Pattern Recognition 46 (2013) 155–162

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

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, No. 95 Zhongguancun East Road, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 19 March 2012
Received in revised form
25 June 2012
Accepted 28 June 2012
Available online 5 July 2012

Keywords:
Handwritten Chinese character recognition
Online
Offline
Databases
Benchmarking

ABSTRACT

Recently, the Institute of Automation of Chinese Academy of Sciences (CASIA) released the unconstrained online and offline Chinese handwriting databases CASIA-OLHWDB and CASIA-HWDB, which contain isolated character samples and handwritten texts produced by 1020 writers. This paper presents our benchmarking results using state-of-the-art methods on the isolated character datasets OLHWDB1.0 and HWDB1.0 (called DB1.0 in general), OLHWDB1.1 and HWDB1.1 (called DB1.1 in general). The DB1.1 covers 3755 Chinese character classes as in the level-1 set of GB2312-80. The evaluated methods include 1D and pseudo 2D normalization methods, gradient direction feature extraction from binary images and from gray-scale images, online stroke direction feature extraction from pen-down trajectory and from pen lifts, classification using the modified quadratic discriminant function (MQDF), discriminative feature extraction (DFE), and discriminative learning quadratic discriminant function (DLQDF). Our experiments reported the highest test accuracies 89.55% and 93.22% on the HWDB1.1 (offline) and OLHWDB1.1 (online), respectively, when using the MQDF classifier trained with DB1.1. When training with both the DB1.0 and DB1.1, the test accuracies on HWDB1.1 and OLHWDB are improved to 90.71% and 93.95%, respectively. Using DFE and DLQDF, the best results on HWDB1.1 and OLHWDB1.1 are 92.08% and 94.85%, respectively. Our results are comparable to the best results of the ICDAR2011 Chinese Handwriting Recognition Competition though we used less training samples.

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) [1,2]. Successful applications have been found in document digitization and retrieval, postal mail sorting, bankcheck processing, form processing, pen-based text input, and so on [3].

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 texts). Handwritten Chinese character recognition has reported accuracies of over 98% on sample datasets of constrained handwriting, but the accuracy on unconstrained handwriting is much lower [4]. Continuous handwritten script recognition is even more difficult because of the ambiguity of character segmentation. The results of the recent Chinese handwriting recognition competition reveal the challenge of both isolated character recognition and handwritten text recognition [5].

To support academic research and benchmarking, the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA), has built new databases of unconstrained Chinese handwriting. The handwritten data was produced using Anoto pen on paper such that both online and offline data were obtained concurrently. The samples include both isolated handwritten characters and continuous scripts. The online handwriting database CASIA-OLHWDB (OLHWDB in brief) and the offline database CASIA-HWDB (HWDB in brief), produced by 1020 writers, were released recently for free use in academic research [6]. Either the OLHWDB or the HWDB contain about 3.9 million isolated character samples and about 5090 handwritten text pages containing 1.35 million characters. The isolated character samples are divided into three datasets DB1.0–1.2, and the handwritten texts are divided into three datasets DB2.0–2.2 (with corresponding writers of DB1.0–1.2). The isolated samples involve 7356 character classes, including 7185 Chinese characters and 171 alphanumerics and symbols. The Chinese characters in DB1.1 (produced by 300 writers) fall in 3755 classes as in the

* Corresponding author. Tel.: +86 10 62558820; fax: +86 10 62551993.
E-mail address: liucl@nlpr.ia.ac.cn (C.-L. Liu).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.06.021

level-1 set of GB2312-80 (called GB1 in brief), which was often taken as a standard set of Chinese character recognition research. The DB1.0 (produced by 420 writers) involves 3866 frequently Chinese characters, with 3740 classes overlapping with the GB1 set. It is recommended to add the samples of DB1.0 to DB1.1 for enhancing the training dataset.

The databases CASIA-OLHWDB and CASIA-HWDB have been used for training in the competitions organized at 2010 Chinese Conference on Pattern Recognition (CCPR 2010) [7] and 11th International Conference on Document Analysis and Recognition (ICDAR 2011) [5]. The results of competition show improvements over time, and involve many different recognition methods. However, there is still a strong need of standard benchmark because the participating systems of competitions used different training datasets though reference datasets were recommended. Thus, this study provides a benchmark of online and offline handwritten Chinese character recognition on the new standard datasets. We only consider isolated handwritten Chinese character recognition in this study since it is still an un-solved problem, while the handwritten text recognition will be considered in-depth in other works.

As done in many previous works, we evaluate the recognition of the 3755 classes of the level-1 set of GB2312-80, as in the DB1.1 of CASIA-OLHWDB and CASIA-HWDB. We implement recognition systems using state-of-the-art methods of character normalization, feature extraction and classification. Specifically, we use 1D and pseudo 2D normalization methods [8], gradient direction feature extraction from binary images and from gray-scale images [9,10], online stroke direction feature extraction from pen-down trajectory and from pen lifts [11,12], classification using the modified quadratic discriminant function (MQDF) [13], nearest prototype classifier [14], discriminative feature extraction (DFE) [15], and discriminative learning quadratic discriminant function (DLQDF) [16]. We first compare normalization and feature extraction methods on the standard dataset DB1.1, then compare different classification methods using the combined training dataset of DB1.0 and DB1.1. The reported results provide some guidelines of methods selection, and serve as a baseline for evaluating the further works.

In the rest of this paper, we briefly introduce the datasets in Section 2, outline the recognition methods in Section 3, present and discuss the experimental results in Section 4, and give a conclusion in Section 5.

2. Datasets

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

The handwritten Japanese character database ETL9B contains 200 samples for each of 3036 classes (including 2965 Kanji characters). Reported accuracies on this database are mostly over 99%. A larger Japanese character database JEITA-HP contains 580 samples for each of 3214 characters, and high accuracies of over 98% have been reported [8]. In 2000, Beijing University of Posts and Telecommunications released a large database called HCL2000, which contains 1000 samples for each of 3755 characters [17]. This database is not challenging either, because high accuracies over 98% can be obtained [18].

For online character recognition, Tokyo University of Agriculture and Technology (TUAT) released two large databases called Kuchibue and Nakayosi [19], 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 [11]).

Table 1
Specifications of the isolated character datasets.

Dataset	Total			GB1			
	# writer	# class	# sample	# class	# sample	Training	Test
OLHWDB1.0	420	4037	1,694,741	3740	1,570,051	1,256,009	314,042
HWDB1.0	420	4037	1,680,258	3740	1,556,675	1,246,991	309,684
OLHWDB1.1	300	3926	1,174,364	3755	1,123,132	898,573	224,559
HWDB1.1	300	3926	1,172,907	3755	1,121,749	897,758	223,991

The South China University of Technology (SCUT) released a comprehensive online Chinese handwriting database SCUT-COUCH2009 [20]. 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 3755 classes (level-1 set of GB2312-80 standard), produced by 188 writers. A state-of-the-art recognizer achieves 95.27% accuracy on it [20].

The new databases CASIA-OLHWDB and CASIA-HWDB (details can be found in [6]) have some outstanding features compared to the previous ones: unconstrained writing, concurrent online and offline data, combination of isolated samples and continuous scripts, deep annotation of script data, large category set, large number of writers and samples. For the research of isolated character recognition, we recommend to use the datasets OLHWDB1.1 and HWDB1.1 (called DB1.1 in general), OLHWDB1.0 and HWDB1.0 (called DB1.0 in general). The Chinese characters in DB1.1 fall in the 3755 classes of the standard level-1 set of GB2312-80 (GB1 set), while the DB1.0 has 3740 classes overlapping with the GB1 set.

The online datasets provide the sequences of coordinates of strokes. The offline datasets 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. Nevertheless, to exploit the gray level information is generally beneficial. The four datasets, online and offline DB1.0 and DB1.1, which are used in our experiments, are summarized in Table 1. The datasets OLHWDB1.0 and HWDB1.0 are partitioned into training set of 336 writers and test set of 84 writers. The datasets OLHWDB1.1 and HWDB1.1 are partitioned into training set of 240 writers and test set of 60 writers. The training set and the test set are disjoint and produced by totally different writers.

Fig. 1 shows some samples of online and offline data produced by the same writer.

3. Recognition methods

A character recognition system generally consists of three major components: character normalization, feature extraction, and classification. Usually, the classification method does not differ for online or offline recognition, but the normalization and feature extraction methods depend on the type of input data. In the following, we outline the normalization and feature extraction methods for offline recognition and for online recognition separately, and then give the classification methods.

3.1. Normalization and feature extraction for 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 for

overcoming the gray scale variation among different images [21]. We consider two types of gray level normalization: linear and nonlinear. Linear normalization, as used in [21], re-scales the mean and standard deviation (s.d.) of foreground gray levels of the original image to specified values. Denote the mean and s.d. of original image as m and σ, respectively, which are transformed to standard values m_0 and σ_0, respectively, the original pixel gray level g is transformed to g' by

$$g' = (g-m)\cdot\frac{\sigma_0}{\sigma}+m_0. \tag{1}$$

The linear gray level normalization has a shortage that it does not map the original gray level 0 to normalized gray level 0, though we can artificially bound the gray levels in [0,255]. Consider that the gray level mostly lies in $[m-2\sigma,m+2\sigma]$, the linear normalization maps it to $[m_0-2\sigma_0,m_0+2\sigma_0]$. To get a smooth nonlinear gray level mapping, we use a nonlinear function

$$g' = \alpha g^p \tag{2}$$

that maps three values $\{0,m,m+2\sigma\}$ to normalized values $\{0,m_0,m_0+2\sigma_0\}$. The constraints give parameters

$$p = \frac{\log\frac{m_0}{m_0+2\sigma_0}}{\log\frac{m}{m+2\sigma}}$$

$$\alpha = \frac{m_0}{m^p} \tag{3}$$

Fig. 2 shows the curves of linear and nonlinear gray level mapping. We can see that when the gray level variance is big (Fig. 2(a)), the nonlinear mapping tends to enlarge the contrast of foreground pixels with gray level $g<m$; when the gray level variance is small (Fig. 2(b)), the nonlinear mapping tends to moderate the contrast of foreground pixels with gray level $g<m$. The case of Fig. 2(a) occurs often in Chinese character images when there are many strokes and the between-stroke gap is blurred and fail to be separated by binarization. Fig. 3 shows some character images where nonlinear gray level normalization gives better contrast for between-stroke gaps.

We can normalize both binary and gray-scale character images using seven methods: linear normalization (LN), nonlinear normalization (NLN) based on line density equalization [22], moment normalization (MN), bi-moment normalization (BMN) [23], pseudo

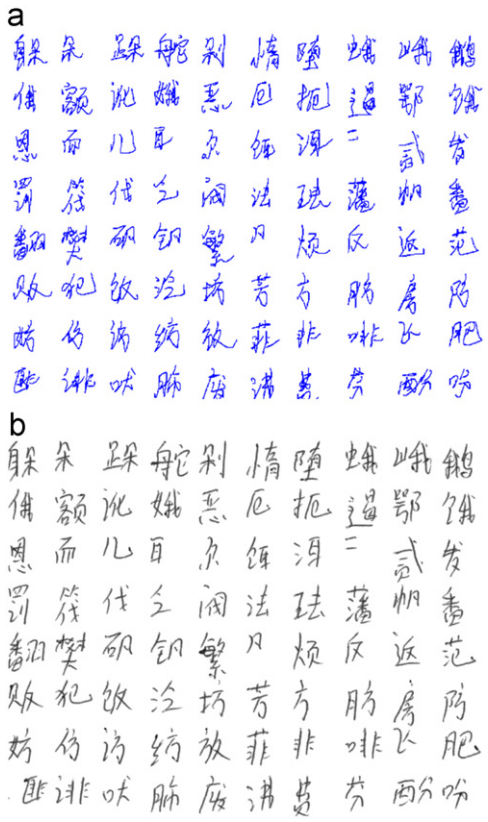

Fig. 1. Online (a) and offline (b) character samples of the same writer.

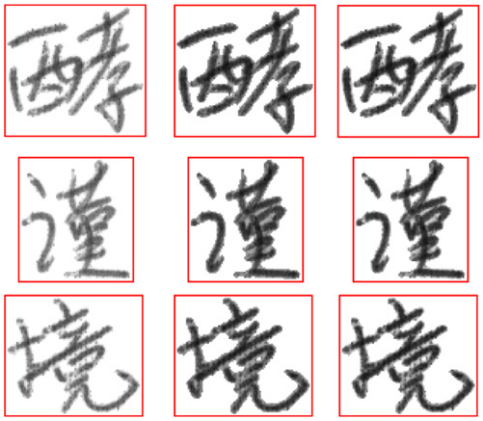

Fig. 3. Left: original image; middle: linear gray level normalization; right: nonlinear gray level normalization. The gray level of original image is reversed before normalization, and the gray level normalized images are reversed again for display.

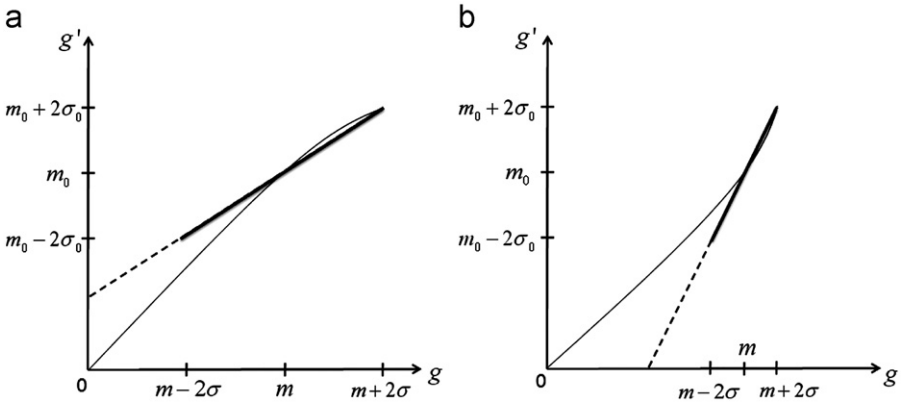

Fig. 2. Linear (thick line) and nonlinear (thin line) gray level normalization.

Fig. 4. Examples of character normalization. From left to right: input (gray level normalized), LN, NLN, MN, BMN, P2DMN, P2DBMN, and LDPI.

2D moment normalization (P2DMN), pseudo 2D bi-moment normalization (P2DBMN) and line density projection interpolation (LDPI) [8]. The details of all these methods can be found in [8]. For normalizing gray-scale images, the foreground pixel gray values are directly used in computing the horizontal/vertical projections in MN, BMN, P2DMN and P2DBMN. For the methods LN, NLN and LDPI, the character boundary and line density functions are calculated from the binary image while the gray values are transferred to the normalized image after the pixel coordinate mapping functions are obtained from the boundary and line density functions. Fig. 4 shows some examples of gray-scale character image normalization using the seven methods.

For binary images, three feature extraction methods are evaluated: normalization-cooperated contour feature (NCCF) [9], normalization-based gradient feature (NBGF) and normalization-cooperated gradient feature (NCGF) [10]. In either case, contour/gradient elements are decomposed into 8 directions and each direction is extracted 8×8 values by Gaussian blurring, and so, the feature dimensionality is 512. A contour element is a contour pixel in one of eight chaincode directions. The gradient is computed by the Sobel operator and its direction is decomposed into its two adjacent standard chaincode directions by the parallelogram rule. The NCCF is implemented based on the improved method of [9], called continuous NCFE. By NCCF or NCGF, the normalized character image is not generated, but instead, the contour/gradient elements of the original image are directly mapped to direction maps incorporating pixel coordinates transformation. By NBGF, the features are extracted from the normalized image. The details of the three feature extraction methods can be found in [9,10].

For gray-scale images, the contour feature extraction method NCCF is not available, but the gradient feature extraction methods NBGF and NCGF are directly applicable.

3.2. Normalization and feature extraction for 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 [11]. In either case, the features are extracted from the original pattern incorporating coordinate transformation without generating the normalized pattern. The coordinate normalization methods include linear normalization (LN), moment normalization (MN), bi-moment normalization (BMN), pseudo 2D moment normalization (P2DMN) and pseudo 2D bi-moment normalization (P2DBMN). The line density based normalization methods in offline character recognition, NLN and LDPI, are not applicable to online trajectory because the line density needs to be calculated on raster image. In feature extraction, the local stroke direction (of the line segment formed by two adjacent points) is decomposed into eight directions and from the feature map of each direction, 8×8 values are extracted by Gaussian blurring. So, the dimensionality of feature vectors is 512. The details of the normalization and direction feature extraction methods can be found in [11].

We also implemented the direction feature of imaginary strokes (pen lifts or called off-strokes) [12]. To minimize the computation overhead, we simply add the direction values of imaginary strokes to the direction histograms of real strokes with a weight of 0.5 to get enhanced direction features. So, the resulting feature vector dimensionality remains 512.

3.3. Classification methods

We evaluate recognition accuracies using four types of classifiers: modified quadratic discriminant function (MQDF) [13], nearest prototype classifier (NPC) [14], NPC with discriminative feature extraction (DFE) [15], and discriminative quadratic discriminant function (DLQDF) [16]. Before classification, the feature vector is reduced to a low-dimensional subspace learned by Fisher discriminant analysis (FDA). Before dimensionality reduction, every feature is transformed by variable transformation $y=x^{0.5}$, also known as Box–Cox transformation [24].

The MQDF is a modification of the multivariate Gaussian-based QDF by regulating the minor eigenvalues of each class to a constant, such that the discriminant function can be calculated from the principal eigenvalues and their corresponding eigenvectors only, and the regulation of minor eigenvalues benefits the generalization performance. Denote the d-dimensional feature vector (after dimensionality reduction) by $\mathbf{x}$, the MQDF of class ω_i ($i=1, \ldots, M$) is computed by

$$g_2(\mathbf{x},\omega_i)=\sum_{j=1}^{k}\frac{1}{\lambda_{ij}}[(\mathbf{x}-\mu_i)^T\phi_{ij}]^2 + \frac{1}{\delta_i}\left\{\|x-\mu_i\|^2-\sum_{j=1}^{k}[(\mathbf{x}-\mu_i)^T\phi_{ij}]^2\right\} + \sum_{j=1}^{k}\log\lambda_{ij}+(d-k)\log\delta_i \quad (4)$$

where μ_i is the mean vector of class ω_i, λ_{ij} and ϕ_{ij}, $j=1, \ldots, d$, are the eigenvalues (sorted in nonascending order) and their corresponding eigenvectors of the covariance matrix of class ω_i. k denotes the number of principal eigenvectors, and the minor eigenvalues are replaced with a constant δ_i.

The nearest prototype classifier has lower runtime computation cost than the MQDF when using few prototypes per class. There are many algorithms for prototype learning, including clustering, learning vector quantization (LVQ) [25] and many generalized discriminative learning methods. We use a recent learning algorithm which optimizes the log-likelihood of hypothesis margin (LOGM) [14]. The LOGM is a modification of the minimum classification error (MCE) criterion [26] for improving the training convergence and generalization performance. The DFE is a discriminative subspace learning method, mostly combined with prototype learning: the subspace (initially learned by FDA) is optimized jointly with the prototypes by stochastic gradient descent, usually under the MCE criterion [15]. In our implementation of DFE, we use the LOGM criterion to replace the MCE criterion.

The DLQDF is a discriminative version of MQDF, with the parameters (class means, eigenvectors and eigenvalues) optimized by stochastic gradient descent under the MCE objective. We also use the LOGM criterion to replace the MCE in learning the DLQDF. Either the MQDF or the DLQDF can be used in the feature subspace learned by DFE, i.e., learning the subspace jointly with the NPC and then applying the learned subspace to MQDF and DLQDF [27].

In our implementation of MQDF, we use a unified minor eigenvalue $\delta_i = \delta_0$ for all classes and optimize this parameter by 5-fold holdout cross validation on training data. Particularly, we set δ_0 as $(\beta/Md)\sum_{i=1}^{M}\sum_{j=1}^{d}\lambda_{ij}$ with β selected from $[0, 1]$. For selecting the value of β, we use 4/5 of training data for training classifier with different values of β and using the remaining 1/5 of training data for validation, and on selecting the best β value, re-train the classifier using the whole training data. For accelerating MQDF classification, we first select 200 top rank classes according to the Euclidean distance to class means, and then compare the MQDF values of 200 classes only. This causes little loss of accuracy for MQDF because the accumulated accuracy of 200 candidate classes is mostly over 99.40% on our datasets. In training the DLQDF, we use the parameters of MQDF as initial values.

Besides the DLQDF, there are some alternative improved versions of MQDF proposed in recent years, including the MQDF with pairwise discrimination [28,29], compact MQDF with subspace parameter compression [30], compressed and discriminatively trained Gaussian model [31], and MQDF estimated from reweighted samples [32]. All these methods effect in improving the classification accuracy of MQDF or reducing the storage and computation complexity. We did not endeavor to implement these methods in our experiments because the DLQDF is already effective to achieve high accuracies. Other classifiers that have shown superiority in many areas, such as the support vector machine (SVM) and the Gaussian process classifier, are not suitable for Chinese character recognition because of their extremely high complexity on large category problems.

4. Recognition results

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

As shown in Table 1, the offline dataset HWDB1.1 has 897,758 training samples and 223,991 test samples of GB1 character set (3755 classes). The online dataset OLHWDB1.1 has 898,573 training samples and 224,559 test samples of GB1 character set.

In all the experiments, we fixed the feature subspace dimensionality as 160. Our experience shows that increasing the subspace dimensionality from 160 leads to slight improvement of recognition accuracy but increases the computational complexity evidently.

4.1. Offline recognition results on HWDB1.1

First, to justify the benefits of nonlinear gray level normalization introduced in Section 3.1, we conducted an experiment to compare the recognition performance of linear and nonlinear gray level normalization, using a standard feature extraction method NCGF combined with character normalization method NLN, classification by MQDF and Euclidean distance classifiers with dimensionality reduction by FDA. We tested some tentative values of standard mean m_0 and deviation σ_0 of foreground gray level which approximately meet $m_0+2\sigma_0<255$ and $m_0-2\sigma_0>50$ such that after gray level normalization, the contrast among foreground pixels and the one between foreground and background are both sufficient. Using the training samples of HWDB1.1 for training the classifiers, the accuracies on the test set of HWDB1.1 are shown in Table 2. We can see that in most settings of m_0 and σ_0, nonlinear gray level normalization yields higher accuracies than linear gray level normalization. Nonlinear gray level normalization with $m_0=200$ and $\sigma_0=30$ achieves the highest accuracies. Consider that when $m_0=180$, the recognition performance is comparable and more stable against different values of σ_0, we choose nonlinear gray level normalization with $m_0=180$ and $\sigma_0=30$ for the remaining experiments.

On binary images of HWDB1.1, we can evaluate three feature extraction methods NBGF, NCGF and NCCF, and different character normalization methods. The test accuracies are shown in Table 3. Comparing the three types of features, we can see that NBGF and NCCF perform comparably, and NCGF shows obvious superiority, especially when combined with nonlinear normalization and pseudo 2D normalization methods. This comparative relationship is consistent with that reported in the literature [10].

Table 4 shows the test accuracies of offline recognition on gray-scale images (in this case, the contour feature does not apply). Again, NCGF yields higher accuracies than NBGF. And comparing with the performance on binary images, feature extraction from gray-scale

Table 2
Test accuracies (%) of linear and nonlinear gray level normalization on HWDB1.1.

m_0	σ_0	Linear		Nonlinear	
		MQDF	Euclid	MQDF	Euclid
180	20	87.69	79.40	88.03	79.69
180	30	87.92	79.60	88.15	79.77
180	40	88.08	79.64	88.13	79.74
200	20	87.65	79.31	87.97	79.62
200	30	87.90	79.56	**88.17**	**79.79**
160	30	87.97	79.62	88.14	79.75
160	40	**88.10**	**79.66**	88.15	79.66
140	40	88.08	79.61	88.12	79.57
140	50	87.97	79.42	87.97	79.40

Table 3
Test accuracies (%) of offline character recognition on binary images of HWDB1.1.

Normalization	NBGF		NCGF		NCCF	
	MQDF	Euclid	MQDF	Euclid	MQDF	Euclid
LN	79.88	66.18	79.89	66.3	79.25	65.41
NLN	85.62	76.91	86.59	78.12	86.09	77.47
MN	85.29	76.48	85.49	76.72	84.97	76.20
BMN	85.61	77.06	85.87	77.32	85.37	76.80
P2DMN	85.73	77.68	86.65	79.13	86.39	78.79
P2DBMN	86.29	78.67	87.23	80.05	87.00	79.68
LDPI	**86.70**	**78.75**	**87.87**	**80.45**	**87.49**	**79.82**

Table 4
Test accuracies (%) of offline character recognition on gray-scale images of HWDB1.1.

Normalization	NBGF		NCGF	
	MQDF	Euclid	MQDF	Euclid
LN	81.93	68.2	81.97	68.30
NLN	87.48	78.83	88.15	79.77
MN	86.73	77.97	86.85	78.10
BMN	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
LDPI	**88.55**	**80.68**	**89.55**	**82.17**

Table 5
Test accuracies (%) of online character recognition on OLHWDB1.1.

Normalization	Original direction		Normalized direction		Original enhanced	
	MQDF	Euclid	MQDF	Euclid	MQDF	Euclid
LN	85.99	72.35	86.13	72.46	87.83	74.83
MN	91.69	85.16	91.58	85.12	92.74	87.13
BMN	91.79	85.18	91.70	85.25	92.80	87.03
P2DMN	92.10	86.41	91.68	85.95	93.08	88.10
P2DBMN	**92.22**	**86.75**	**91.92**	**86.33**	**93.22**	**88.26**

images shows apparent advantage. Specifically, it improves the test accuracy of MQDF from 87.87% to 89.55%.

4.2. Online recognition results on OLHWDB1.1

Table 5 shows the test accuracies of online recognition on dataset OLHWDB1.1, based on five normalization methods, two types of direction features (original direction and normalized direction), as well as the original direction feature enhanced with pen lifts (called "Original enhanced" in Table 5). We can see that the pseudo 2D normalization methods (P2DMN and P2DBMN) yield higher accuracies than the 1D normalization methods, and the feature of original direction outperforms that of normalized direction. Enhancing the feature of original direction with pen lifts, the recognition accuracy is further improved, specifically, from 92.22% to 93.22% by MQDF classification.

4.3. Recognition results on combined DB1.0 and DB1.1

The above experimental results on standard datasets of 3755 classes in GB1 can be used as benchmarks for handwritten Chinese character recognition research. We further report recognition results of various classifiers using larger training datasets combining DB1.0 (HWDB1.0 or OLHWDB1.0) and DB1.1. Combining the training samples of GB1 set in DB1.0 and DB1.1 (Table 1), we obtained training datasets of 2,144,749 offline samples and 2,154,582 online samples, respectively.

In experiments on the larger datasets, we used the best combination of normalization and feature extraction methods, i.e., LDPI normalization and NCGF for offline samples, P2DBMN and enhanced original direction feature for online samples. The feature dimensionality remains 512, which is reduced to 160 by FDA and DFE. We report the recognition accuracies on the test datasets of both DB1.0 and DB1.1.

For the nearest prototype classifier (NPC) and DFE, we evaluated the recognition performance with variable number of prototypes per class. Prototype learning by k-means clustering (using cluster centers of each class as prototypes) and LOGM [14], as well as joint prototype and subspace (initialized by FDA) learning by DFE (LOGM criterion) were evaluated. For MQDF and DLQDF, we give the results of MQDF with numbers of principal eigenvectors per class as $k=40$ and $k=50$, and DLQDF with $k=40$. We did not implement the DLQDF with $k=50$ because the MQDF/DLQDF versions with $k=40$ and $k=50$ give comparable performance, and the training of DLQDF on large dataset is very time consuming. On a PC with Quad CPU Q9550 2.83 GHz, the training time with 2.1 million samples of 160D was 46.5 h even though we used acceleration technique based on hierarchical classification [27].

First, we see the results of offline character recognition. The test accuracies of NPC on HWDB1.0 and HWDB1.1 are shown in Table 6, and the results of MQDF and DLQDF are shown in Table 7. For the NPC, we used 1, 2, 3, 4, and 5 prototypes per class. Increasing prototypes further was shown to give little improvement for LOGM and LOGM+DFE. It is not a surprise that the supervised prototype learning algorithm LOGM yields higher accuracies than k-means clustering, and LOGM+DFE further improves the accuracy. However, the accuracies of LOGM+DFE are lower than those of MQDF and DLQDF as shown in Table 7, which also gives the accuracies on the test dataset of ICDAR 2011 Competition [5]. We can see that using the same classifier, the test dataset of DB1.0 has higher accuracy than the DB1.1. This confirms that the DB1.0 has better writing quality than the DB1.1 (in database building, the DB1.0 abandoned more samples of low quality than the DB1.1). The quality of the competition data is between the DB1.0 and the DB1.1.

Table 6
Test accuracies (%) of offline character recognition by NPC trained with HWDB1.0 and HWDB1.1.

# prototype per class	Cluster		LOGM		LOGM+DFE	
	DB1.0	DB1.1	DB1.0	DB1.1	DB1.0	DB1.1
1	85.70	81.83	89.23	85.82	90.92	87.89
2	87.56	84.02	90.30	87.07	91.47	88.58
3	88.28	84.83	90.50	87.39	91.42	88.46
4	88.65	85.32	90.53	**87.46**	91.39	88.49
5	**88.81**	**85.58**	**90.59**	87.43	**91.47**	**88.59**

Table 7
Test accuracies (%) of offline character recognition by MQDF and DLQDF trained with HWDB1.0 and HWDB1.1.

Subspace	Classifier (k)	DB1.0	DB1.1	Competition
FDA	MQDF(40)	92.94	90.68	91.51
	MQDF(50)	93.00	90.71	91.57
	DLQDF(40)	93.33	91.00	91.78
DFE	MQDF(40)	93.95	91.89	92.60
	MQDF(50)	93.94	91.92	92.64
	DLQDF(40)	**94.20**	**92.08**	**92.72**

Using classifiers trained with the merged training set of HWDB1.0 and HWDB1.1, the MQDF ($k=50$) gave accuracy 90.71% on HWDB1.1 test set, which is higher than the accuracy 89.55% of the classifier trained with HWDB1.1 only. The DLQDF improved the accuracy of MQDF ($k=40$) from 90.68% to 91.00%. Using the feature subspace learned by DFE instead of that of FDA, the test accuracy of MQDF ($k=50$) was improved from 90.71% to 91.92%. In the subspace learned by DFE, DLQDF improves the accuracy of MQDF ($k=40$) from 91.89% to 92.08%, which is the highest accuracy that we achieved in this study on the HWDB1.1 test set.

As for online character recognition, the test accuracies of NPC on OLHWDB1.0 and OLHWDB1.1 are shown in Table 8, and the results of MQDF and DLQDF are shown in Table 9. The NPC again used 1, 2, 3, 4, and 5 prototypes per class, and again, the joint prototype and subspace learning algorithm LOGM+DFE outperforms the k-means clustering and LOGM prototype learning. Using classifiers trained with the merged training data of OLHWDB1.0 and OLHWDB1.1, the MQDF ($k=50$) gave test accuracy 93.95% on OLHWDB1.1, which is higher than the accuracy 93.22% of the classifier trained with OLHWDB1.1 only. The DLQDF improved the accuracy of MQDF ($k=40$) from 93.90% to 94.29%. Using the feature subspace learned by DFE instead of that of FDA, the test accuracy of MQDF ($k=50$) was improved from 93.95% to 94.68%. In the subspace learned by DFE, DLQDF improves the accuracy of MQDF ($k=40$) from 94.65% to 94.85%.

We give the accuracies on the competition test dataset of ICDAR 2011 Competition [5] (60 writers, 224,419 offline samples, and 224,590 online samples) to demonstrate that the methods used in this study are state-of-the-art. The ICDAR 2011 Competition reported

Table 8
Test accuracies (%) of online character recognition by NPC trained with OLHWDB1.0 and OLHWDB1.1.

# prototype per class	Cluster		LOGM		LOGM+DFE	
	DB1.0	DB1.1	DB1.0	DB1.1	DB1.0	DB1.1
1	89.00	87.99	91.73	90.99	92.73	92.15
2	90.38	89.38	92.46	91.74	92.68	92.05
3	90.88	89.98	92.61	91.95	92.77	92.15
4	91.19	90.31	92.73	**92.09**	92.95	92.33
5	**91.38**	**90.52**	**92.77**	92.05	**93.11**	**92.49**

Table 9
Test accuracies (%) of online character recognition by MQDF and DLQDF trained with OLHWDB1.0 and OLHWDB1.1.

Subspace	Classifier (k)	DB1.0	DB1.1	Competition
FDA	MQDF(40)	94.39	93.90	94.26
	MQDF(50)	94.45	93.95	94.31
	DLQDF(40)	94.81	94.29	94.69
DFE	MQDF(40)	95.09	94.65	95.08
	MQDF(50)	95.12	94.68	95.12
	DLQDF(40)	**95.28**	**94.85**	**95.31**

Fig. 5. Some misrecognized samples in HWDB1.1. Below each image are the ground-truth class and classification outcome.

the highest accuracy 92.18% in offline character recognition and 95.77% in online character recognition, both were achieved by training with larger number of samples (including distorted samples and all the samples of DB1.0 and DB1.1). Our best result of 92.72% in offline recognition is superior to that of the ICDAR 2011 Competition. Our accuracy of 95.31% in online recognition is slightly lower than that of the ICDAR2011 Competition. Consider the training set size of the winning recognizers of the competition, our recognizers trained with much less samples are sufficiently competitive.

4.4. Examples of recognition errors

We show some misrecognized offline samples by the best classifier DLQDF in DFE subspace trained with combined HWDB1.0 and HWDB1.1. The confusion of online samples is similar to that of offline samples since the online and offline samples were produced concurrently by the same writers. Some misrecognized samples are shown in Fig. 5, where we can see that the ground-truth class and the assigned class (classification outcome) are indeed confusing in shape, particularly the samples in the top two rows. The presence of many cursive and confusing samples in the new databases CASIA-HWDB and CASIA-OLHWDB makes it very difficult to achieve high accuracies.

Fig. 6. Some miswritten samples in HWDB1.1. Below each image are the labeled class and classification outcome.

There are also some miswritten characters in the databases, though very few, that were not removed during ground-truthing. Fig. 6 shows some examples of such characters. Compared to the remarkable error rates (7.91% on HWDB1.1 and 5.15% on OLHWDB1.1), the very small number of miswritten samples does not affect the confidence of performance evaluation.

5. Conclusion

We evaluated state-of-the-art online and offline handwritten character recognition methods on the new large scale, unconstrained Chinese handwriting databases CASIA-HWDB and CASIA-OLHWDB. The results on the isolated character datasets of 3755 classes can serve as benchmarks for evaluating recognition methods. On the new datasets, the highest accuracies achieved by the state-of-the-art methods (92.08% on offline dataset HWDB1.1 and 94.85% on online dataset OLHWDB1.1) are far lower than the accuracies on previous constrained datasets (mostly over 98%). This indicates that isolated handwritten Chinese character recognition (HCCR) is still a challenge, and it opens a large room for research and improvement. We achieved best performance using pseudo 2D character normalization, normalization-cooperated gradient/trajectory direction feature extraction, subspace learning by discriminative feature extraction (DFE) and classification by discriminative learning quadratic discriminant function (DLQDF). The community is expected to propose more effective methods and achieve higher performance in the future.

Our recommendations for evaluating isolated HCCR (either online or offline) are as follows. (1) Using the standard dataset DB1.1 for evaluating character normalization and feature extraction methods with standard classifiers of MQDF and Euclidean distance. The DB1.1 has handwritten samples of 3755 classes of GB2312-80 level-1 set (GB1) produced by 300 writers. (2) For training discriminative classifiers using larger dataset, combining the training samples of DB1.0 and DB1.1. The DB1.0 has samples of 420 writers, with 3740 classes falling in GB1. (3) More samples can be generated by synthesizing or distorting handwritten samples. Other research issues have been discussed in [6].

For evaluating classification algorithms on standard feature data, we have also release our extracted feature data of CASIA-HWDB1.0–1.1 and CASIA-OLHWDB1.0–1.1 on our database webpage.[1]

[1] http://www.nlpr.ia.ac.cn/databases/handwriting/Download.html.

Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program) Grant 2012CB316302, the National Natural Science Foundation of China (NSFC) Grants 60933010 and 60825301, and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06030300).

References

[1] T.H. Hildebrandt, W. Liu, Optical recognition of Chinese characters: advances since 1980, Pattern Recognition 26 (2) (1993) 205–225.
[2] C.-L. Liu, S. Jaeger, M. Nakagawa, Online recognition of Chinese characters: the state-of-the-art, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2) (2004) 198–213.
[3] H. Fujisawa, Forty years of research in character and document recognition—an industrial perspective, Pattern Recognition 41 (8) (2008) 2446–2453.
[4] C.-L. Liu, Handwritten Chinese character recognition: effects of shape normalization and feature extraction, in: S. Jaeger, D. Doermann (Eds.), Arabic and Chinese Handwriting Recognition, Lecture Notes in Computer Science, vol. 4768, Springer, 2008, pp. 104–128.
[5] C.-L. Liu, F. Yin, Q.-F. Wang, D.-H. Wang, ICDAR 2011 Chinese handwriting recognition competition, in: Proceedings of the 11th ICDAR, Beijing, China, 2011, pp. 1464–1469.
[6] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, CASIA online and offline Chinese handwriting databases, in: Proceedings of the 11th ICDAR, Beijing, China, 2011, pp. 37–41.
[7] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, Chinese handwriting recognition contest 2010, in: Proceedings of the 2010 Chinese Conference on Pattern Recognition (CCPR), Chongqing, China, 2010.
[8] C.-L. Liu, K. Marukawa, Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition, Pattern Recognition 38 (12) (2005) 2242–2255.
[9] C.-L. Liu, K. Nakashima, H. Sako, H. Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, Pattern Recognition 37 (2) (2004) 265–279.
[10] C.-L. Liu, Normalization-cooperated gradient feature extraction for handwritten character recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (8) (2007) 1465–1469.
[11] C.-L. Liu, X.-D. Zhou, Online Japanese character recognition using trajectory-based normalization and direction feature extraction, in: Proceedings of the 10th IWFHR, 2006, pp. 217–222.
[12] K. Ding, G. Deng, L. Jin, An investigation of imaginary stroke technique for cursive online handwriting Chinese character recognition, in: Proceedings of the 10th ICDAR, Barcelona, Spain, 2009, pp. 531–535.
[13] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 9 (1) (1987) 149–153.
[14] X.-B. Jin, C.-L. Liu, X. Hou, Regularized margin-based conditional log-likelihood loss for prototype learning, Pattern Recognition 43 (7) (2010) 2428–2438.
[15] C.-L. Liu, R. Mine, M. Koga, Building compact classifier for large character set recognition using discriminative feature extraction, in: Proceedings of the 8th ICDAR, Seoul, Korea, 2005, pp. 846–850.
[16] C.-L. Liu, H. Sako, H. Fujisawa, Discriminative learning quadratic discriminant function for handwriting recognition, IEEE Transactions on Neural Networks 15 (2) (2004) 430–444.
[17] H. Zhang, J. Guo, G. Chen, C. Li, HCL2000—a large-scale handwritten Chinese character database for handwritten character recognition, in: Proceedings of the 10th ICDAR, Barcelona, Spain, 2009, pp. 286–290.
[18] H. Liu, X. Ding, Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes, in: Proceedings of the 8th ICDAR, 2005, pp. 19–23.
[19] K. Matsumoto, T. Fukushima, M. Nakagawa, Collection and analysis of on-line handwritten Japanese character patterns, in: Proceedings of the 6th ICDAR, 2001, pp. 496–500.
[20] L. Jin, Y. Gao, G. Liu, Y. Li, K. Ding, SCUT-COUCH2009—a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation, International Journal on Document Analysis and Recognition 14 (1) (2011) 53–64.
[21] C.-L. Liu, C.Y. Suen, A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters, Pattern Recognition, 42 (12) (2009) 3287–3295.
[22] J. Tsukumo, H. Tanaka, Classification of handprinted Chinese characters using non-linear normalization and correlation methods, in: Proceedings of the 9th ICPR, Rome, 1988, pp. 168–171.
[23] C.-L. Liu, H. Sako, H. Fujisawa, Handwritten Chinese character recognition: alternatives to nonlinear normalization, in: Proceedings of the 7th ICDAR, Edinburgh, Scotland, 2003, pp. 524–528.
[24] R.V.D. Heiden, F.C.A. Gren, The Box–Cox metric for nearest neighbor classification improvement, Pattern Recognition 30 (2) (1997) 273–279.
[25] T. Kohonen, Improved versions of learning vector quantization, in: Proceedings of the 1990 IJCNN, Washington, DC, vol. I, 1990, pp. 545–550.
[26] B.-H. Juang, W. Chou, C.-H. Lee, Minimum classification error rate methods for speech recognition, IEEE Transactions on Speech and Audio Processing 5 (3) (1997) 257–265.
[27] C.-L. Liu, High accuracy handwritten Chinese character recognition using quadratic classifiers with discriminative feature extraction, in: Proceedings of the 18th ICPR, Hong Kong, vol. 2, 2006, pp. 942–945.
[28] T.-F. Gao, C.-L. Liu, High accuracy handwritten Chinese character recognition using LDA-based compound distances, Pattern Recognition 42 (11) (2008) 3442–3451.
[29] K.C. Leung, C.H. Leung, Recognition of handwritten Chinese characters by critical region analysis, Pattern Recognition 43 (3) (2010) 949–961.
[30] T. Long, L. Jin, Building compact MQDF classifier for large character set recognition by subspace distribution sharing, Pattern Recognition 41 (9) (2008) 2916–2925.
[31] Y. Wang, Q. Huo, Building compact recognizers of handwritten Chinese characters using precision constrained Gaussian model, minimum classification error training and parameter compression, International Journal on Document Analysis and Recognition 14 (3) (2011) 255–262.
[32] Y. Wang, X. Ding, C. Liu, MQDF discriminative learning based offline handwritten Chinese character recognition, in: Proceedings of the 11th ICDAR, Beijing, China, 2011, pp. 1100–1104.

Cheng-Lin Liu is a professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences, Beijing, China, and is now the deputy director of the laboratory. He received the BS degree in electronic engineering from Wuhan University, Wuhan, China, the ME degree in electronic engineering from Beijing Polytechnic University, Beijing, China, the PhD degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, China, in 1989, 1992 and 1995, respectively. He was a postdoctoral fellow at Korea Advanced Institute of Science and Technology (KAIST) and later at Tokyo University of Agriculture and Technology from March 1996 to March 1999. From 1999 to 2004, he was a research staff member and later a senior researcher at the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan. His research interests include pattern recognition, image processing, neural networks, machine learning, and especially the applications to character recognition and document analysis. He has published over 140 technical papers at prestigious international journals and conferences.

Fei Yin is an assistant professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. He received the BS degree in computer science from Xidian University of Posts and Telecommunications, Xi'an, China, the ME degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology, Wuhan, China, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1999, 2002 and 2010, respectively. His research interests include document image analysis, handwritten character recognition and image processing. He has published over 20 papers at international journals and conferences.

Da-Han Wang received the BS degree in automation science and electrical engineering from Beihang University, Beijing, China, in 2006. He is currently pursuing a PhD degree in pattern recognition and intelligent systems at the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, handwriting recognition and retrieval, and probabilistic graphical models.

Qiu-Feng Wang received the BS degree in computer science from Nanjing University of Science and Technology, Nanjing, China, in 2006. He is currently pursuing a PhD degree in pattern recognition and intelligent systems at the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, handwritten text recognition, and language models.