Chinese Handwriting Recognition Contest 2010

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

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

Abstract: Chinese handwriting recognition remains a challenge. Research works have reported very high accuracies on neatly handwritten characters yet the performance on unconstrained handwriting remains quite low. To promote the recognition technology, new databases of unconstrained handwriting have been constructed for academic research and benchmarking. This paper reports the contest results of online and offline handwritten Chinese character recognition using the new generation of databases, targeting 3,755 Chinese characters of the GB2312-80 first level set. Nine systems from four groups were submitted for evaluation. The best results are 92.39% accuracy for online character recognition and 89.99% accuracy for offline character recognition. Detailed analysis of results on data of different writers reveals the diversity of writing quality. The future contests will consider continuous script recognition as well as isolated character recognition.

Key Words: Chinese handwriting recognition contest, online handwritten character recognition, offline handwritten character recognition.

1 INTRODUCTION

Handwritten Chinese character recognition, including online (stroke trajectory-based) and offline (image-based) recognition, have received intensive attention. Despite the tremendous advances and successful applications, there still remain big challenges, particularly in unconstrained handwriting. Handwritten Chinese character recognition has reported accuracies as high as 98% on sample databases of constrained handwriting but the accuracy on unconstrained handwriting is much lower [1]. Continuous handwritten script recognition is an even more difficult problem. To promote the performance, research efforts are needed to design new methods and databases of unconstrained handwriting are needed for benchmarking.

In recent years, many competitions of handwriting recognition have been organized, such as the IWFHR 2006 online Tamil handwritten character recognition competition [2], the Arabic handwriting recognition competitions of ICDAR 2007 and 2009 [3][4], the handwriting segmentation competitions of ICDAR 2007 and 2009 [5][6]. Competitions have been effective to attract research attention and promote the technology. Particularly, the series of contests see evident increase of performance over time. Competitions of Chinese handwriting recognition, however, were not seen in recent years. In China, Chinese character recognition competitions were organized by the National High-Tech Program (863) Office for several times in 1990s, but the results were not opened.

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 handwriting 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 [7].

To evaluate the state of the art of Chinese handwriting recognition, we organize this contest. Since this is the first contest of a new series, we publicized only to Chinese researchers, and confine the target of recognition to isolated characters. Continuous Chinese script recognition is not yet undertook widely, and we postpone its evaluation to the future contests.

Isolated handwritten Chinese character recognition deserves attention and participation because it is an un-solved problem and is an integral part of continuous script recognition. We confine the character set to the 3,755 characters in the GB2312-80 level-1 set (GB1 in brief). This is meaningful in several respects. First, the characters in GB1 are among the most frequently used, occupy over 99% of usage in modern Chinese language. Second, the characters of high frequency tend to be written cursively and are more difficult to recognize. Third, increasing the character class number from 3,755 to a larger number does not alter the recognition methods.

In this contest 2010, we received nine systems submitted by four groups, including three for online character recognition and six for offline character recognition. On evaluation on a test sample set written by 60 persons, the best results are 92.39% accuracy for online character recognition and 89.99% accuracy for offline character recognition. This reveals a big gap between computer recognition performance and human recognition and leaves an opportunity for the research community to improve. We also analyze the performance on the datasets of different writers to investigate the diversity of writing quality.

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

978-1-4244-7210-9/10/$26.00 ©2010 IEEE

2 DATABASES

Many databases of handwritten Chinese and Japanese characters have been released. Among them the most famous is the ETL9B database of handwritten Japanese characters (200 samples for each of 3,036 classes, including 2,965 Kanji characters), which has been evaluated by many research works and has resulted in accuracies over 99%.

In 1990, a handwritten Chinese character image database was constructed by the Institute of Automation of Chinese Academy of Sciences (CASIA), which contains 300 samples for each of 3,755 characters (in GB1 set) but was not made public. The reported accuracy on this database is over 98% [1]. 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 [8]. This database is not challenging either because high accuracies over 98% can be obtained [9].

For online character recognition, Tokyo University of Agriculture and Technology (TUAT) released two databases Kuchibue and Nakayosi [10], produced by 120 writers and 163 writers, respectively. The recognition of Kanji characters in these databases is not challenging [11]. South China University of Technology (SCUT) released a large online Chinese handwriting database SCUT-COUCH2009 [12] consisting of 11 datasets of isolated characters, Chinese Pinyin and words. The dataset of GB1 contains 188 samples for each of 3,755 classes, and a state-of-the-art recognizer achieves 95.27% accuracy on it [12].

2.1. CASIA Databases

The NLPR of CASIA has been constructing new databases of unconstrained Chinese handwriting from 2007. The number of involved writers is over 1,000. 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 were written on printed forms with spacious intervals, while texts were written without form. Online ink documents were segmented into text lines and characters according to stroke gaps and transcript mapping. Paper documents were scanned in 300DPI to acquire color images, from which dot patterns (pre-printed on Anoto paper) were separated by pixel classification. The foreground pixels are converted to gray scale and segmented into text lines and characters [13].

At the time of contest announcement 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 were produced by the same 420 writers, and the samples in OLHWDB1.1 and HWDB1.1 were produced by the same 300 writers.

The samples of OLHWDB1.0 and HWDB1.0 were written on forms with 4,037 pre-printed characters, and the samples of OLHWDB1.1 and HWDB1.1 were written on forms with 3,926 pre-printed characters. The character set of OLHWDB1.0 and HWDB1.0 includes 3,866 Chinese characters, 3,740 of which are contained in GB2312-80 level-1 set (GB1). The character set of OLHWDB1.1 and HWDB1.1 includes exactly 3,755 Chinese characters of GB1. The Chinese characters of each set were pre-printed in six different orders to guide the writing order. During annotation, 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 number of samples.

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. The four databases are summarized in Table 1.

Table 1. Specifications of released databases.

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

The contest participants were recommended to use the released CASIA databases for training recognizers. For contest evaluation, we use a new sample set produced by 60 writers. The samples were written on papers of the same form as OLHWDB1.1 and HWDB1.1, and the Chinese characters of 3,755 classes in GB1 were extracted for contest. The sample numbers of online and offline contest datasets are 224,590 and 224,419, respectively.

3 PARTICIPATING SYSTEMS

We received nine participating systems from four groups, including three for online character recognition and six for offline character recognition.

3.1 Online Character Recognition Systems

CASIA-CREC: The Character Recognition Engineering Center of CASIA (CASIA-CREC, jointly owned by CASIA and Hanvon Technology Ltd.) submitted a system. The system extracts 1,024-dimensional direction feature after input pattern scaling by centroid alignment, and reduces feature vector to 128D subspace by principal component analysis (PCA). The classifier is a nearest prototype classifier discriminatively trained using the maximum mutual information (MMI) criterion. The training dataset contains 1,231,362 samples of 315 writers from CASIA-OLHWDB1.0 and 1,174,364 samples of 300 writers from CASIA-OLHWDB1.1. The total number of classes is 4,052.

SCUT-HCII: The Human-Computer Communication and Intelligent Interface Lab of SCUT (SCUT-HCII) submitted two systems, contributed by Yan Gao, Lingyu Liang, and Lianwen Jin. The underlying method extracts 8-direction

features (1,024D) of both real strokes and imaginary strokes [14], reduces to 160D subspace by linear discriminant analysis (LDA), and classifies using the modified quadratic discriminant function (MQDF) classifier [15]. The parameters of MQDF are compressed by splitVQ technique [16]. The two systems differ in the classification stage that SCUT-HCII-1 uses 12 principal components for MQDF while SCUT-HCII-2 uses 30 principal components. Moreover, SCUT-HCII-1 uses one MQDF classifier while SCUT-HCII-2 combines two MQDF classifiers. Both systems used all the GB1 samples of CASIA-OLHWDB1.0 and CASIA-OLHWDB1.1 for training.

The specifications of the online recognition systems are summarized in Table 2, where the last column shows the size of dictionary file (storing parameters).

3.2 Offline Character Recognition Systems

CASIA-CREC: The CASIA-CREC submitted three systems, using the same method but training with different datasets. The character image is normalized using the modified centroid-boundary alignment (MCBA) method [17]. 896D peripheral direction contributivity (PDC) feature is extracted [18] and is reduced to 128D by LDA. For classification, nearest prototype classifiers were trained using the learning vector quantization (LVQ3) algorithm of Kohonen. The training datasets of three systems are: (1) GB1 samples of CASIA-HWDB1.0 and CASIA-HWDB1.1; (2) samples of (1) plus Hanvon dataset 1 (about 10M samples); (3) samples of (1) plus Hanvon dataset 2 (about 10M samples).

HKU: The Department of Electrical and Electronic Engineering of University of Hong Kong (HKU) submitted a system, contributed by K.C. Leung and C.H. Leung based on the method in [19]. The input binary character image is normalized using the 2D nonlinear normalization method [20], 4-orientation chaincode feature (256D) is extracted and reduced to 216D subspace by LDA. The classifier is a regularized version of QDF (MQDF1 [15]) trained with both raw samples and artificially generated samples using a distortion model. The raw samples are the ones of GB1 in CASIA-HWDB1.0 and CASIA-HWDB1.1.

SCUT-HCII: The system submitted by SCUT-HCII uses linear normalization, 8-direction gradient feature extraction (512D) and dimensionality reduction by LDA (160D). The reduced vector is classified using the MQDF classifier (12 principal components), with parameters compressed by splitVQ. The training dataset contains the GB1 samples in CASIA-HWDB1.1.

WHU: The Department of Communications Engineering of Wuhan University (WHU) submitted a system contributed by Yankai Tu and Qinghu Chen. The recognition method is based on multiple features extraction and MQDF classification. On linear normalization of input image and skeletonization, three types of features are extracted: Gabor feature with elastic mesh, direction element feature and gradient direction feature. The obtained 1,024D feature vector is reduced to 256D by LDA. The MQDF classifier was trained with the GB1 samples in CASIA-HWDB1.1.

The specifications of the offline recognition systems are summarized in Table 3.

Table 2. Specifications of online character recognition systems.

	Normalization	Feature extraction	Dimension reduction	Classifier	#class	#training sample	Dictionary size
CASIA-CREC	Centroid alignment	Direction histograms, 1,024D	PCA, 128D	Prototypes, MMI training	4,052	2,405,726	2.86MB*
SCUT-HCII-1	Linear, elastic mesh	Direction of real & imaginary strokes, 1,024D	LDA, 160D	MQDF, splitVQ	3,755	2,693,183	4.23MB
SCUT-HCII-2	Linear, elastic mesh	Direction of real & imaginary strokes, 1,024D	LDA, 160D	Two MQDF classifiers combined	3,755	2,693,183	30.06MB

*Size of executive file embedding dictionary.

Table 3. Specifications of offline character recognition systems.

	Normalization	Feature extraction	Dimension reduction	Classifier	#class	#training sample	Dictionary size
CASIA-CREC-1	MCBA	PDC, 1,024D	LDA, 128D	Prototypes, LVQ3 training	3,755	2,678,424	5.71MB
CASIA-CREC-2	MCBA	PDC, 1,024D	LDA, 128D	Prototypes, LVQ3 training	3,755	12.6M	10.33MB
CASIA-CREC-2	MCBA	PDC, 1,024D	LDA, 128D	Prototypes, LVQ3 training	3,755	12.6M	12.17MB
HKU	2D NLN	Chaincode orientation, 256D	LDA, 216D	MQDF1, with distorted samples	3,755	2,678,424 (raw)	339.06M
SCUT-HCII	Linear, elastic mesh	Gradient direction, 512D	LDA, 160D	MQDF, splitVQ	3,755	1,121,749	4.15MB
WHU	Linear, elastic mesh	Three types of features, 1,024D	LDA, 256D	MQDF	3,755	1,121,749	29.34MB

4 RECOGNITION RESULTS

The submitted systems were evaluated on the contest evaluation dataset (60 writers, 224,590 online samples and 224,419 offline samples). The recognition systems were executed on a personal computer with Intel Core2-Duo-3.0GHz CPU, 2G RAM, and MS Windows XP OS. Each system loads the test samples from hard disc and stores the recognition results (10 candidate classes in decreasing order of confidence or increasing order of distance) of all samples in a result file of specified format. We count the correct rate of top candidate and the accumulated accuracy of 10 candidates. The average processing time is the division of total time (from system start to termination) by the number of test samples. The evaluation results of online recognition systems and offline recognition systems are listed in Table 4 and Table 5, respectively.

From the results of online recognition in Table 4, the system SCUT-HCII-2 gives the highest accuracy but the CASIA-CREC follows closely. The CASIA-CREC system has higher accumulated 10-candidate accuracy than the SCU-HCII systems and also runs faster. The high speed of CASIA-CREC system is attributed to its simple classifier structure (nearest prototype classifier). Yet by discriminative training, the classifier still yields fairly high accuracy. The MQDF classifier, used in SCUT-HCII systems, always yields high accuracies but its computation cost is appreciable even if after compression by splitVQ. All the three systems extract stroke direction histogram feature, which is widely acknowledged to be a superior feature in character recognition.

Table 4. Evaluation results of online recognition systems.

	Accuracy (top)	Accuracy (10)	Average time
CASIA-CREC	92.28%	**98.95%**	1.15ms
SCUT-HCII-1	91.48%	96.36%	2.97ms
SCUT-HCII-2	**92.39%**	97.55%	7.81ms

Table 5. Evaluation results of offline recognition systems.

	Accuracy (top)	Accuracy (10)	Average time
CASIA-CREC-1	83.02%	97.15%	1.78ms
CASIA-CREC-2	82.02%	96.75%	1.69ms
CASIA-CREC-3	82.45%	96.97%	2.72ms
HKU	**89.99%**	**98.64%**	250ms
SCUT-HCII	84.36%	93.52%	2.33ms
WHU	60.07%	88.14%	100ms

From the results of offline recognition in Table 5, the system of HKU gives by far the highest accuracy and accumulated accuracy. This is due to its MQDF1 classifier trained with large number of distorted samples. However, compared to the MQDF with reduced principal components, the MQDF1 has the same complexity with the QDF and thus costs much more storage and computation. The three CASIA-CREC systems yield comparable accuracies and speed though they were trained with different datasets. The system trained with CASIA-HWDB1.1 alone performs fairly well because the contest samples were written in similar environment with those of CASIA-HWDB1.1. The SCUT-HCII system yields higher accuracy due to the more powerful feature (gradient direction feature) and classifier (MQDF). The inferior performance of the WHU system indicates that the implementation was not optimized.

In both online recognition and offline recognition, the highest accuracy is quite low compared to the reported accuracies in the literature on other datasets (e.g., [1][11][19]). This justifies that the samples of unconstrained handwriting are indeed hard to recognize. To reveal the variation of writing quality over different writers, we give the accuracies on the dataset of each of 60 writers by the best offline recognition system of HKU, as shown in Fig. 1.

We can see that for most writers, the recognition accuracy is around 90%. In details, there are 5 writers with accuracy below 80%, 14 writers between 80%--90%, 26 writers between 90%--95%, and 15 writers with accuracy over 95%. The highest, medium (30th highest) and lowest accuracy are 98.05% (writer no.43), 92.90% (writer no.27) and 57.03% (writer no.34). Some samples of the three writers are shown in Fig. 2. We can see that even for the datasets of low accuracy, most samples are still human recognizable. This indicates a gap between computer handwriting recognition and human recognition, and an opportunity for research to improve the performance.

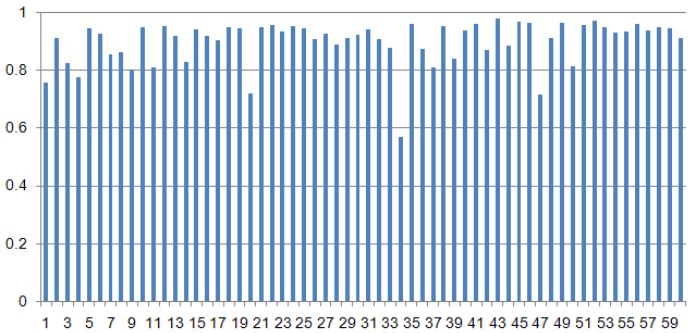
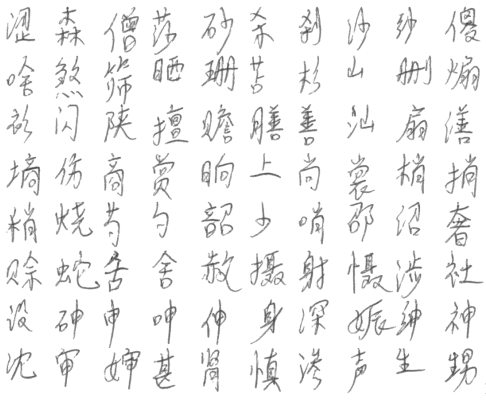


Fig. 1: The accuracies of offline recognition of datasets of 60 writers by the HKU system.



no.43

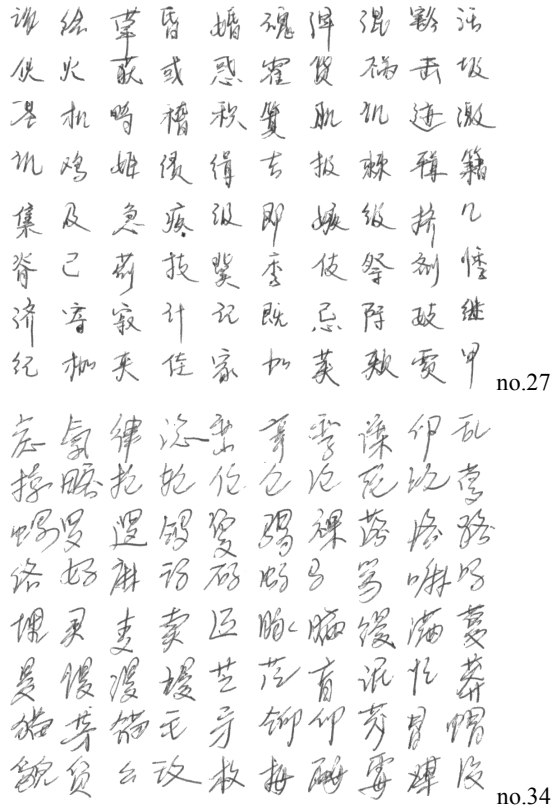


Fig.2. Samples of three writers of different quality.

5 CONCLUSION

We report the recognition results of Chinese Handwriting Recognition Contest 2010 based on new generation of databases of unconstrained handwriting. The contest focuses on online and offline isolated handwritten character recognition this year, but will extend to consider continuous handwritten scripts. The results of this year reveal that the performance of computer recognition of unconstrained handwriting is still far behind human recognition and application needs. This leaves an opportunity for the research community to improve the technology.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) under grants no.60825301 and no.60933010. We thank the contest participants for their active participation and feedbacks.

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] S. Madhvanath, S.M. Lucas, IWFHR 2006 online Tamil handwritten character recognition competition, *Proc. 10th IWFHR*, La Baule, France, 2006, pp.239-242.

[3] V. Margner, H. El Abed, ICDAR 2007 Arabic handwriting recognition competition, *Proc. 9th ICDAR*, Curitiba, Brazil, 2007, pp.1274-1278.

[4] V. Margner, H. El Abed, ICDAR 2009 Arabic handwriting recognition competition, pp.1381-1387.

[5] B. Gatos, N. Stamatopoulos, G. Lououdis, ICDAR 2007 handwriting segmentation contest, *Proc. 9th ICDAR*, Curitiba, Brazil, 2007, pp.1284-1288.

[6] B. Gatos, N. Stamatopoulos, G. Lououdis, ICDAR 2009 handwriting segmentation contest, pp.1393-1397.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] C.-L. Liu, K. Marukawa, Global shape normalization for handwritten Chinese character recognition: A new method, *Proc. 9th IWFHR*, Tokyo, Japan, 2004, pp.300-305.

[18] N. Hagita, S. Naito, I. Masuda, Handprinted Chinese characters recognition by peripheral direction contributivity feature, *Trans. IEICE Japan*, J66-D(10): 1185-1192, 1983 (in Japanese).

[19] K.C. Leung, C.H. Leung, Recognition of handwritten Chinese characters by combining regularization, Fisher's discriminant and distorted sample generation, *Proc. 10th ICDAR*, Barcelona, Spain, 2009, pp.1026–1030.

[20] T. Horiuchi, R. Haruki, H. Yamada, K. Yamamoto, Two-dimensional extension of nonlinear normalization method using line density for character recognition, *Proc. 4th ICDAR*, Ulm, Germany, 1997, pp.511-514.