Online and Offline Handwritten Chinese Character Recognition: Benchmarking on New Databases

Cheng-Lin Liu, Fei Yin, Da-Han Wang, Qiu-Feng Wang

National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P.R. China
E-mail : {liucl, fyin, dhwang, wangqf}@nlpr..ia.ac.cn

Abstract: By 2010, the National Laboratory of Pattern recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA) has released four databases of online and offline handwritten characters, among its database series of unconstrained Chinese handwriting. This paper introduces the composition of the databases and reports the recognition performance on them using state-of-the-art methods. The online database CASIA-OLHWDB1.0 and offline database CASIA-HWDB1.0 (DB1.0 in general)were produced by the same 420 writers, and the online database CASIA-OLHWDB1.1 and offline database CASIA-HWDB1.1 (DB1.1 in general) were produced by another 300 writers. Handwriting was produced using Anoto Pen on paper, so online trajectory data and offline image data can be obtained simultaneously. The character classes include 171 alphanumeric characters and symbols, and 3,866 Chinese characters (DB1.0) or 3,755 Chinese characters (DB1.1). The offline databases consist of gray-scale images with background removed. Experimental results show that feature extraction from gray-scale images yield significantly higher accuracies than that from binary images, and the recognition performance on both online and offline data reveals a big challenge to attack.

Keywords Handwritten Chinese character recognition, online, offline, databases, benchmarking

1. Introduction

Handwritten Chinese character recognition, including online (stroke trajectory-based) and offline (image-based) recognition, have received intensive attention since the early works in 1960s and 1970s. Particularly, there have been a boom of research from the 1980s owing to the popularity of personal computers and handy devices for data acquisition (laser scanners, writing tablets and PDAs). Successful applications have been found in document digitization and retrieval, postal mail sorting, bankcheck processing, form processing, pen-based text input, and so on.

Despite the tremendous advances and successful applications, there still remain big challenges, particularly, the recognition of unconstrained handwriting, including isolated characters and continuous scripts. Handwritten Chinese character recognition has reported accuracies of over 98% on sample databases of constrained handwriting but the accuracy on unconstrained handwriting is much lower [1]. Continuous handwritten script recognition is even more difficult because of the ambiguity of character segmentation.

To support academic research and benchmarking, the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA), has collected new databases of unconstrained Chinese handwriting. The handwritten data was generated using Anoto pen on paper such that both online and offline data can be obtained. The samples include both isolated handwritten characters and continuous scripts. A portion of online handwritten characters, in the database called CASIA-OLHWDB1[1], have been released at ICDAR 2009 [2]. To organize the Chinese handwriting recognition contest at 2010 Chinese Conference on Pattern Recognition (CCPR) [3], we have released four databases of isolated handwritten characters: online databases CASIA-OLHWDB1.0 and CASIA-OLHWDB1.1, offline databases CASIA-HWDB1.0 and CASIA-HWDB1.1.

To provide a benchmark for academic research in handwritten Chinese character recognition, we evaluate some state-of-the-art methods on the new unconstrained handwriting databases. The implemented methods include advanced character normalization methods [4], feature extraction methods for offline recognition [5][6]

[1] The database CASIA-OLHWDB1 was recently renamed as CASIA-OLHWDB1.0.

and online recognition [7][8], as well as classifier design methods [9-11]. The reported results thus provide a baseline for benchmarking of further research.

2. Databases

Many databases of handwritten Chinese and Japanese characters have been released but only the very recent ones target unconstrained handwriting.

The handwritten Japanese character database ETL9B contains 200 samples for each of 3,036 classes (including 2,965 Kanji characters). Reported accuracies on this database are mostly over 99%. A larger Japanese character database JEITA-HP contains 580 samples for each of 3,214 characters. High accuracies of over 98% can be obtained on this database [4].

In 2000, Beijing University of Posts and Telecommunications released a large database called HCL2000, which contains 1,000 samples for each of 3,755 characters [12]. This database is not challenging either, because high accuracies over 98% can be obtained [13]. In 2006, Harbin Institute of Technology (HIT) released a database of handwritten text pages called HIT-MW [14], which has 853 page images containing 186,444 characters produced by 780 writers. This is the first database of continuous Chinese handwriting but its scale is small and the images were not annotated.

For online character recognition, Tokyo University of Agriculture and Technology (TUAT) released two large databases Kuchibue and Nakayosi [15], containing samples written in boxes but in sequences of sentences, produced by 120 writers and 163 writers, respectively. The recognition of Kanji characters in these databases is not challenging, however (see the results in [7]). The South China University of Technology (SCUT) released a comprehensive online Chinese handwriting database SCUT-COUCH2009 [16]. It consists of 11 datasets of isolated characters (Chinese simplified and traditional, English letters, digits and symbols), Chinese Pinyin and words. The dataset GB1 contains 188 samples for each of 3,755 classes (level-1 set of GB2312-80 standard), produced by 188 writers. A state-of-the-art recognizer achieves 95.27% accuracy on it [16].

The NLPR of CASIA has been constructing new generation of unconstrained Chinese handwriting databases from 2007. The number of involved writers is over 1,300. Each writer wrote 3,661~4,037 isolated characters (including 171 alphanumeric characters and symbols) and five pages of texts (each page consists of 200~300 characters). Handwriting was produced using Anoto pen on paper such that online and offline data can be acquired concurrently. The isolated characters written on printed forms with spacious intervals, while texts were written without form. Online ink pages of isolated characters were segmented into characters according to between-stroke intervals and aligned with text transcripts. Online handwritten texts were segmented into text lines according to stroke gaps and into characters by transcript mapping embedding a character recognizer. Paper documents were scanned in 300DPI to acquire color images, from which dot patterns (pre-printed on Anoto paper) were separated by pixel classification based on color and local configuration. The foreground pixels are converted to gray scale and segmented into text lines and characters [17].

At the time of announcing Chinese Handwriting Contest in May 2010, we had released four databases of isolated characters: online databases CASIA-OLHWDB1.0 and CASIA-OLHWDB1.1, offline databases CASIA-HWDB1.0 and CASIA-HWDB1.1. The samples in OLHWDB1.0 and HWDB1.0 (DB1.0 in general) were produced by the same 420 writers, and the samples in OLHWDB1.1 and HWDB1.1 (DB1.1 in general) were produced by another 300 writers.

The samples of DB1.0 were written on forms with 4,037 pre-printed characters, and the samples of DB1.1 were written on forms with 3,926 pre-printed characters. In either case, the character set contains 171 alphanumeric characters and symbols. The character set of DB1.0 also contains 3,866 Chinese characters, 3,740 of which are contained in GB2312-80 level-1 set (GB1). The character set of DB1.1 contains exactly 3,755 Chinese characters of GB1. The Chinese characters of each set were pre-printed in six different orders to balance the writing quality variation of each writer through the writing process.

After annotation, the samples of each writer maybe less than the pre-printed character set because miswritten samples and those of ill-acquired signals (incomplete stroke trajectory or degraded scanned image) were removed. So, the online and offline sample sets of the same writer may have different numbers of samples. The offline datasets were removed more samples because of the low contrast of some scanned images.

The online databases provide the sequences of coordinates of strokes. The offline databases provide gray-scaled images with background pixels labeled as 255. So, it is easy to convert the gray-scale images to

binary images by simply labeling all the foreground pixels as 1 and background pixels as 0. The four databases are summarized in Table 1. Fig. 1 shows some samples of online and offline data produced by the same writer.

Table 1. Specifications of released databases.

	#writer	Total		GB1	
		#class	#sample	#class	#sample
OLHWDB1.0	420	4,037	1,694,741	3,740	1,570,051
HWDB1.0	420	4,037	1,680,258	3,740	1,556,675
OLHWDB1.1	300	3,926	1,174,364	3,755	1,123,132
HWDB1.1	300	3,926	1,172,907	3,755	1,121,749

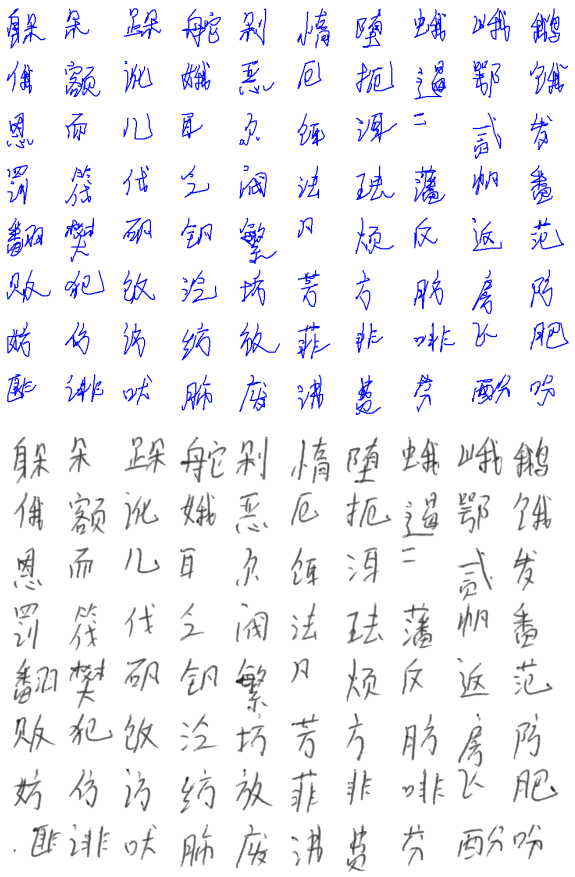

Figure 1. Online and offline samples of the same writer.

The databases OLHWDB1.0 and HWDB1.0 are partitioned into training set of 336 writers and test set of 84 writers. The databases OLHWDB1.1 and HWDB1.1 are partitioned into training set of 240 writers and test set of 60 writers.

We recommend researchers to use the databases OLHWDB1.1 and HWDB1.1 for basic research because they contain Chinese characters of a standard set, GB2312-80 level-1 set (GB1), and so, the recognition results on them can be compared with many previous results (e.g. [1][3][6][7]). To train classifiers using large dataset, the samples of 3,740 Chinese characters of GB1 in databases DB1.1 can be combined with the databases DB1.0.

3. Recognition Methods

3.1 Feature Extraction from Offline Samples

We evaluate recognition performance on both binary images and gray-scale images. For gray-scale images, the gray levels are reversed: background as 0 and foreground in 0~254, and foreground gray-levels are normalized to a specified range [18].

Both binary and gray-scale images are normalized using seven methods: linear normalization, nonlinear normalization [19], moment normalization, bi-moment normalization [20], pseudo 2D moment normalization (P2DMN), pseudo 2D bi-moment normalization (P2DBMN) and line density interpolation (LDI) [4].

For binary images, three feature extraction methods are evaluated: normalization-cooperated contour feature (NCCF) [5], normalization-based gradient feature (NBGF) and normalization-cooperated gradient feature (NCGF) [6]. In either case, contour/gradient elements are decomposed into 8 directions and each direction is extracted 8x8 values by Gaussian blurring. The NCCF is implemented based on the improved method of [5], called continuous NCFE. In any case, the feature dimensionality is 512.

From gray-scale images, two types of gradient features are extracted: NBGF and NCGF [6]. The methods are the same as for binary images.

3.2 Feature Extraction from Online Samples

From online character samples (sequences of stroke coordinates), we extract two types of direction features: histograms of original stroke direction and normalized direction [7]. The coordinate normalization methods include linear normalization, moment normalization, bi-moment normalization, pseudo 2D moment normalization (P2DMN) and pseudo 2D bi-moment normalization (P2DBMN). In all cases, the local stroke direction is decomposed into 8 directions and from the feature map of each direction, 8x8 values are extracted by Gaussian blurring. So, the dimensionality of feature vectors is 512.

We also implemented the direction feature of imaginary strokes (off-strokes) [8]. To minimize the

computation overhead, we simply add the direction values of off-strokes to real strokes with a weight of 0.5. So, the resulting feature vector dimensionality remains 512.

3.3 Classification Methods

We evaluate recognition accuracies using two classifiers: modified quadratic discriminant function (MQDF) [9] and nearest prototype classifier with discriminative feature extraction (DFE) [10]. For MQDF, we optimize the parameter of unified minor eigenvalue by holdout cross validation on training data. For accelerating MQDF, we first select 200 top rank classes according to Euclidean distance to class means, and then compare the MQDF values of 200 classes only. The accumulated accuracy of 200 candidate classes is mostly over 99.40%. For prototype classifier training, either with or without DFE, we take the new training criterion called logarithm of hypothesis margin (LOGM) [11].

For all classifiers, the input feature vector is first reduced from 512D to 160D by Fisher linear discriminant analysis (FLDA). For prototype classifier with DFE, the subspace parameters are further adjusted with the prototypes during discriminative learning.

4. Recognition Results

We first evaluated the recognition methods on databases OLHWDB1.1 and HWDB1.1. On selecting the best normalization and feature extraction methods, we then trained classifiers using the merged training data of DB1.0 and DB1.1.

The online database OLHWDB1.1 has 898,573 training samples and 224,559 test samples of GB1 character set. OLHWDB1.0 has 1,256,009 training samples and 314,042 test samples of GB1.

The offline database HWDB1.1 has 897,758 training samples and 223,991 test samples of GB1 character set. HWDB1.0 has 1,246,991 training samples and 309,684 test samples of GB1.

4.1 Online Recognition Results

Table 2 shows the test accuracies of online recognition on database OLHWDB1.1, based on combinations of five normalization methods and two types of direction features. The MQDF classifier gives results of both MQDF and Euclidean distance (minimum distance to class means). We can see that the 1D moment and bi-moment normalization methods yield higher accuracies than linear normalization, and pseudo 2D normalization methods can further improve the accuracies.

Table 3 shows the test accuracies of recognition with imaginary stroke features (original direction). Compared with Table 1, we can see that adding imaginary stroke features can improve recognition accuracies significantly. For example, when using normalization method P2DBMN, the accuracy of MQDF is promoted from 92.22% to 93.22%.

Based on the combination of P2DBMN and real+imaginary stroke features, MQDF and prototype classifiers were trained using merged training data of databases OLHWDB1.0 and OLHWDB1.1. The accuracies on test data of two databases are shown in Table 4, where Prototype-1 indicates prototype classifier with one prototype per class, Prototype-2 indicates classifier with two prototypes per class, Prototype-1-DFE and Prototype-2-DFE indicate prototype classifiers with DFE. Comparing with the accuracy of training with OLHWDB1.1 only, the accuracy on test set of OLHWDB1.1 is further improved from 93.22 to 93.95%. For prototype classifiers, DFE can improve the accuracy substantially, but its accuracy is still lower than that of MQDF. Prototype classifiers, however, have much lower complexity than the MQDF classifier.

Table 2. Test accuracies of online character recognition on OLHWDB1.1.

	Original direction		Normalized direction	
	MQDF	Euclid	MQDF	Euclid
Linear	85.99	72.35	86.13	72.46
Moment	91.69	85.16	91.58	85.12
Bi-moment	91.79	85.18	91.70	85.25
P2DMn	92.10	86.41	91.68	85.95
P2DBMN	**92.22**	86.75	**91.92**	86.33

Table 3. Test accuracies of online character recognition with imaginary stroke features on OLHWDB1.1.

	MQDF	Euclid
Linear	87.83	74.83
Moment	92.74	87.13
Bi-moment	92.80	87.03
P2DMn	93.08	88.10
P2DBMN	**93.22**	88.26

4.2 Offline Recognition Results

Table 5 shows the test accuracies of offline recognition on binary images of HWDB1.1. Comparing the three types of features, NBGF and NCCF perform

comparably, and NCGF shows obvious superiority, especially when combined with nonlinear and pseudo 2D normalization methods. This comparative relationship is consistent with that of [6].

Table 6 shows the test accuracies of offline recognition on gray-scale images. Again, NCGF yields higher accuracies than NBGF. And comparing with the performance on binary images, feature extraction from gray-scale images shows advantage. Specifically, it improves the test accuracy from 87.87% to 89.55%.

Based on the combination of pseudo 2D LDI normalization and NCGF on gray-scale images, MQDF and prototype classifiers were trained using merged training data of databases HWDB1.0 and HWDB1.1. The accuracies on test data of two databases are shown in Table 7. Comparing with the accuracy of training with HWDB1.1 only, the accuracy on test set of HWDB1.1 is further improved from 89.55 to 90.71%. Again, DFE effectively improves the accuracy of prototype classifiers, but its accuracy is still lower than that of MQDF.

Note that the accuracies of offline recognition are evidently lower than those of online recognition though the samples were produced by the same writers. This is because online data provides sequences of stroke coordinates such that local stroke direction features can be extracted more accurately, and further, the imaginary stroke feature effectively improves the recognition accuracy.

Table 6. Test accuracies of offline character recognition on gray-scale images of HWDB1.1.

	NBGF		NCGF	
	MQDF	Euclid	MQDF	Euclid
Linear	81.93	68.20	81.97	68.30
NLN	87.48	78.83	88.15	79.77
Moment	86.73	77.97	86.85	78.10
Bi-moment	87.09	78.50	87.28	78.68
P2DMN	87.11	79.06	88.01	80.41
P2DBMN	87.69	80.06	88.59	81.36
LDI	**88.55**	80.68	**89.55**	82.17

5. Concluding Remarks

We evaluated state-of-the-art online and offline handwritten character recognition methods on new databases of unconstrained Chinese handwriting. The results show that online data yields higher accuracies than offline data. For online data, imaginary stroke feature effectively improves the accuracy. For offline data, feature extraction from gray-scale images yields higher accuracies than that from binary images. The overall highest accuracies on both online and offline data are much lower than those on constrained data reported in the literature. This reveals a big challenge and leaves opportunities for basic research and improvement. The reported accuracies in this paper provide a benchmark for evaluating future research.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under grants no.60825301 and no.60933010. We thank the members of the PAL Group at the NLPR for their efforts in checking the sample databases.

References

[1] 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.

[2] D.-H. Wang, C.-L. Liu, J.-L. Yu, X.-D. Zhou, CASIA-OLHWDB1: A database of online handwritten Chinese characters, *Proc. 10th ICDAR*, Barcelona, Spain, 2009, pp.1206-1210.

[3] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, Chinese Handwriting Recognition Contest 2010, *Proc. 2010 Chinese Conference on Pattern Recognition (CCPR)*, Chongqing, China, 2010.

[4] C.-L. Liu, K. Marukawa, Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition, *Pattern Recognition*, 38(12): 2242-2255, 2005.

[5] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: Investigation of normalization and feature extraction techniques, *Pattern Recognition*, 37(2): 265-279, 2004.

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

[7] 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.

[8] K. Ding, G. Deng, L. Jin, An investigation of imaginary stroke technique for cursive online handwriting Chinese character recognition, *Proc. 10th ICDAR*, Barcelona, Spain, 2009, pp.531-535.

[9] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9(1): 149-153, 1987.

[10] C.-L. Liu, R. Mine, M. Koga, Building compact classifier for large character set recognition using discriminative feature extraction, *Proc. 8th ICDAR*, Seoul, Korea, 2005, pp.846-850.

[11] X.-B. Jin, C.-L. Liu, X. Hou, Regularized margin-based conditional log-likelihood loss for prototype learning, *Pattern Recognition*, 43(7): 2428-2438, 2010.

[12] H. Zhang, J. Guo, G. Chen, C. Li, HCL2000 – A large-scale handwritten Chinese character database for handwritten character recognition, *Proc. 10th ICDAR*, Barcelona, Spain, 2009, pp.286-290.

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

[14] 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.

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

[16] L. Jin, Y. Gao, G. Liu, Y. Li, K. Ding, SCUT-COUCH2009 – A comprehensive online unconstrained Chinese handwriting database and benchmark evaluation, *Int. J. Document Analysis and Recognition*, advanced version, 2010.

[17] F. Yin, Q.-F. Wang, C.-L. Liu, A tool for ground-truthing text lines and characters in off-line handwritten Chinese documents, *Proc. 10th ICDAR*, Barcelona, Spain, 2009, pp.951-955.

[18] C.-L. Liu, C.Y. Suen, A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters, *Pattern Recognition*, 42(12): 3287-3295, 2009.

[19] J. Tsukumo, H. Tanaka, Classification of handprinted Chinese characters using non-linear normalization and correlation methods, *Proc. 9th ICPR*, Rome, 1988, pp.168-171.

[20] C.-L. Liu, H. Sako, H. Fujisawa, Handwritten Chinese character recognition: Alternatives to nonlinear normalization, *Proc. 7th ICDAR*, Edinburgh, Scotland, 2003, pp.524-528.

Table 4. Test accuracies of online character recognition with merged training data.

	MQDF	Euclid	Prototype-1	Prototype -2	Prototype -1-DFE	Prototype -2-DFE
OLHWDB1.0-test	**94.45**	89.00	91.73	92.46	**92.73**	92.68
OLHWDB1.1-test	**93.95**	87.99	90.99	91.74	**92.15**	92.05

Table 5. Test accuracies of offline character recognition on binary images of HDWB1.1.

	NCCF		NBGF		NCGF	
	MQDF	Euclid	MQDF	Euclid	MQDF	Euclid
Linear	79.25	65.41	79.88	66.18	79.89	66.30
NLN	86.09	77.47	85.62	76.91	86.59	78.12
Moment	84.97	76.20	85.29	76.48	85.49	76.72
Bi-moment	85.37	76.80	85.61	77.06	85.87	77.32
P2DMN	86.39	78.79	85.73	77.68	86.65	79.13
P2DBMN	87.00	79.68	86.29	78.67	87.23	80.05
LDI	**87.49**	79.82	**86.70**	78.75	**87.87**	80.45

Table 7. Test accuracies of offline character recognition with merged training data.

	MQDF	Euclid	Prototype-1	Prototype-2	Prototype-1-DFE	Prototype-2-DFE
HWDB1.0-test	**93.00**	85.70	89.26	90.30	90.92	**91.43**
HWDB1.1-test	**90.71**	81.83	86.48	87.07	87.87	**88.57**