpus, the online recognition of Chinese handwriting character are still big challenging problems for most of real applications. The unconstrained character recognition remains one of the most challenging tasks [1]. One of the most critical bottlenecks for improving its recognition performance is the short of available large-scale unconstrained handwriting database [2].

Recently, several online Chinese handwriting databases have been published, which include SCUT-COUCH2009 [3], CASIA-OLHWDB1 [2], and HIT-OR3C [4]. In short, we will call these three databases SCUT, CASIA and HIT respectively in the below. SCUT [3] consists of 11 datasets of different vocabularies, named GB1, GB2, TradGB1, Big5, Pinyin, Letters, Digit, Symbol, Word8888, Word17366, Word44208, the total number of character samples is over 3.6 million. The samples were collected using personal digit assistant (PDA) and smart phones with touch screens and were contributed by more than 190 persons. CASIA [2] is another recent published corpus that contains handwriting characters of 4,037 categories produced by 420 persons via the device called "Anoto pen", and 1,694,741 samples in total. HIT [4] is a Chinese handwriting character and document corpus that are inputted through handwriting pad. HIT consists of 5 subsets, namely Digit, Letter, GB1, GB2, and Document. The first 4 subsets contain 6,825 categories, and totally 832,650 samples produced by 122 persons. The document corpus is corresponding to 10 news articles that contain 2,442 categories and 77,168 samples in total. Each news article is produced by 2 persons.

Table I
Number of Writers for All Subsets in Different Chinese Databases.

Subsets	SCUT	CASIA	HIT
Digit	195	420	122
Letter	195	420	122
Pinyin	130	0	0
Symbol	195	420	0
GB1	188	420	122
GB2	195	0	122
TradGB1	130	0	0
Big5	65	0	0
Word8888	130	0	0
Word44208	5	0	0
Word17366	10	0	0
Document	0	0	20

SCUT, CASIA and HIT are collected through PDA, Anoto pen and handwriting pad respectively, and the pre-processing tools are different too. To help the researchers choosing the proper database for different recognition algorithms and applications, we compare the property of the handwriting characters in these databases, and evaluate these databases with the same experimental setup and handwriting recognizer. Moreover, we analyze the connection between the property and the corresponding recognition accuracy for the handwriting characters in different databases.

The rest of the article is as follows. Section II compare the property of the handwriting characters in these three handwriting databases. Section III test the performance of the state-of-the-art handwriting recognizer in these three databases with same experimental setup, and analyze the connection between the property and the corresponding recognition accuracy for the handwriting characters in different databases. The paper is closed with conclusion.

II. Property Comparison

In this part, we analyze the variations of these three databases in the number of strokes, size variation, and aspect ratio variation.

The comparison of subsets composition and number of writers for different databases can be seen in Table I, the

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.19

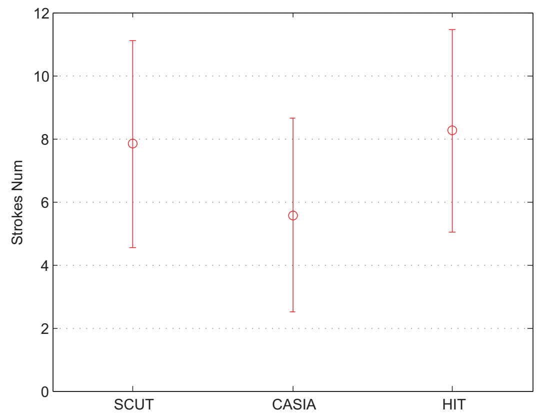

Figure 1. Average and variance of stroke numbers for GB1 subsets in different databases.

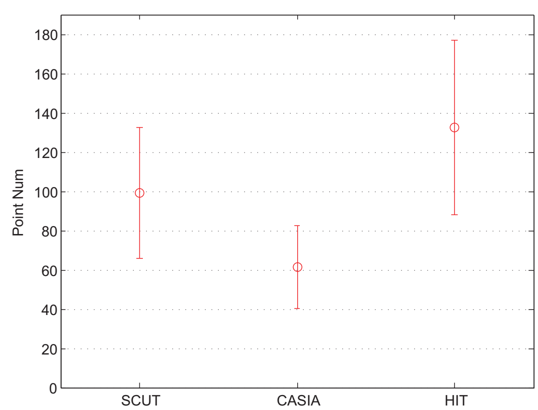

Figure 2. Average and variance of point numbers for GB1 subsets in different databases.

number 0 indicate that this database is not include the corresponding subset. Through the table, we can see that SCUT is the most complete database including most of subsets, CASIA contains the most largest subset written by 420 writers, and HIT is the only database which include document subset.

There are only 3 subsets (Digit, Letter and GB1) which are all included by these three databases. Because Digit and Letter subsets are not include Chinese characters, so we choose GB1 subset to evaluate the property of these three databases. For GB1 subset in CASIA database, some handwriting characters are not collected for some writers because of the poor handwriting quality, so we use the corresponding characters of other writer's to fill in this blank space. Moreover, CASIA database just include 3740 frequently used GB1 characters. So we use 3740×420 handwriting characters for CASIA database, 3755×188 handwriting characters for SCUT database, and 3755×122 handwriting characters for HIT database.

The average and variance of stroke numbers for GB1 subsets in these three databases can be seen in Fig. 1.

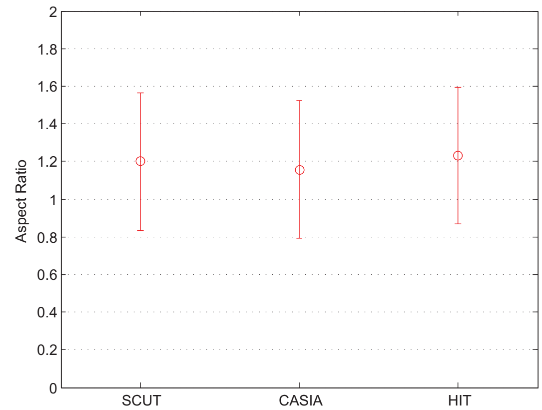

Figure 3. Average and variance of aspect ratios for GB1 subsets in different databases.

The average number of strokes for HIT is more than other two databases, and CASIA contains least number of strokes comparing with other two databases. The variance of stroke numbers for these three databases are almost the same. This means that the quality of handwriting characters in HIT is the best, there are more continuous strokes for handwriting characters than SCUT and CASIA.

The average and variance of point numbers for GB1 subsets in these three databases can be seen in Fig. 2. Similar with the number of strokes, the average number of points for HIT is more than other two databases, and CASIA contains least number of points comparing with other two databases. However, the variance of point numbers for these three databases are different. Comparing with SCUT and CASIA, the variance of point numbers for HIT is the biggest one. This means that the size and size variation for HIT are both bigger than other two databases, more sampling points are used to represent these handwriting characters in HIT, and the writing speed variation for HIT is bigger than other two databases.

The average and variance of aspect ratios for GB1 subsets in these three databases can be seen in Fig. 3. Both the average and variance of aspect ratios for these three databases are almost same, the aspect ratio for HIT is slightly bigger than other two databases. This means that most handwriting characters in HIT database are longer and thinner than the characters in other databases.

Through Fig. 1, Fig. 2 and Fig. 3, we can find out that HIT has the highest aspect ratio, and is the biggest database which use more sampling points to save the handwriting information. CASIA is the most cursive database when collecting the handwriting characters, because of the more continuous strokes in it. All these three targets for SCUT are in the middle.

Table II
TEST ACCURACIES (%) OF DIFFERENT DATABASES WITH THEIR CORRESPONDING CLASSIFIER FOR ONLINE RECOGNITION.

Dataset	1 candidate	5 candidates	10 candidates
SCUT	78.56	89.57	91.24
CASIA	83.06	92.31	93.33
HIT	**87.48**	**94.23**	**94.93**

III. EXPERIMENTAL COMPARISON

In this part, we compare the performance of a handwriting recognizer on the same subset of SCUT, CASIA and HIT databases. Handwriting characters of 122 writer's in GB1 subsets are used for every database, characters of 100 writers are used as training data, and the rest characters are used as test data. First, we compare the performance of recognizer on the online dataset of these three databases with same experimental setup. Second, we compare the performance of recognizer on the pseudo-offline dataset of these three databases with same experimental setup.

A. Experiment with Online Dataset

We use a classical recognizer for online handwriting Chinese character recognition, the experimental setting is similar with [2]. 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 [5], 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.

The test accuracies of GB1 subsets in different databases with classifiers trained by their corresponding database for online recognition can be seen in Table II, 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 this subset in HIT are better than in other databases, and the subset in SCUT get the worst recognition results.

As shown in Section II, although HIT is higher in aspect ratios, these different aspect ratios for different character categories can help recognizer to classify these characters accurately. Moreover, it has more sampling points and fewer connecting stokes, so we can get the best performance on subsets in HIT. CASIA has different aspect ratios for different character categories too, however, it has fewer sampling points and more connecting stokes than HIT, so its recognition results are worse than HIT. For SCUT, it has fewer sampling points and more connecting stokes than HIT too, so we get the worst recognition results on the subsets in it.

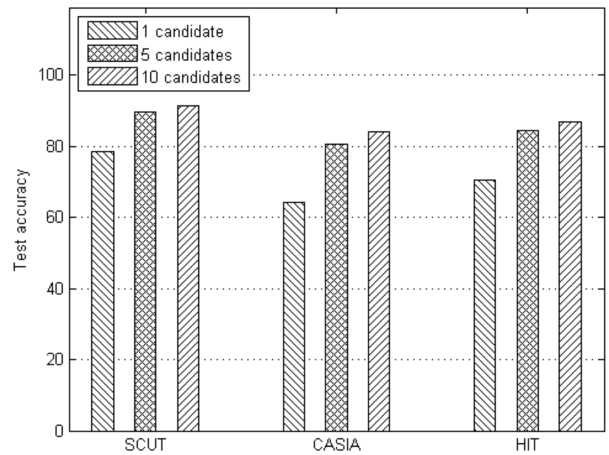

Figure 4. Test accuracies (%) of GB1 subset in SCUT database with classifiers trained by different databases for online handwriting recognition.

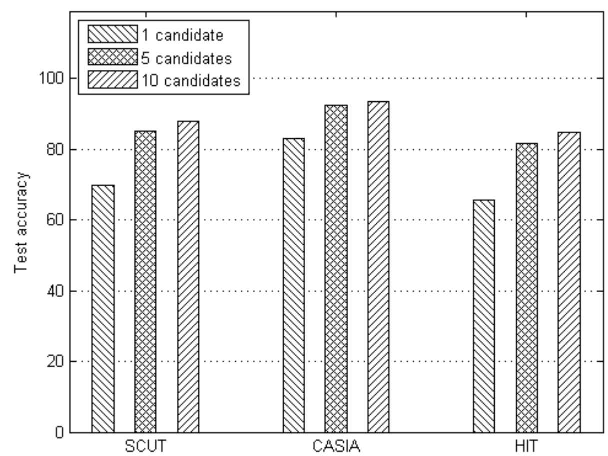

Figure 5. Test accuracies (%) of GB1 subset in CASIA database with classifiers trained by different databases for online handwriting recognition.

To test the writing and sampling difference between SCUT, CASIA and HIT databases, we use the classifier trained by these three databases test every database respectively. The test accuracies of GB1 subset in SCUT, CASIA, and HIT databases with classifiers trained by different databases for online handwriting recognition can be seen in Fig. 4, Fig. 5 and Fig. 6 respectively. In Fig. 4, we can see that classifier trained by SCUT database has the best performance on SCUT database. Through Fig. 5 and Fig. 6, we can also find out that classifier trained by CASIA database has the best performance on CASIA database, and classifier trained by HIT database has the best performance on HIT database. This validate that the handwriting characters from the same database have the best similarity.

In Fig. 4, on test data from SCUT, the performance of

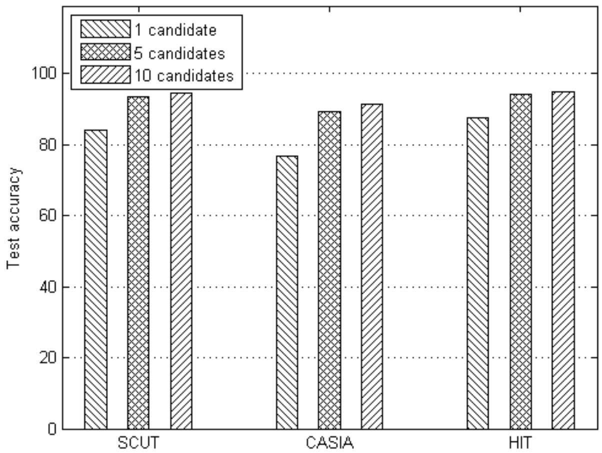

Figure 6. Test accuracies (%) of GB1 subset in HIT database with classifiers trained by different databases for online handwriting recognition.

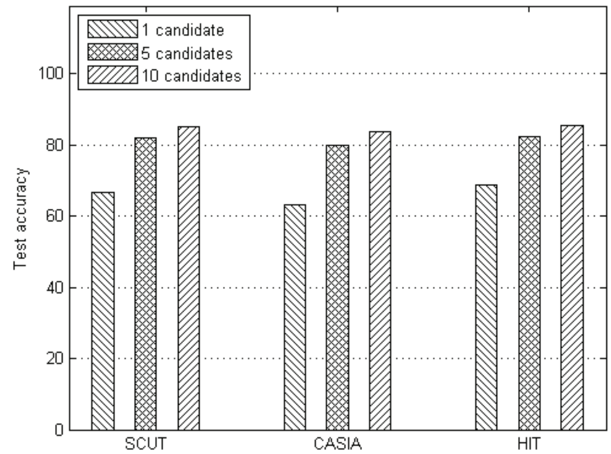

Figure 7. Test accuracies (%) of document subset in HIT database with classifiers trained by different databases for online handwriting recognition.

the classifier trained by SCUT is much better than the classifier trained by CASIA, especially for the test accuracy for 1 candidate. The performance of the classifier trained by HIT is worse than the classifier trained by SCUT, but better than the classifier trained by CASIA. This means that there are much difference between SCUT and CASIA. In Fig. 5, on test data from CASIA, the performance of the classifier trained by CASIA is much better than the classifier trained by SCUT and HIT, the performance of the classifier trained by HIT is slight worse than the classifier trained by SCUT. This means that there are much difference between HIT and CASIA. In Fig. 6, on test data from HIT, the performance of the classifier trained by HIT is much better than the classifier trained by CASIA, and slight better than the classifier trained by SCUT. This means that there are much difference between HIT and CASIA too. Through Fig. 4, Fig. 5 and Fig. 6, we can see that there are much difference between HIT and CASIA, and HIT is similar with SCUT.

We also test the Document subset in HIT database with classifiers trained by different databases for online handwriting recognition, which can be seen in Fig. 7. The contextual processing method is not used for the test. The performance of the classifier trained by HIT is much better than the classifier trained by CASIA, and slight better than the classifier trained by SCUT. This validate that there are much difference between HIT and CASIA, and few difference between HIT and SCUT.

B. Experiment with Offline Dataset

For offline handwriting recognition, 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 [6]. Second, gradient direction feature is extracted with Sobel operator [7], 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.

Table III
TEST ACCURACIES (%) OF DIFFERENT DATABASES WITH THEIR CORRESPONDING CLASSIFIER FOR OFFLINE RECOGNITION.

Dataset	1 candidate	5 candidates	10 candidates
SCUT	66.52	82.53	85.4
CASIA	72.34	87.17	89.33
HIT	**77.71**	**89.97**	**91.55**

The test accuracies of GB1 subsets in different databases with classifiers trained with their corresponding database for offline recognition can be seen in Table III, these results are worse than online recognition results in Table II, because the offline feature extraction method MCBA deform the character patterns. Same as online recognition results, we can see that the test accuracies of GB1 subset in HIT are better than in other databases, and the subset in SCUT gets the worst recognition results.

To test the writing difference on offline dataset of SCUT, CASIA and HIT databases, we use the classifier trained by these three databases test every database respectively. The test accuracies of GB1 subset in SCUT, CASIA, and HIT databases with classifiers trained by different databases for offline handwriting recognition can be seen in Fig. 8, Fig. 9 and Fig. 10 respectively. Same as online handwriting recognition, the classifiers training and test by the same database have the best performance.

IV. CONCLUSION

To help the researchers choose the right database for different applications and algorithm evaluation, we compare the property of the handwriting characters in three handwriting Chinese character databases SCUT, CASIA and HIT,

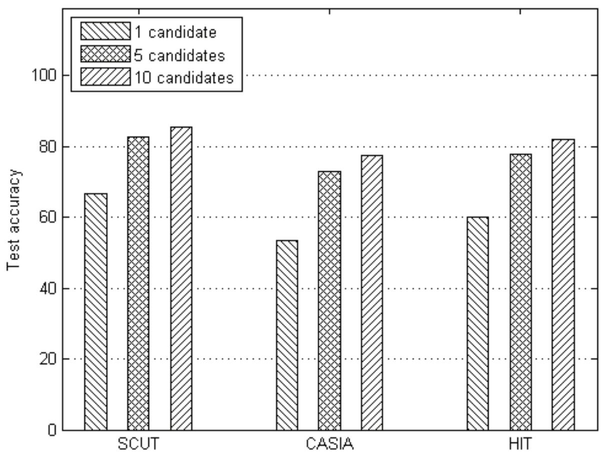

Figure 8. Test accuracies (%) of GB1 subset in SCUT database with classifiers trained by different databases for offline handwriting recognition.

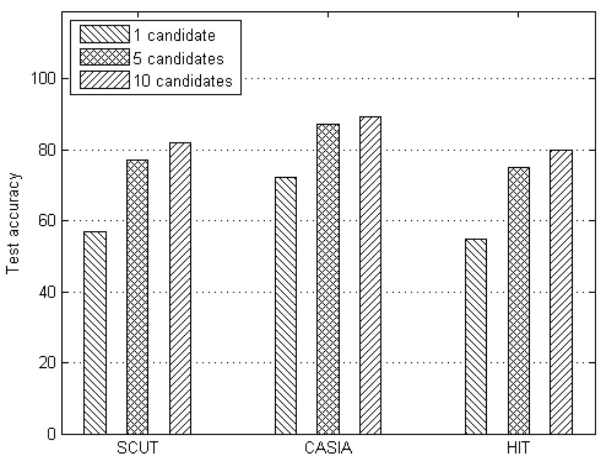

Figure 9. Test accuracies (%) of GB1 subset in CASIA database with classifiers trained by different databases for offline handwriting recognition.

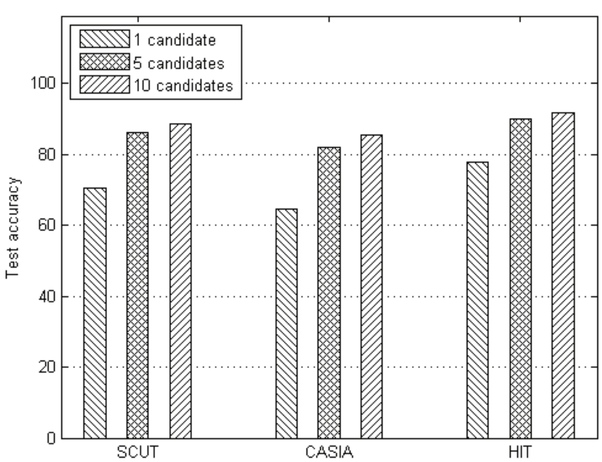

Figure 10. Test accuracies (%) of GB1 subset in HIT database with classifiers trained by different databases for offline handwriting recognition.

and evaluate these databases with the same experimental setup and handwriting recognizer. We analyze the connection between the property and the corresponding recognition accuracy for the handwriting characters in different databases, and find out that there are much difference between handwriting characters in HIT and CASIA, and few difference between handwriting characters in HIT and SCUT. Users can combine the handwriting characters in HIT and SCUT for training and test the new recognition algorithms, and use these combined characters for applications with higher sampling rate. Comparing with HIT and SCUT, CASIA has fewer sampling points, which can be used in the applications with lower sampling rate.

ACKNOWLEDGMENT

We would like to thank Xinggang Fu, Xinyi Guo, Hui Li and all those volunteers for their contributions in HIT-OR3C. We would also like to thank Chenglin Liu et al. in Chinese Academy of Sciences, and Lianwen Jin et al. in South China University of Technology, for their providing of CASIA-OLHWDB1 and SCUT-COUCH2009 databases. This work was supported in part by the National Natural Science Foundation of China (No. 61173075 and No. 60973076).

REFERENCES

[1] 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.

[2] 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.

[3] L. W. Jin, Y. Gao, G. Liu, Y. Y. Li, and K. Ding, "Scut-couch2009-a comprehensive online unconstrained chinese handwriting database and benchmark evaluation," *International Journal on Document Analysis and Recognition*, vol. 14, no. 1, pp. 53–64, 2011.

[4] 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.

[5] 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.

[6] 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.

[7] C. L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques," *Pattern Recognition*, vol. 36, no. 10, pp. 2271–2285, 2003.