IJDAR (2007) 10:27–38
DOI 10.1007/s10032-006-0037-6

ORIGINAL PAPER

Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text

Tonghua Su · Tianwen Zhang · Dejun Guan

Received: 1 July 2006 / Accepted: 27 November 2006 / Published online: 7 March 2007

Abstract A Chinese handwriting database named HIT-MW is presented to facilitate the offline Chinese handwritten text recognition. Both the writers and the texts for handcopying are carefully sampled with a systematic scheme. To collect naturally written handwriting, forms are distributed by postal mail or middleman instead of face to face. The current version of HIT-MW includes 853 forms and 186,444 characters that are produced under an unconstrained condition without preprinted character boxes. The statistics show that the database has an excellent representation of the real handwriting. Many new applications concerning real handwriting recognition can be supported by the database.

Keywords Standardization · Data acquisition · Optical character recognition · Handwritten Chinese text

1 Introduction

Offline recognition of Chinese handwritten character still remains one of the most challenging problems in pattern recognition domain after nearly three decades of research. Good results can be achieved by many algorithms as long as the Chinese characters are ideally written. However, when real-world characters are fed, the performance degrades greatly [29]. Usually, this kind of imperfection is attributed to characteristics of Chinese characters, e.g., complex structure, highly similar characters, great writing variations, and a large set of characters. Besides, one of the crucial factors is omitted: there are few real handwriting databases to fully explore the problem from different viewpoints.

In fact, standard databases play fundamental roles in handwriting recognition research. On the one hand, they provide a large number of training and testing data, resulting in high model fit and reliable confidence in statistic. On the other, they offer a means by which evaluation among different recognition algorithms can be performed. More and more handwriting researchers begin to pay much attention to the database standardization and evaluate their work using standard databases.

Dozens of handwriting databases have been released in literature for offline handwriting recognition. We tabulate some of them in Table 1. From the table, we can infer some insights. First, most databases have been published since 1990s. At the first year of this time, a Chinese handwritten character database entitled ITRI [31] was done which was hand-printed by 3,000 people in Taiwan. In 1992, CENPARMI [30] and PE92 [11] were reported. The former consists of unconstrained handwritten postcodes sampled from real mail pieces. The latter is a Korean character database written by 1,000 writers (an alternative Korean database is KU-1 [23]). Two years later, CEDAR [9] and CAMBRIDGE [26] were released. Similar to CENPARMI, CEDAR is also collected from real mail pieces. What's more, it includes a subset of handwritten city words extracting from mail

T. Su (✉) · T. Zhang
Harbin Institute of Technology, School of Computer Science and Technology, Harbin, China
e-mail: tonghuasu@hit.edu.cn

T. Zhang
e-mail: twzhang@hit.edu.cn

D. Guan
Harbin Engineering University, School of Computer Science and Technology, Harbin, China
e-mail: guandejun@hl.chinamobile.com

Table 1 Standard databases for offline handwriting recognition

Database	Language	Unit	Year	Source
Highleyman	English	Alphanum	1961	[8]
Munson		Alphanum	1968	[20]
Suen		Numeral	1972	[28]
CENPARMI		Postcode	1992	[30]
CEDAR		City name	1994	[9]
CAMBRIDGE		Sentence	1994	[26]
IAM		Sentence	1998	[16]
IAAS-4M	Chinese	Character	1985	[15]
ITRI		Character	1991	[31]
HCL2000		Character	2000	[36]
HK2002		Character	2002	[5]
ETL-8	Japanese	Character	1976	[19]
ETL-9		Character	1985	[24]
PE92	Korean	Character	1992	[11]
KU-1		Character	2000	[23]
IRONOFF	French	Character	1999	[32]
GRUHD	Greek	Character	2000	[10]
ISI	Indian	Alphanum	2005	[2]

addresses. CAMBRIDGE is the first handwritten English text database with a large vocabulary, which is written by a single writer in an unconstrained domain and used for writer-dependent handwriting recognition. Following that, the first version of IAM was put forward in 1998 [16], then the second version in 2002 [17], adapting some ideas from CAMBRIDGE. It is written by multiple writers and the texts for handcopying are progressively taken from the Lancaster-Oslo/Bergen (LOB) corpus. In 1999, a French handwritten database, IRONOFF [32], was released. While the characters are recorded in an online manner, they can be transformed into offline versions. In 2000, a hand-printed Chinese character database named HCL2000 [36] and a handwritten Greek database named GRUHD [10] were published. Writers in HCL2000 are asked to write a comprehensive set of the First Level Chinese characters of GB2312-80 [6] and the characters should be carefully written within a preprinted character box. GRUHD consists of two subsets. One includes hand-printed Greek characters and digits, the other an unconstrained Greek poem that can be used to conduct text-line segmentation experiments. More recently, a Chinese character database named HK2002 [5] and an Indian database named ISI [2], are published.

Second, English handwriting recognition is one of the most thoroughly studied branches not only in recognition strategy but also in database standardization. There are three different recognition strategies to English handwriting: segmentation-based recognition, segmentation-free recognition, and holistic recognition [3]. When arranging the English handwritten databases chronologically, we find that the handwritten unit has transmitted from digit or letter [8,20,28,30] to city name [9], further to sentence [26,16,17] and that application fields have expanded from small lexicon domains, such as bank check reading [7] and address recognition [35], to large lexicon and general unconstrained domains [12, 33,37].

Third, six offline databases are available for Chinese character recognition up to now, namely ETL-8/ETL-9 [19,24], IAAS-4M [15], ITRI, HCL2000 and HK2002 and all of them follow the same paradigm: each participant is requested to write a large set of Chinese characters, and each character should be carefully written within a preprinted character box. As a result, each character class contains the same number of samples, no matter whether it is rarely or frequently used in daily life. Meanwhile, samples in those databases are far from real-world ones, given that they are hand-printed within character boxes. In real-world applications, the input to handwriting reader is multiple lines of handwritten text even running up and down or with outliers (for instance, crossing off a character/word with special marks), instead of isolated characters. So, not only character recognition, but the text-line segmentation, outlier modeling and linguistic promotion are needed in real-world handwriting recognition. Moreover, since these databases are character-level, the recognition must be performed after character segmentation. Just as Sayre's paradox [25] goes, segmentation is prone to error and difficult to make correction afterward. Generally, much of the error rate can be attributed to imperfect segmentation. In addition, there are not enough data to support segmentation experiments, since the standard Chinese databases include only characters.

As a tradeoff, such experiments are conducted on Chinese mail addresses [14], though the number of them is limited. Indeed, a large handwritten Chinese text-level database is in great need.

We motivate to research the general purpose Chinese character recognition from segmentation-free perspective. After 3 years work, we compiled HIT-MW (HIT is the abbreviation of Harbin Institute of Technology, and MW means it is written by Multiple Writers), a handwritten Chinese text database for the first time. Comparing to CAMBRIDGE and IAM, our database has at least three distinctions. First, the handwriting is naturally written with no rulers that can be used to make the text-line straight by and large. This feature makes it suitable for conducting experiments on Chinese text-line segmentation. Second, the underlying texts for handcopying are sampled from China Daily corpus in a systematic way and the writers are carefully chosen to give a balanced distribution. Third, it is collected by mail or middleman instead of face to face, resulting in some real handwriting phenomena, such as miswriting and erasing. Besides text-line segmentation, the HIT-MW is fit to research segmentation-free recognition algorithms, to verify the effect of statistical language model (SLM) in real handwriting situation, and to study the nerve mechanism of Chinese handcopying activity.

This paper is an expanded version of [27]. Many new contents are incorporated here: (1) the thoughts underlying HIT-MW are fully stated for the first time; (2) a mechanism for text-line extraction is provided which makes HIT-MW ready for offline segmentation-free Chinese text recognition; (3) two important real handwriting phenomena, miswriting and erasing, are discussed which are special features of HIT-MW.

The flowchart of developing HIT-MW is illustrated in Fig. 1. The next section describes the sampling strategy. Then the handwriting collection and handwriting processing are discussed in Sect. 3 and 4, respectively. Section 5 first analyzes the basic statistics of the database to verify the effectiveness of our sampling strategy, and then presents two real handwriting phenomena, i.e., miswriting and erasing. Potential applications of our database are explored in Sect. 6. Finally, discussions and concluding remarks are given in Sect. 7.

2 Sampling strategy

Our database is to make a reasonable representation of the real Chinese handwriting, so it is important to carefully design sampling schemes. In this section, we describe two sampling schemes, dealing with objective writers and electronic data, respectively.

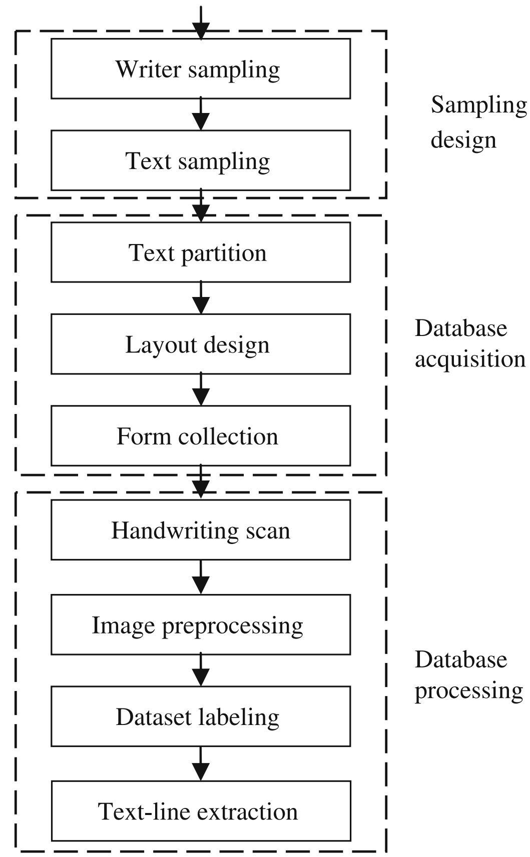

Fig. 1 Flowchart of HIT-MW database

2.1 Writer sampling

We determine our potential users to be college students, government clerk graduated from university, and senior students in high school who are potential college students in the next year. There are three reasons. First, according to the handwriting theory, the handwriting goes into a stable and consistent state at 25 years old, and after that there is little change. Second, the college students are enrolled throughout the country, so the handwriting by them can be seen as samples from the whole country. This diminishes the sampling bias to some degree. Third, it is mainly the well-educated people that are potential users of handwriting recognition in China, such as personal notes and manuscripts transcription.

Due to special users oriented, we need not sample the writers randomly. Instead, we divide the country into three regions, i.e., north region, middle region, and south region, and select one city handy from each region. Just using this simple sampling method, we obtain balanced writer samples (see Sect. 5 for more details).

2.2 Text sampling

We choose China Daily corpus as the data source of our database. In the natural language processing field, China Daily is extensively used as Chinese written language corpus, covering comprehensive topics such as politics, economics, science and technology, and culture. Using corpus as our data source instead of chaotic electronic texts demonstrates three advantages: linguistic context is automatically built in; Database can be easily expanded with tremendous texts to sample from; More frequently a character occurs, more training samples it possesses. Thereby, our database can be collected in a progressive way and is helpful to conduct the linguistic post-processing after the recognition stage.

We sample texts with a stratified random manner. To reserve more data for future expansion, we only use texts of the China Daily 2004 (news ranging from January to October is chosen at this stage). We first divide texts into ten groups according to month. Then we randomly draw 25 texts without replacement from each group. Using this method, we obtain a compact and sound approximation to Chinese written language (the verification is put aside in Sect. 5).

3 Database acquisition

As soon as the texts are extracted, it's time to start the collection process. Initially, we split each text into smaller and manageable segments. After several trials, we make each of them about 200 characters consisting of a few complete sentences. Next we format them into a clear and uniform layout. To design an informative layout, some considerations have been taken. Whenever all those have been done, we distribute forms to writers. Finally, we select forms according to special criteria.

3.1 Text partition

Texts previously sampled from corpus should be split into smaller segments. The number of characters in texts ranges from tens to thousands, which is inconvenient to distribute. In order to split each of them into a series of reasonable-size text segments, we consider the following two factors. First, it is wise to avoid breaking each complete sentence, in which as much linguistic context as possible can be held. Some punctuation marks—the period, the exclamation mark, the question mark, and combination of them with quotation marks—serve as sentence end. Others, such as the semicolon, the dash, and ellipsis mark can also be selected as sentence end if necessary.

Second, segment should have a reasonable number of characters. If it is too short, the writer's style and handwriting variability are hardly obtained. In the opposite case, it makes tired the writer's hand-muscle and vision-muscle, which in turn mostly makes the handwriting illegible. Moreover, we will not collect the handwriting completely when big-size characters are presented.

Based on these two factors, we conduct simulated experiments several times. It seems that segments between 50 and 400 characters are acceptable. The further discussion is presented in the next subsection.

3.2 Layout design

When we print text segments as forms, it is the layout that serves as an interface to writers. It is a nontrivial task to make it friendly and informative. The design of layout follows three criteria. First, the layout is simple and clear. Each form is divided into three distinct blocks: guideline block, text block, and writing block. The horizontal lines are used to separate the adjacent blocks and the faces of font to discern different information within block.

Second, we compress the writing guidelines to give more space reserved for handwriting. We make our commands concise by using short phrases and arrange them within five text lines with small font.

Third, we make use of implicit restrictions. In some cases, we want the writer to follow a special pattern, but it has difficulties to express in words. For example, we expect that the handwriting has a relatively small skew angle, but if we express it as a command, it will make the writer too careful to write naturally. Then we use horizontal lines both at top and bottom as references. It can help the writer know whether his handwriting is skew or not, and make some remedies reduce the skew adaptively. (In our opinion, totally freedom without any restrictions in handwriting collection is intractable).

After several recursions of feedback and modification, the final layout is illustrated in Fig. 2 (the writing block shown here is scaled down vertically to make the graph smaller). Each form is identified by a 4-pair-digit code and each pair stands for certain meaning, e.g., 04090902 means that it is the second text segment of the ninth text sampled from September 2004.

3.3 Form collection

Forms are distributed by mail or middleman instead of face to face. This makes the writers impossible to tailor the handwriting for easy recognition, not exactly knowing what their handwriting will be used. Naturally written handwriting is more likely to acquire.

样本编号：04090902　　此手写样本授予哈尔滨工业大学人工智能实验室研究之用。
性别 男□女□　年龄　　职业　　签名
书写要求：
保持纸张勿折、正反面均清洁，规范书写、勿潦草，蓝、黑或蓝黑色笔均可，尽量少连笔少涂污，行间留出空隙。
请抄写下面的印刷文本到空白区内，谢谢您的合作！

郑培民生前是中共湖南省委副书记、湖南省人大常委会副主任，2002 年 3 月 11 日因心脏病突发，牺牲在工作岗位上。2003 年 3 月 11 日，中共中央总书记胡锦涛作出重要批示，号召向郑培民同志学习。2004 年 3 月，潇湘电影集团、中国电影集团和大成公司投拍影片《郑培民》，并于国庆前夕奉献给全国观众。

该片取材于郑培民同志生前的生活小事，以修建公路为主线，集中体现了他权为民所用、情为民所系、利为民所谋的情怀，成功地塑造了一个党的好干部的典型形象。

Fig. 2 An illustration of layout

Once a pile of handwriting forms are collected, we accept the legible ones, and the illegible or lost ones are reprinted and distributed again. Handwriting is thought as legible, if it runs from left to right, its contents are what we have appointed (a little miswriting and erasing are allowed), and a majority of it can be read correctly by human.

4 Database processing

The accepted handwriting is scanned into computer as digital image and then pixel-level processing is applied on it. The processing includes frame eliminating and binarization to give a clean and compact registration of the handwriting. Next, we transcribe the handwriting's ground truth that will serve as standard answers when calculating the recognition rate. Eventually, the database is ready for segmentation-free recognition by separating the text lines.

4.1 Handwriting digitalization

Each writing block of legible forms is scanned into computer by Microtek ScanMaker 4180. The resolution is set to 300dpi. Images are saved as gray-scale BMP files with no compression and named after their forms' code. The average storage space of each image is about 2.1M bytes.

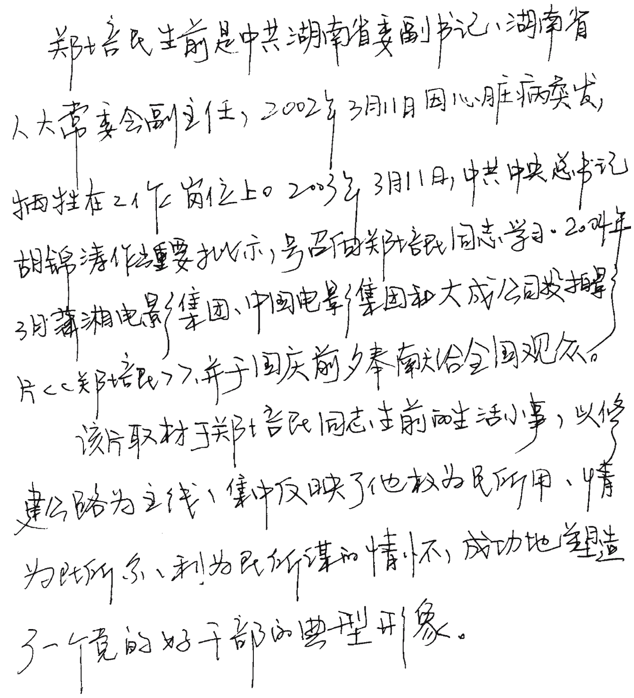

Fig. 3 Binary image of handwriting sample named 04090902

4.2 Image preprocessing

We perform image preprocessing on each scanned image. First, we eliminate the frame lines enclosing the writing block. We deal with them in an automatic way, and manually eliminate them once the lines are off standard positions. We pay special attention to preserving the smoothness of its strokes intersecting the frame lines.

Then, we binarize handwriting image using Otsu algorithm [22]. The binary image is named after the gray-scale image and a letter "b" is inserted as the prefix. The black–white version of the handwriting image named 04090902 is shown in Fig. 3.

4.3 Database labeling

The ground truth acts as the standard answers to the handwriting image. To evaluate the performance, transcription from recognition engine is compared with the ground truth. That is to say, labeling the database to generate its ground truth is the preliminary stage for the development of the recognition system.

Generating the ground truth file involves two different level alignments: a text-line level alignment and a character level alignment. The former makes text segment produce a new line where corresponds to the end of each handwriting text line. The latter crosses off the deleted characters from each segment, key in the inserted characters and modify the substituted characters. An

郑培民生前是中共湖南省委副书记、湖南省
人大常委会副主任，2002 年 3 月 11 日因心脏病突发，
牺牲在工作岗位上。2003 年 3 月 11 日，中共中央总书记
胡锦涛作出重要批示，号召向郑培民同志学习。2004 年
3 月，潇湘电影集团、中国电影集团和大成公司投拍影
片《郑培民》，并于国庆前夕奉献给全国观众。
该片取材于郑培民同志生前的生活小事，以修
建公路为主线，集中反映了他权为民所用、情
为民所系、利为民所谋的情怀，成功地塑造
了一个党的好干部的典型形象。

Fig. 4 The ground truth on document level of Fig. 3

example of the labeled ground truth on document level is illustrated in Fig. 4. Further, each row of the text is extracted and saved as a separate file (ground truth on text-line level).

Note that, we don't label the ground truth character by character. This is determined by our research goal. Our recognition engine follows a segmentation-free strategy, that is, there is no character segmentation stage in our system and the output of recognition engine is a string of Chinese characters which are transcriptions of (at least) one textline. By comparing the transcription with the corresponding ground truth, the recognition rate can be calculated. As a result, labeling each character's location is needless.

4.4 Text-line extraction

Many skewed handwritten documents even with strokes touching and overlapping between adjacent text lines are presented in HIT-MW database (as shown in Fig. 3). We have developed a fast skew correction algorithm to improve the recall rate of text lines. We employ the angular histogram of the horizontal strokes.

In GB2312-80, the national encoding standard used in China, each Chinese character consists of about 15.17 basic strokes and horizontal strokes account for 39.51 % of them [34]. In other words, there are averagely six horizontal strokes in a Chinese character. Those statistics mean that horizontal strokes are stable features in Chinese characters.

We calculate the horizontal runlength of each document and only keep the strokes whose runlength greater than T_s, where T_s is a threshold relating to the average stroke width. The kept strokes of Fig. 3 are shown in Fig. 5. We can see that the long downward strokes are stripped out. Further, we select one representative point for each remained stroke, as in Fig. 6.

The skew detection can be modeled as a classification problem and the skew angle of a document can be identified as the class maximizing a cost function as follows:

$$\theta' = \arg\max_{\theta} \Omega_\theta. \tag{1}$$

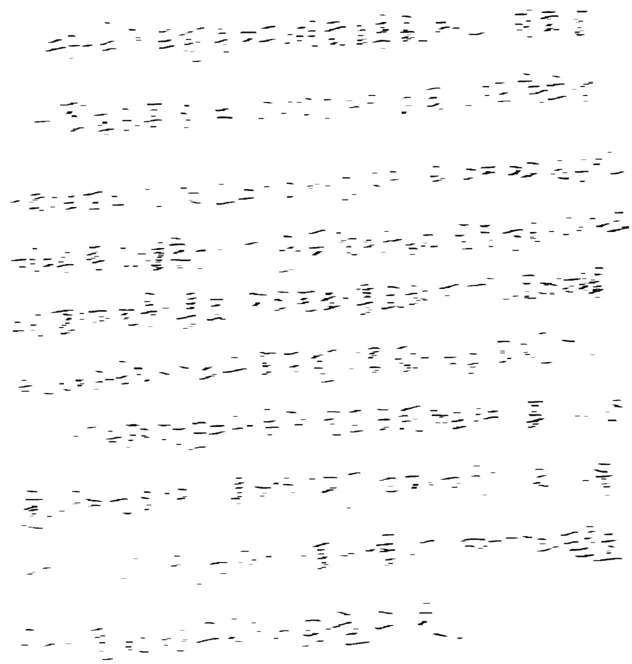

Fig. 5 The horizontal pseudo-stroke map of Fig. 3

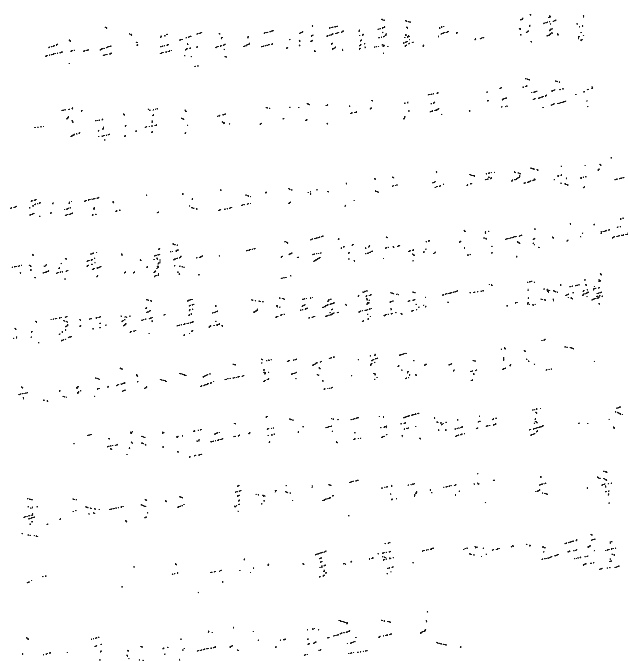

Fig. 6 The representative points corresponding to Fig. 3. Each point is expanded nine times to give a clear view

The cost function defines as a weighted sum of three terms, has the following form:

$$\Omega_\theta = a_1\sigma_\theta + a_2\phi_\theta + a_3\varpi_\theta, \tag{2}$$

where σ_θ refers to the variance of the horizontal projection histogram, ϕ_θ the total gaps (the number of

Table 2 The number of recalled text lines (recall rate) in two setups

No skew correction	With skew correction
4414 (56.47%)	5394 (69.01%)

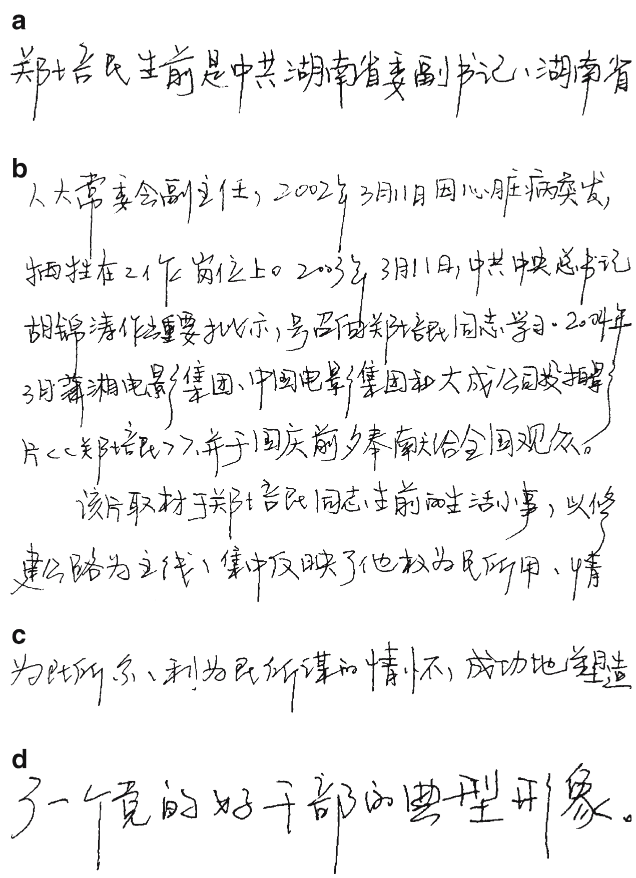

Fig. 7 The segmentation result by global projection after skew correction **a** The first text line is extracted successfully, **b** The second part is failed to segment, **c** The second to last text line is extracted successfully, **d** The last text line is extracted successfully

positions with zero histogram value) divided by document height, and ϖ_θ the normalized maximum of the histogram. The weights, a_is' are determined from experiments.

After the skew correction, the recall rate of the text lines by global horizontal projection is improved by 12.7%, as indicated in Table 2. Three out of ten text lines of the handwriting in Fig. 3 can be successfully recalled by global projection following the skew correction (as in Fig. 7a, c, d).

In order to handle the complex text lines (just as in Fig. 7b), we currently adapt the genetic algorithm based HIDER method (an improved version to [1])to find the failure blocks (those can not successfully separated by partial projection) and then heuristic based thinning algorithm (similar to [13]) will be used to extract the text border lines.

Table 3 Lexicon of HIT-MW database vs GB2312-80 character set (unit: characters)

Within GBset				
flGBset	slGBset	ASCII	others	Beyond GBset
2,746	215	48	27	5

5 Database statistics

The HIT-MW database is the first collection of Chinese handwritten texts in handwriting recognition domain. More than 780 participants produce their handwriting naturally. In this section we will present HIT-MW's features by a data-driven way. First, we describe the basic statistics, which show sound writer distributions and an appealing lexical coverage on the China Daily corpus. Next, we focus our attention on two key handwriting phenomena, i.e., miswriting and erasing, and analyze them, respectively.

5.1 Basic statistics

We have collected 853 legible Chinese handwriting samples. There are 186,444 characters in total including letters, punctuations besides Chinese characters, and these characters lead to 8,664 text lines. By simple computation, we get following statistics: Each sample has 10.16 text lines; each text line has 21.51 characters; each sample includes 218.57 characters.

Mining the ground truth files of our database, we derive following results. The lexicon of the database has 3,041 entries. In other words, each character averagely occurs 61.31 times. Most of the entries fall into GB2312-80 character set (hereafter, abbreviated as GBset), and details are summarized in Table 3. Chinese handwritten character databases (such as HCL2000, IAAS-4M) only consist of the first level Chinese characters of GBset (flGBset in short, and similarly slGBset for the second level Chinese characters of GBset). Unlike them, our database samples characters by their real use in daily life. As a result, not only most of flGBset but a quantity of slGBset are included (even several characters beyond GBset are included).

Moreover, to check its representative capability, we plot its coverage over China Daily corpus with 79,509,778 characters in Fig. 8. Note that, the corpus has already excluded the data of China Daily 2004 to give

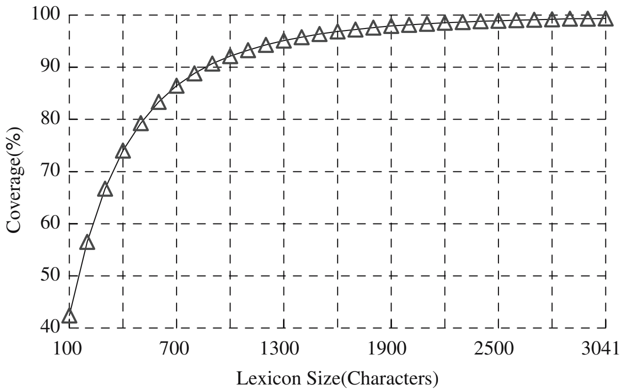

Fig. 8 Lexicon size of HIT-MW versus coverage of China Daily corpus

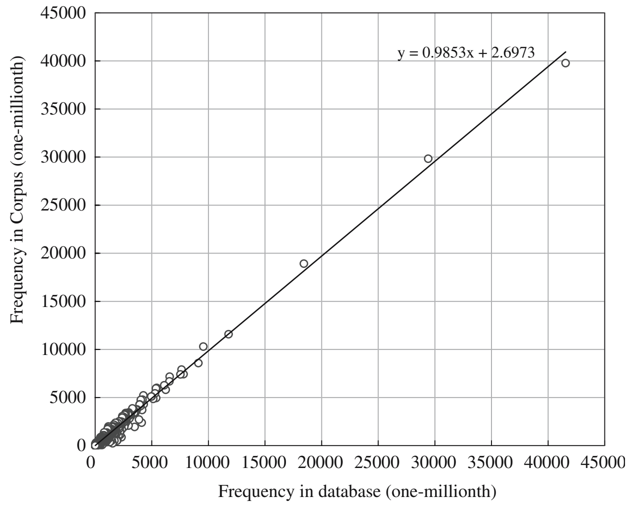

Fig. 9 The scatter map of the character frequency occurred in HIT-MW and China Daily corpus

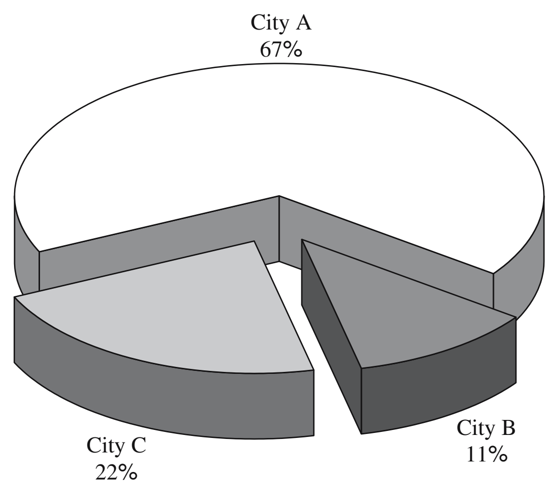

Fig. 10 Sampling percentage of three regions

objective coverage estimation. From the graph, we can see that a 1,800 character lexicon covers 97.60% of the corpus, and the full-size lexicon 99.33% of the corpus. The lexicon is extracted from the database according to the character frequency. For example, a 100 character lexicon consists of 100 most frequently occurred characters in the database. In another way, we plot the scatter map in Fig. 9 between lexicon of database and that of corpus. Each dot in the figure, (x, y), means that a character appears x times in database and y times in corpus. We can see from the figure that the cloud of dots is mainly spread along the auxiliary diagonal. Minimizing the least squares, we obtain a regression line as follow:

$$y = 0.9853x + 2.6973 \tag{3}$$

We can see that the x value is approximately the same of y. By correlation analysis, we get a high coefficient of 0.9936 (the number of dots is 3,037).

Further, we calculate the writer's distribution. We mark the three sampled cities as City A, City B, and City C, respectively. From the view of city distribution in Fig. 10, the sampled writers are mainly from City A with a proportion of 67%. Seen from Table 4, the department distribution of writers is near to that calculated from real data of college students of 2004 [21]. Similarly, Table 5 shows that the sex distribution of our database has a good coincidence with that calculated from real educational statistics of 1998 [18].

In summary, both the distribution of writer and the coverage of lexical entry show the effectiveness of the proposed sampling schemes.

5.2 Erasing statistics

The handcopying activity relates three interactive processes. Initially, the vision perceives the stimuli and transmits them as signals to the brain. Then, the brain stores the information in memory. And as the last step, the brain makes certain muscles active and further those muscles drive the writing instrument to run on the paper. Errors in any process will result in erasing or miswriting. For example, there is an erasure marked by a black dot at the eighth character of the last line in Fig. 11.

We group erasures by erasing mark and month in Table 6. In real handwriting, writers use erasing marks to express the marked character is discarded. To different persons, their marks may vary in some way, for example, writer A may use a double slash (//), while writer B uses a black dot (•). From Table 6, we can learn some points. First, erasures are common in our database. There are 382 instances of erasure totally, and about one instance appears out of every two handwriting samples. If we do not model it properly in recognition

Table 4 Writers from science and engineering departments versus college students of 2004 from that

Of 2004	Sampled
61.37%	60.69%

Table 5 Gender distribution comparison between writers sampled and students of year 1998

Boy writers sampled		Boy students of 1998	
High school	University	High school	University
57.25%	62.54%	57.26%	63.29%

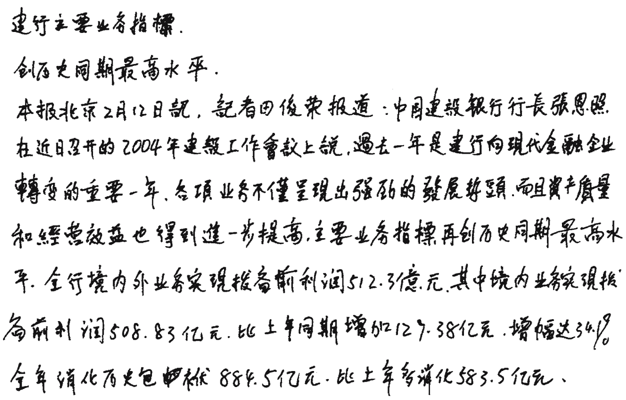

Fig. 11 A piece of handwriting with an erasure

stage, it may decrease the recognition rate by 0.20% solely. Moreover, when SLM is used as postprocessing, it may make things worse. As an extreme, if the recognition is based on segmentation-free strategy, about 4.39% of the characters will be under threat.

Second, analyzing the occurrences in each month, we can also infer that erasures are stable phenomena. Averagely, there are 38–39 occurrences per month with a concentrated derivation.

Third, the erasing marks show high possibility to be modeled by clustering them. There are 12 types of marks, however, the most commonly used ones are mainly fall in 4 types and the sum of them makes 88% of all.

In summary, erasing is a common and natural phenomenon stemming from real handwriting, and we should properly model them in order to acquire a sound recognition performance. It is good news that the erasing marks manifest an excellent grouping possibility and that gives a promise for erasure modeling.

5.3 Miswriting statistics

Miswriting in handwriting means what have been written are different from the appointed ones. It can be classified into three types: deletion, insertion, and substitution. Miswriting may hurt the linguistic context. However, it may not necessarily do that, and in some settings it even facilitates the context. For example, miswriting "建行工作" (in PinYin: jian-hang-gong-zuo) as "建设工作" (in PinYin: jian-she-gong-zuo) will improve the performance in tri-gram environment (see Fig. 12 to get a illustration).

We calculate the miswriting occurrences excluding punctuation, since there present no punctuation in some applications, for example, automatic document image summarizing. At this stage, we integrate the decisions from three local language holders to determine whether the miswriting hurt the linguistic context or not. The term "context" here refers to two characters before and after the miswriting block. The result is summarized in Table 7. From Table 7a, we can see that the deleted characters are the most frequently occurred among the three classes. This fact leads us to infer: In handcopying activity, it's easier to miss characters than other miswriting cases. In Table 7b, there are 824 miswriting blocks totally, however, only 274 out of them hurt linguistic context.

Such imperfect situation has never happened yet in optical character recognition (OCR) history, since all of the recognition algorithms are evaluated in ideal handwriting environment. Whether we should use SLM or not will not be as obvious as before. Supposing the recognition rate without SLM is 65%, 80% after SLM, and there is no rejection. It's interesting to see that the role of SLM is mainly determined by the degree of context hurting. If the recognition rate of hurting portion is larger than 35%, SLM will be an essential stage; otherwise, there is no simple answer.

If we further analysis the substituted blocks, we may infer some tips concerning nerve mechanism of Chinese handcopying [4] which is out of the scope of this paper.

6 Application of HIT-MW database

Our database can support experiments in a more real aspect than character-level database. At least but not limited to following four research directions can be emerged. Most of them are rarely or never explored yet.

Real text-line segmentation Each piece of handwriting in our database is produced naturally by participant with no rules, resulting in a great number of real text lines. As expressed in Subsect. 4.4, using global projection method directly, only 56.47% of them can be correctly separated. The failure lies in irregular text lines. As soon as single text line is concerned, irregularity

Table 6 Statistics on erasure

Mark	January	February	March	April	May	June	July	August	September	October	*Total*
\	19	20	7	27	3	13	6	15	15	8	133
\\	12	12	9	11	5	3	11	21	10	11	105
•	4	7	12	7	10	5	10	3	1	4	63
\ \ \	8	5	3	6	1	6	–	–	2	3	34
≡	–	5	1	–	5	1	–	1	–	3	16
=	1	7	–	–	–	–	–	2	–	3	13
∘	2	1	1	–	1	2	1	1	2	–	11
– –	–	–	–	1	–	–	–	–	–	2	3
//	1	–	–	–	–	–	–	–	–	–	1
×	–	–	–	1	–	–	–	–	–	–	1
()	–	–	–	1	–	–	–	–	–	–	1
/	–	–	–	–	–	–	–	–	–	–	1
Total	47	57	33	54	25	30	28	43	30	35	382

$$\begin{aligned}
&\frac{p(\text{04年建设工作})}{p(\text{04年建行工作})}\\
&\approx \frac{p(\text{设}\mid\text{建年})p(\text{工}\mid\text{设建})p(\text{作}\mid\text{工设})}{p(\text{行}\mid\text{建年})p(\text{工}\mid\text{行建})p(\text{作}\mid\text{工行})}\\
&\approx \frac{C(\text{年建设})C(\text{建设工})C(\text{设工作})C(\text{建行})C(\text{行工})}{C(\text{年建行})C(\text{建行工})C(\text{行工作})C(\text{建设})C(\text{设工})}\\
&= \frac{156\times1885\times635\times561\times1354}{6\times3\times1146\times109621\times1905} = 32.93 >> 1
\end{aligned}$$

Fig. 12 An example when miswriting facilitating the context

mainly comes from skew line or undulate line. When considering adjacent text lines, there exist overlapping lines and touching ones. So, HIT-MW can be used to develop fine text-line segmentation algorithms.

Real and general handwriting recognition Our database is produced with linguistic context and it is sampled from natural handwriting. Besides hand-printed characters, slant and cursive ones are of great quantity. In addition, erasures are presented. As manifested in Subsect. 5.2, without modeling them, the recognition rate will suffer a bit. In this complex environment, more advanced techniques are needed.

SLM in real situation As we known, SLM is essential for general domain recognition. However, in our database, whether we should use SLM or not is not as clear as before due to the miswriting and outlier (such as erasures). In addition, how to efficiently incorporate the SLM into the handwritten text recognition framework raises a new problem.

Segmentation-free recognition Current Chinese character recognition algorithms are all segmentation-based. As mentioned in Sect. 1, character recognition is a prone-to-error step. Unlikely, segmentation-free recognition deals with segmentation and recognition together and good optimal results may be gained easily. There are good reasons to explore the Chinese handwriting recognition from segmentation-free strategy. HIT-MW database provides such possibility.

7 Discussion and conclusion

HIT-MW database inherits data sparseness from natural language, since texts are sampled from corpus. Character frequency of database is shown in Table 8. We can see that only a small portion of characters occur frequently. For example, only 1,853 ones out of 3,041 characters occur more than five times. This phenomenon can save our time and resource by pouring most efforts on most frequently used characters. However, as soon as the seldom-occurred characters concerned, there are too small number of samples for training. To overcome the data sparseness of our database and obtain complete flGBset, we can incorporate character-level databases (such as IAAS-4M, HCL2000) into our database.

The handwritten Chinese text database discussed in this paper addresses several important aspects not covered by most other databases. It is naturally written by multiple writers, hence, there are real text lines and real handwriting phenomena. In addition, not only texts are well sampled, but also writers are carefully determined, resulting in a sound sampling of Chinese handwriting.

The original purpose of HIT-MW database is to facilitate the fundamental study on offline Chinese handwriting recognition from a brand new perspective. Many new research directions can be emerged, such as real text-line segmentation, real and general handwriting recognition, SLM in real situation, segmentation-free

Table 7 Statistics of miswriting

(a) Miswriting characters (unit: character)			
Total	Deletion	Insertion	Substitution
2,280	1,884	110	274
(b) Miswriting blocks and its linguistic effect (unit: block)			
Total	Hurting context	Favoring context	
824	274	281	

Table 8 Character frequency of HIT-MW database

Occurrences	*Characters*
≥1,000	16
≥100	456
≥10	1,469
≥5	1,853

recognition. Study on them may promote the real-world Chinese handwriting recognition greatly.

The database and the latest details are available at http://hitmwdb.googlepages.com/. They can be downloaded freely. In addition, the ground truth and the gray-scale version of the database are also available upon request (Please contact hitmwdb@gmail.com).

Acknowledgments We would like to thank Yiping Deng, Hui Xia, Ling Song, Di Zhang, Xuecai Yu, Haidan Xie, Yufeng Sun, Cuan Su and Guangjin Shao for their collaborations. We would also like to thank Haijing Wang and Yu Zhou for their valuable suggestions. This work is supported by the National Natural Science Foundation of China (No. 60475011) and the Natural Science Foundation of Heilongjiang Province (No. F0322).

References

1. Aguilar-Ruiz, J.S., Riquelme, J.C., Toro, M.: Evolutionary learning of hierarchical decision rules. IEEE Trans. Syst. Man Cybern. B. **33**(2), 324–331 (2003)
2. Bhattacharya, U., Chaudhuri, B.B.: Databases for research on recognition of handwritten characters of Indian scripts. In: The 8th International Conference on Document Analysis and Recognition, Seoul, pp. 789–793 (2005)
3. Casey, R.G., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(7), 690–706 (1996)
4. Fu, S., Chen, Y., Smith, S., Iversen, S., Matthews, P.M.: Effects of word form on brain processing of written Chinese. Neuroimage **17**(3), 1538–1548 (2002)
5. Ge, Y., Huo, Q.: A comparative study of several modeling approaches for large vocabulary offline recognition of handwritten Chinese characters. In: The 16th International Conference on Pattern Recognition, Quebec, pp. 85–88 (2002)
6. General Administration of Technology of the People's Republic of China: Code of Chinese Graphic Character Set for Information Interchange—Primary Set. Standard Press of China, Beijing (1980) (in Chinese)
7. Guillevic, D., Suen, C.Y.: Recognition of legal amounts on bank cheques. Pattern Anal. Appl. **1**(1), 28–41 (1998)
8. Highleyman, W.: An analog method for character recognition. IRE Trans. Electron. Comput. EC **10**, 502–512 (1961)
9. Hull, J.: A database for handwritten text recognition research. IEEE Trans. Pattern Anal. Mach. Intell. **16**(5), 550–554 (1994)
10. Kavallieratou, E., Liolios, N., Koutsogeorgos, E., Fakotakis, N., Kokkinakis, G.: The GRUHD database of Greek unconstrained handwriting. In: The 6th International Conference on Document Analysis and Recognition, Seattle, pp. 561–565 (2001)
11. Kim, D.-H., Hwang, Y.-S., Park, S.-T., Kim, E.-J., S.-H, P., Bang, S.-Y.: Handwritten Korean character image database PE92. IEICE Trans. Inf. Syst. E **79-D**(7), 943–950 (1996)
12. Kim, G., Govindaraju, V., Srihari, S.N.: An architecture for handwritten text recognition systems. Int. J. Doc. Anal. Recognit. **2**(1), 37–44 (1999)
13. Liang, Z., Shi, P.: A metasynthetic approach for segmenting handwritten Chinese character strings. Pattern Recognit. Lett. **26**, 1498–1511 (2005)
14. Liu, C.-L., Koga, M., Fujisawa, H.: Lexicon–driven segmentation and recognition of handwritten character strings for Japanese address reading. IEEE Trans. Pattern Anal. Mach. Intell. **24**(11), 1425–1437 (2002)
15. Liu, Y.J., Tai, J.W., Liu, J.: An introduction to the 4 million handwriting Chinese character samples library. In: Proceedings of the International Conference on Chinese Computing and Orient Language Processing, Changsha, pp. 94–97 (1989)
16. Marti, U.V., Bunke, H.: A full English sentence database for off-line handwriting recognition. In: The 5th International Conference on Document Analysis and Recognition, Bangalore, pp. 705–708 (1999)
17. Marti, U., Bunke, H.: The IAM-database: an English sentence database for off-line handwriting recognition. Int. J. Doc. Anal. Recognit. **5**(1), 39–46 (2002)
18. Ministry of Education: Educational Statistics Yearbook of 1998. People's Education Press, Beijing (1998) (in Chinese)
19. Mori, S., Yamamoto, K., Yamada, H., Saito, T.: On a handprinted kyoiku-kanji character data base. Bull. Electrotech. Lab. **43**(11–12), 752–773 (1979)
20. Munson, J.H.: Experiments in the recognition of handprinted text: Part I-character recognition. In: Proceedings of Fall Joint Computer Conference. Thompson Books, Washington, DC, December 1968 pp. 1125–1138 (1968)
21. National Bureau of Statistics of China: China Statistical Yearbook 2004. China Statistics Press, Beijing (2005)
22. Otsu, N.: A threshold selection method from gray-level histogram. IEEE Trans. Syst. Man Cybern. SMC **9**(1), 62–66 (1979)
23. Park, J.S., Kang, H.J., Lee, S.W.: Automatic quality measurement of gray-scale handwriting based on extended average entropy. In: The 15th International Conference on Pattern Recognition, Barcelona, pp. 426–429 (2000)
24. Saito, T., Yamada, H., Yamamoto, K.: On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis. IEICE Trans. J. **68**(D(4), 757–764 (1985)
25. Sayre, K.: Machine recognition of handwritten words: a project report. Pattern Recognit. **5**(3), 213–228 (1973)

26. Senior, A.W., Robinson, A.J.: An off-line cursive handwriting recognition system. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 309–321 (1998)
27. Su, T., Zhang, T., Guan, D.: HIT–MW dataset for offline Chinese handwritten text recognition. In: The 10th International Workshop on Frontiers in Handwriting Recognition. (2006)
28. Suen, C.Y., Berthod, M., Mori, S.: Automatic recognition of handprinted characters—the state of the art. Proc. IEEE **68**(4), 469–487 (1980)
29. Suen, C.Y., Mori, S., Kim, S.H., Leung, C.H.: Analysis and recognition of Asian scripts—the state of the art. In: The 7th International Conference on Document Analysis and Recognition, Edinburgh, pp. 866–878 (2003)
30. Suen, C.Y., Nadal, C., Legault, R., Mai, T.A., Lam, L.: Computer recognition of unconstrained handwritten numerals. Proc. IEEE **80**(7), 1162–1180 (1992)
31. Tang, Y.Y., Tu, L.-T., Liu, J., Lee, S.-W., Lin, W.-W.: Offline recognition of Chinese handwriting by multifeature and-multilevel classification. IEEE Trans. Pattern Anal. Mach. Intell. **20**(5), 556–561 (1998)
32. Viard-Gaudin, C., Lallican, P.M., Knerr, S., Binter, P.: The IRESTE on/off (IRONOFF) dual handwriting database. In: The 5th International Conference on Document Analysis and Recognition, Bangalore, pp. 455–458 (1999)
33. Vinciarelli, A., Bengio, S., Bunke, H.: Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 709–720 (2004)
34. Wu, Y., Ding, X.: Character Recongition—Theory, Method and Implementation. Higher Education Press, Beijing (1992) (in Chinese)
35. Yacoubi, M.E., Gilloux, M., Bertille, J.M.: A statistical approach for phrase location and recognition within a text line: an application to street name recognition. IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 172–188 (2002)
36. Zhang, H., Guo, J.: Introduction to HCL2000 database. In: Proceedings of Sino-Japan Symposium on Intelligent Information Networks, Beijing (2000)
37. Zimmermann, M., Bunke, H.: N–gram language models for offline handwritten text recognition. In: The 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, pp. 203–208 (2004)