

Chapter 2

HIT-MW Database

Abstract Standard databases play a fundamental part in handwriting recognition research. This chapter presents a Chinese handwriting database named HIT-MW, designed to facilitate Chinese handwritten text recognition. Both the writers and the texts for handcopying are carefully sampled using a systematic approach. To collect naturally written handwriting, the forms were distributed by postal mail or middleman instead of face to face. The current version of HIT-MW includes 853 forms and 1,86,444 characters that were produced under natural and unconstrained conditions without preprinted character boxes. The statistics show that the database provides an excellent representation of the realistic Chinese handwriting. Many new applications concerning realistic handwriting recognition can be supported by the database. Hundreds of institutes and universities have begun using the HIT-MW database in their experiments over the world.

2.1 Introduction

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 hand, 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 were released in literature for handwriting recognition before 2006. We tabulate some of them in Table 2.1. From the table, we just highlight the following two insights. First, 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

Table 2.1 Standard databases for handwriting recognition

Database	Language	Scope	Years	Source
Highleyman	English	Alphanum	1961	Highleyman (1961)
Munson		Alphanum	1968	Munson (1968)
Suen		Numerical	1972	Suen et al. (1980)
CENPARMI		Postcode	1992	Suen et al. (1992)
CEDAR		City name	1994	Hull (1994)
CAMBRIDGE		Sentence	1994	Senior and Robinson (1998)
UNIPEN		Alphanum, Word, Sentence	1994	Guyon et al. (1994)
IAM	Chinese	Sentence	1998	v1 (Marti and Bunke 1999)
				v2 (Marti and Bunke 2002)
IAAS-4M		Character	1985	Liu et al. (1989)
THOCR-HCD		Character	~1985	Fu et al. (2008)
ITRI		Character	1991	Tu et al. (1991)
SCUT-IRAC		Character	1996	Jin (1996)
HCL2000		Character	2000	Zhang and Guo (2000)
HK2002		Character	2002	Ge et al. (2002)
ETL-8		Japanese	1976	Mori et al. (1979)
ETL-9			1985	Saito et al. (1985)
TUAT			1997	Nakagawa et al. (1997)
PE92		Korean	1992	Kim et al. (1996)
KU-1			2000	Park et al. (2000)
IRONOFF	French	Character	1999	Viard-Gaudin et al. (1999)
GRUHD	Greek	Character	2000	Kavallieratou et al. (2001)
ISI	Indian	Alphanum	2005	Bhattacharya and Chaudhuri (2005)

English handwriting: segmentation-based recognition, segmentation-free recognition, and holistic recognition (Casey and Lecolinet 1996). When arranging the English handwritten databases chronologically, we find that the handwritten scope has transmitted from digit or letter (Highleyman 1961; Munson 1968; Suen et al. 1980, 1992) to city name (Hull 1994) or words (Guyon et al. 1994), further to sentence (Senior and Robinson 1998; Marti and Bunke 1999, 2002; Guyon et al. 1994) and that application fields have expanded from small lexicon domains, such as bank check reading (Guillevic and Suen 1998) and address recognition (Yacoubi et al. 2002), to large lexicon and general unconstrained domains (Kim et al. 1999; Vinciarelli et al. 2004; Zimmermann and Bunke 2004).

Second, nine databases were mentioned in literature for Chinese or Kanji character recognition before 2006, namely ETL-8/ETL-9 (Mori et al. 1979; Saito et al. 1985), IAAS-4M (Liu et al. 1989), THOCR-HCD (Fu et al. 2008), ITRI (Tu et al. 1991), SCUT-IRAC (Jin 1996), TUAT (Nakagawa et al. 1997), HCL2000 (Zhang and Guo 2000), and HK2002 (Ge et al. 2002) and all of them follow the same paradigm: each participant is requested to write a large set of Chinese characters [commonly 3,755 characters in the First Level GB2312-1980 (General Administration of Technology of the People's Republic of China 1980)],

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 handprinted 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 constraint are needed in real-world handwriting recognition. Moreover, since these databases are character-level, the recognition must be performed after explicit character segmentation. Just as Sayre's paradox (Sayre 1973) 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 (Liu et al. 2002), though the number of them is limited. Indeed, a large handwritten Chinese text-level database is in great need.

We have motivated to study the general purpose Chinese handwriting recognition. 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 People's 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.

With the HIT-MW database, the recognition scope has shifted to the unconstrained Chinese handwriting. Chinese researchers can set out to overcome the barrier associated with the practical reading machine for Chinese handwriting. Inspired by HIT-MW database, another three Chinese handwriting databases are built recently: SCUT-COUCH (Jin et al. 2011), HIT-OR3C (Zhou et al. 2010) and CASIA-OLHDB/CASIA-HWDB (Liu et al. 2011). Among them, HIT-MW database and CASIA-HWDB can facilitate the off-line tasks while others can support the online tasks. Provided with these training samples, sophisticated learning algorithms can be investigated.

This chapter presents the overall process for creation of HIT-MW database. The flowchart of developing HIT-MW is illustrated in Fig. 2.1. The next section describes the sampling strategy. Then the handwriting collection and handwriting

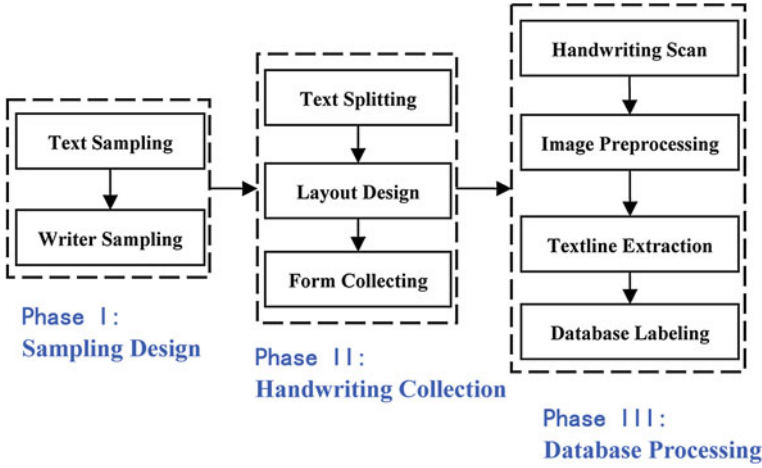


Fig. 2.1 Flowchart of HIT-MW database

processing are discussed in [Sects. 2.3](#) and [2.4](#), respectively. [Section 2.5](#) 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. 2.6](#). Finally, discussions and concluding remarks are given in [Sect. 2.7](#).

2.2 Sampling Strategy

Our database is to make a reasonable representation of the realistic 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.

2.2.1 Writer Sampling

We determine our potential users to be college students, government clerk graduated from university, and senior students in high school who will be 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 who 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 (Cf. [Sect. 2.5](#)).

2.2.2 Text Sampling

We choose People's Daily corpus as the data source of our database. In the natural language processing field, People's 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 People's 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. 2.5](#)).

2.3 Database Acquisition

Once the texts are extracted, we 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.

2.3.1 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 the writer's hand-muscle and vision muscle tired, 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. Further discussion is presented in the next subsection.

2.3.2 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 to 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. 2.2 (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.

样本编号：04090902 此手写样本授予哈尔滨工业大学文字识别研究之用

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，感谢您的合作！

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. 2.2 An illustration of layout

2.3.3 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 for what their handwriting will be used. Naturally written handwriting is more likely to acquire.

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.

2.4 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.

2.4.1 Handwriting Digitalization

Each writing block of legible forms is scanned into computer by Microtek ScanMaker 4180. The resolution is set to 300 dpi. 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 2.1 M bytes.

2.4.2 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 preserve the smoothness of its strokes intersecting the frame lines. Then, we binarize handwriting image using Otsu algorithm (Otsu 1979). The binary image is named after the grayscale image and a letter "b" is inserted as the prefix. The black and white version of the handwriting image named 04090902 is shown in Fig. 2.3.

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21个岗位上。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拍摄影片《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. 2.3 Binary image of a handwritten sample named 04090902. © (2007) IEEE. Reprinted, with permission, from Ref. Su et al. (2007b)

2.4.3 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 which corresponds to the end of each handwriting text line. The latter crosses off the deleted characters from each segment, key in the inserted

Fig. 2.4 The ground truth on document level of Fig. 2.3. Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

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。

characters and modify the substituted characters. An example of the labeled ground truth on document level is illustrated in Fig. 2.4. Further, each row of the text is extracted and saved as a separate file (ground truth on text line level).

For the segmentation-free recognition systems, there is no need to annotate character boundaries, while the segmentation-based or integrated segmentation-recognition systems must provide character level alignment. A recognition-based ground truthing tool is employed to annotate the HIT-MW database (Yin et al. 2009). An annotated segment is shown in Fig. 2.5.

2.4.4 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. 2.3). We have developed a fast skew correction algorithm to improve the recall rate of text lines.

Each Chinese character can be decomposed into a group of strokes. Those strokes fall into four basic patterns: horizontal stroke, vertical stroke, left-falling stroke, and right-falling stroke. In GB2312-1980, 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 (Wu and Ding 1992). In other words, there are on average six horizontal strokes in a Chinese character. In addition, about 99.8 % of characters in GB2312-1980 include horizontal strokes. Those statistics mean that horizontal strokes are stable features in Chinese characters.

We employ the angular histogram of the horizontal strokes. Since they encode rich information of horizontal strokes, we incorporate them into projection profile to estimate the skew angles of handwritten documents from HIT-MW database in a fast and robust way.

We calculate the horizontal run-length of each document and only keep the strokes whose run-length is greater than T_s , where T_s is a threshold relating to the

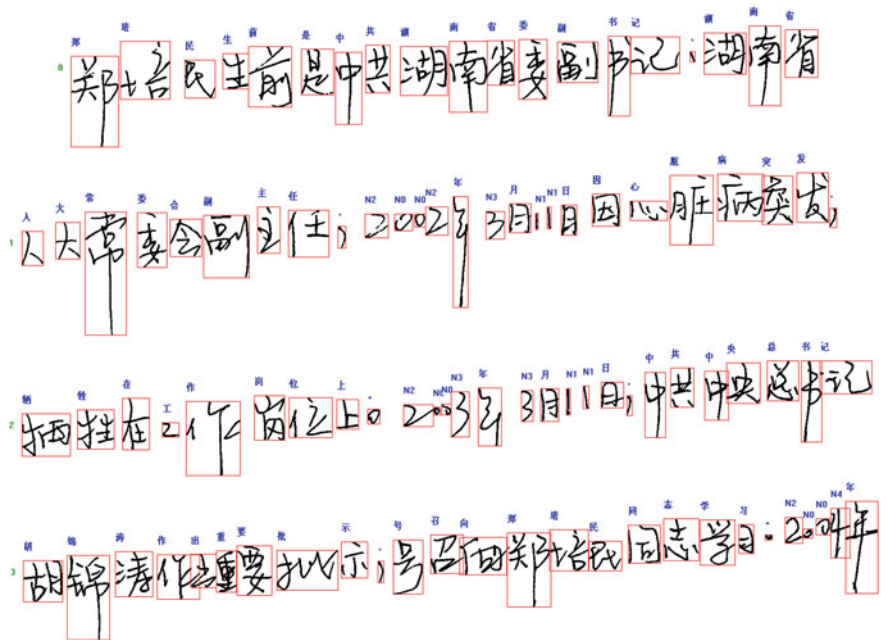
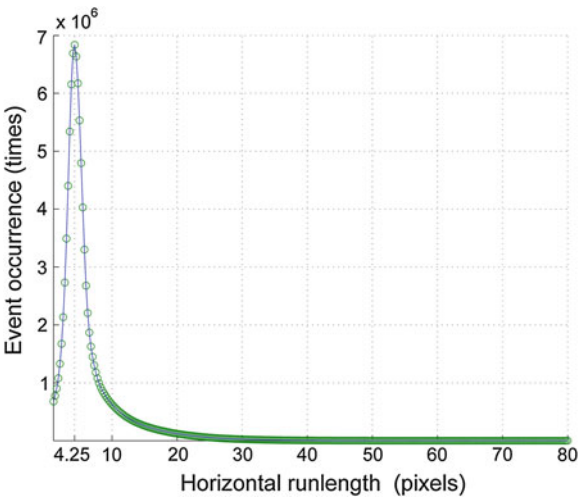


Fig. 2.5 An annotated segment in character level

Fig. 2.6 Horizontal run-length histogram of HIT-MW database. The curve is interpolated by cubic spline. © (2007) IEEE. Reprinted, with permission, from Ref. Su et al. (2007b)



average stroke width (ASW). The value of ASW is estimated from horizontal run-length histogram. Figure 2.6 illustrates the run-length histogram of HIT-MW database. The curve is smoothed by cubic spline. We set ASW as the runlength with maximum peak. From the curve, we can see that ASW is about 4.25 pixels.

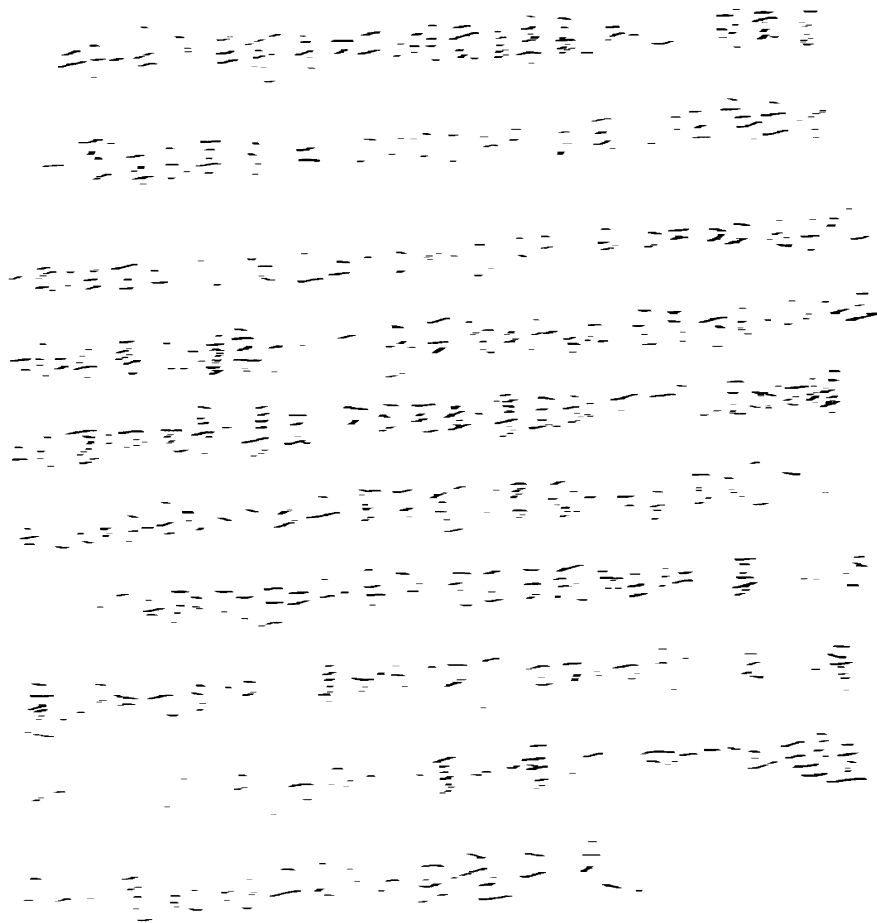


Fig. 2.7 The horizontal stroke map of Fig. 2.3. © (2007) IEEE. Reprinted, with permission, from Ref. Su et al. (2007b)

As a rule of thumb, T_s is bigger than double ASW. If we set $T_s = 10$ pixels, only 37.14 % of foreground pixels are retained. The kept strokes of Fig. 2.3 are shown in Fig. 2.7. We can see that the long downward strokes are stripped out. It benefits the horizontal projection analysis as it presents more acute transitions between peaks and valleys. In addition, small noise can be discarded with ease.

Further, we extract the center point of each remaining stroke and we call it representative point. This step discards at least nine-tenth foreground pixels of above horizontal strokes. In total, more than 96 % foreground pixels are discarded and the bias from long pseudo-strokes is reduced greatly. Figure 2.8 shows the center point version of Fig. 2.7.

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:

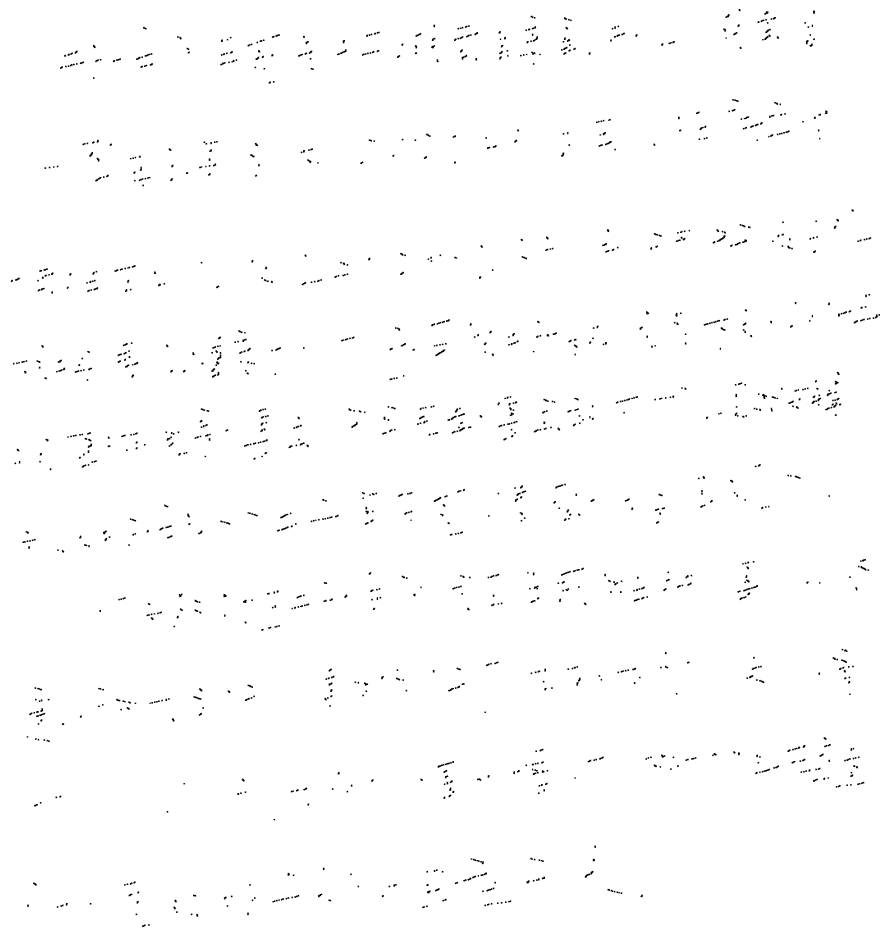


Fig. 2.8 The representative points of Fig. 2.7, where each point is expanded nine times to give a clear view. © (2007) IEEE. Reprinted, with permission, from Ref. Su et al. (2007b)

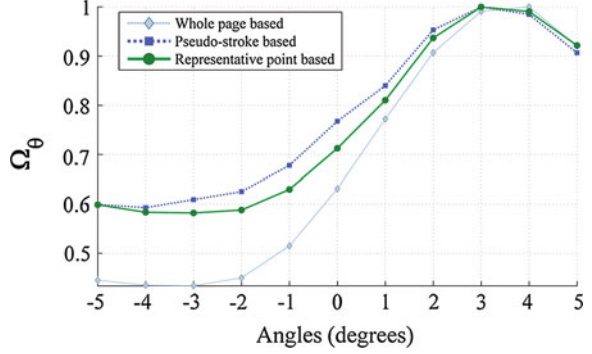
$$\theta' = \arg \max_{\theta} \Omega_{\theta}. \quad (2.1)$$

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

$$\Omega_{\theta} = a_1 \sigma_{\theta} + a_2 \phi_{\theta} + a_3 \varpi_{\theta}, \quad (2.2)$$

where σ_{θ} refers to the variance of the horizontal projection histogram, ϕ_{θ} the total gaps (the number of positions with zero histogram value) divided by document height, and ϖ_{θ} the normalized maximum of the histogram. The weights, a_i 's are determined from experiments.

Fig. 2.9 Curves of Ω_0 for Fig. 2.3 based on whole document, horizontal strokes, and representative points respectively when $\theta \in [-5, 5]$. The maximum of each curve corresponds to the estimated skew angle. © (2007) IEEE. Reprinted, with permission, from Ref. Su et al. (2007b)



The cost function curves for Fig. 2.3 based on the whole document, horizontal strokes, and representative points respectively are plotted in Fig. 2.9 with $a_1 = 1$ (the variances in the figure are normalized to one to put those curves into the same axis). From the figure, the angular projection of representative points has a comparative ability with that of horizontal strokes to extract the right skew angle. The decision by the angular projection of whole document is one degree off the right angle.

Now we show the range of skew angles which can be addressed by our method. A horizontal stroke can be modeled as a rectangle and we define: l is the width of the stroke, h the height of the stroke, γ the acute angle between the diagonal and horizontal direction, and α the skew angle of the stroke.

Commonly l is bigger than T_s . Due to the value of α , three cases can be reached:

- (1) If $\alpha < \gamma$, the longest horizontal runlength is bigger than l and the horizontal stroke can be located (as in Fig. 2.10a);
- (2) If $\alpha = \gamma$, the longest horizontal runlength is l and the horizontal stroke can also be located (see Fig. 2.10b);
- (3) If $\alpha > \gamma$, the problem becomes a little complex. The longest horizontal runlength is $h/\sin\alpha$. Thus the relationships between T_s and $h/\sin\alpha$ determinate whether the horizontal stroke can be located. When $h/\sin\alpha$ is bigger than or equal to T_s , the horizontal stroke can be located (showed in Fig. 2.10c); otherwise, it will be lost (shown in Fig. 2.10d).

Specifically, the maximum of skew angle should be smaller than 25 degrees when: $T_s = 10$, $h = 4.25$, and $l > 10$. By carefully tuning the ratio between h and T_s , we can determine the range of skew angles to be examined.

After the skew correction, the recall rate of the text lines by global horizontal projection is improved by 17.09 %, as indicated in Table 2.2. Three out of ten text lines of the handwriting in Fig. 2.3 can be successfully recalled by global projection following the skew correction (as in Fig. 2.11a, c, d).

In order to handle the complex text lines (just as in Fig. 2.11b), we adapt the genetic algorithm-based HIDER method (Su et al. 2008) to find the failure blocks (those cannot successfully separated by partial projection) and then heuristic-based thinning algorithm [similar to (Liang and Shi 2005)] will be used to extract the text

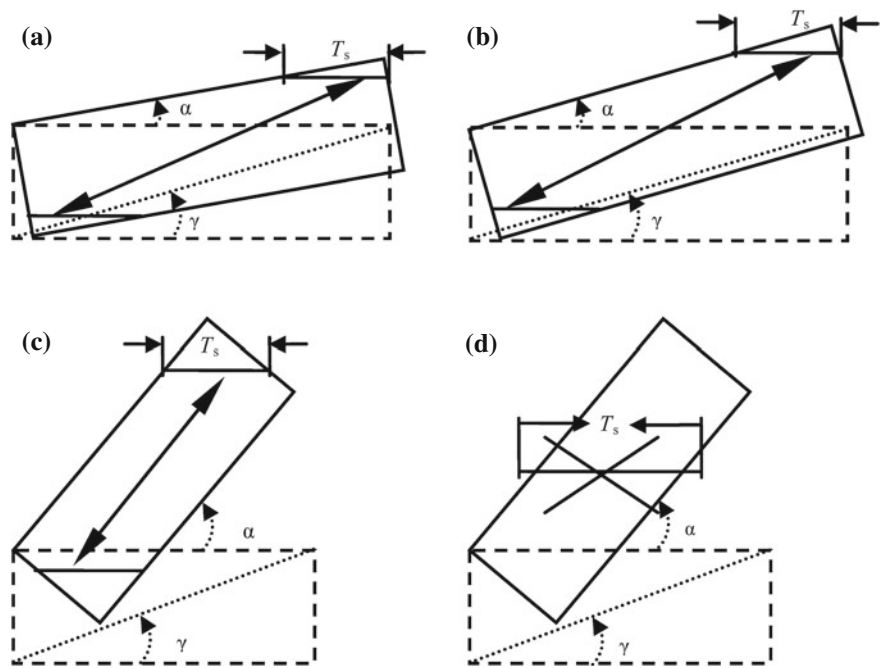


Fig. 2.10 A horizontal stroke with a skew of α will be located (a-c) and lost (d). The arrow indicates the zone containing horizontal runlength bigger than T_s . © [2007] IEEE. Reprinted, with permission, from Ref. Su et al. (2007b). **a** $\alpha < \gamma$ **b** $\alpha = \gamma$ **c** $\alpha > \gamma$, $h/\sin \alpha \geq T_s$ **d** $\alpha > \gamma$, $h/\sin \alpha < T_s$

Table 2.2 The number of recalled text lines (recall rate) in two setups

No skew correction	With skew correction
4062 (51.92 %)	5394 (69.01 %)

© (2007) IEEE. Reprinted with correction, with permission, from Ref. Su et al. (2007b)

border lines. As a key preprocessing step for Chinese handwriting recognition, following researchers have studied the text line segmentation problem from different perspectives and steady progress can be observed (Yin and Liu 2009a, b; Koo and Cho 2012).

2.5 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

- (a) 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拍摄影片《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。
- (b)
- (c)
- (d)

Fig. 2.11 The segmentation result by global projection after skew correction. Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media. **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

Table 2.3 Lexicon of HIT-MW database versus GB2312-1980 character set (unit: characters)

Within GBset				Beyond GBset
fIGBset	sIGBset	ASCII	Others	
2746	215	48	27	5

Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

distributions and an appealing lexical coverage on the People’s Daily corpus. Next, we focus our attention on two key handwriting phenomena, i.e., miswriting and erasing, and analyze them, respectively.

2.5.1 Basic Statistics

We have collected 853 legible Chinese handwriting samples. There are 1,86,444 characters in total including letters, punctuations besides Chinese characters, and these characters lead to 8,664 text lines. By simple computation, we get the 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, on average each character occurs 61.31 times. Most of the entries fall into GB2312-1980 character set (hereafter, abbreviated as GBset), and details are summarized in Table 2.3. Chinese handwritten character databases (such as HCL2000, IAAS-4M) only consist of the first level Chinese characters of GBset (fIGBset in short, and similarly sIGBset 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 fIGBset but a quantity of sIGBset are included (even several characters beyond GBset are included).

Moreover, to check its representative capability, we plot its coverage over People’s Daily corpus with 7,95,09,778 characters in Fig. 2.12. Note that, the corpus has already excluded the data of People’s Daily 2004 to give 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. 2.13 between lexicon of database and that of corpus. Each dot in the figure, (x, y), means that a character

Fig. 2.12 Lexicon size of HIT-MW versus coverage of people's daily corpus. Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

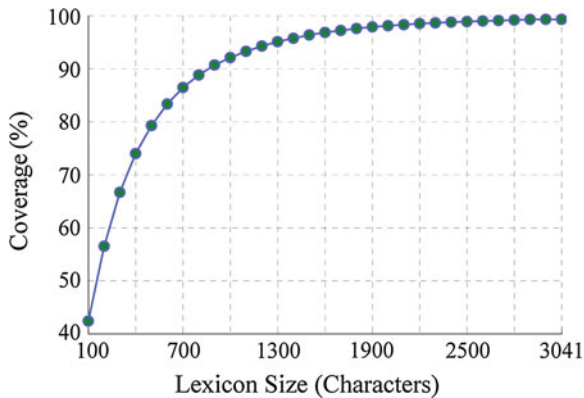
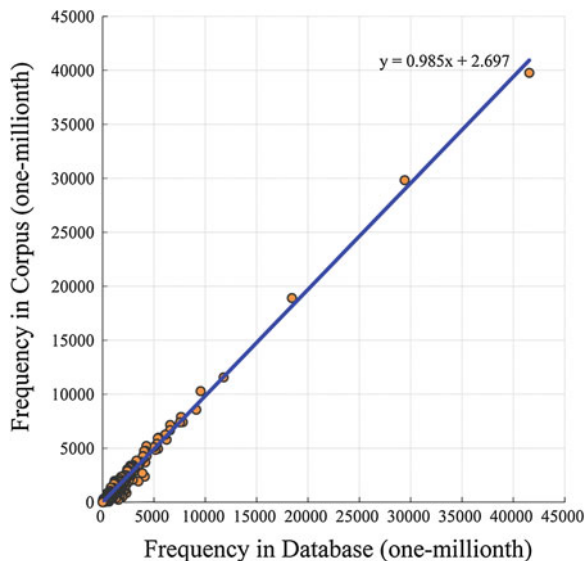


Fig. 2.13 The scatter map of the character frequency occurred in HIT-MW and people's daily corpus. Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media



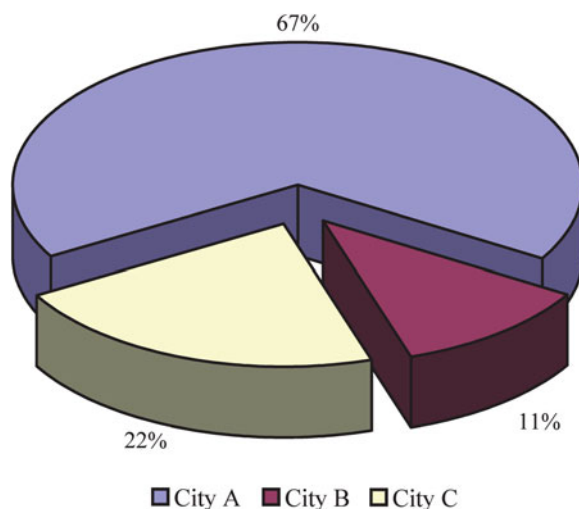
appears x times in database and y times in corpus. We can see from the figure that the cloud of dots is mainly spread along the auxiliary diagonal. Minimizing the least squares, we obtain a regression line as follow:

$$y = 0.9853x + 2.6973. \quad (2.3)$$

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. 2.14, the sampled writers are mainly from City A with a proportion of 67 %. Seen from Table 2.4, the department distribution of writers is near to that

Fig. 2.14 Sampling percentage of three regions. Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media



calculated from real data of college students of 2004 (National Bureau of Statistics of China 2005). Similarly, Table 2.5 shows that the sex distribution of our database has a good coincidence with that calculated from real educational statistics of 1998 (Ministry of Education 1998).

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

2.5.2 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. 2.15.

We group erasures by erasing mark and month in Table 2.6. 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 2.6, 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 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.

Table 2.4 Writers from science and engineering departments versus college students of 2004 from that reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

Of 2004	Sampled
61.37 %	60.69 %

Table 2.5 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 %

Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

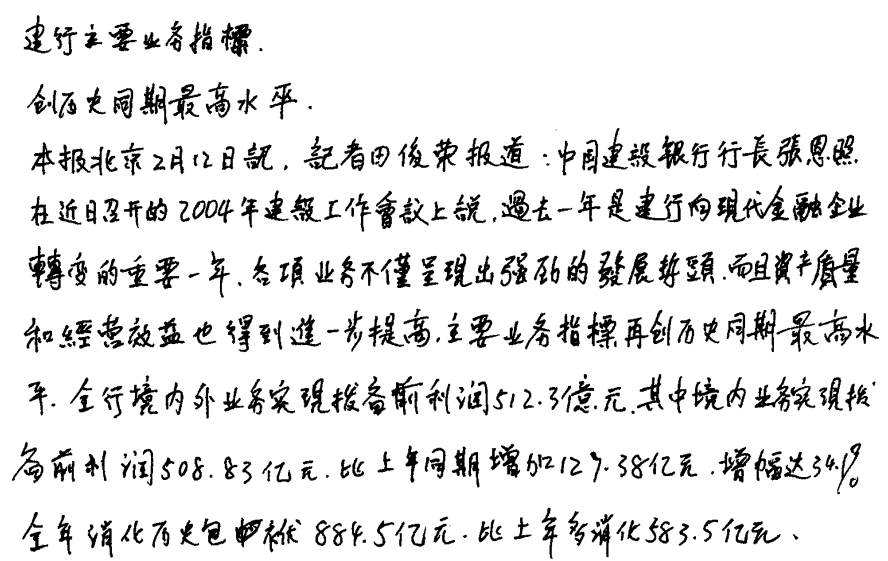


Fig. 2.15 A piece of handwriting with an erasure. Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

Second, analyzing the occurrences in each month, we can also infer that erasures are stable phenomena. On average, 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 mainly fall in four 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

Table 2.6 Statistics on erasure

Mark	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	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

Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

recognition performance. It is good news that the erasing marks manifest an excellent grouping possibility and that gives a promise for erasure modeling.

2.5.3 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 Chinese Pinyin: jian-hang-gong-zuo) as “建设工作” (in Chinese Pinyin: jian-she-gong-zuo) will improve the performance in tri-gram environment (see Eq. (2.4) to get a illustration).

$$\begin{aligned} & \frac{p(04\text{年建设工作})}{p(04\text{年建行工作})} \\ & \approx \frac{p(\text{设}|\text{建年})p(\text{工}|\text{设建})p(\text{作}|\text{工设})}{p(\text{行}|\text{建年})p(\text{工}|\text{行建})p(\text{作}|\text{工行})} \\ & \approx \frac{C(\text{年建设})C(\text{建设工})C(\text{设工作})C(\text{建行})C(\text{行工})}{C(\text{年建行})C(\text{建行工})C(\text{行工作})C(\text{建设})C(\text{设工})} \\ & = \frac{156 \times 1885 \times 635 \times 561 \times 1354}{6 \times 3 \times 1146 \times 109621 \times 1905} = 32.93 \gg 1 \end{aligned}$$

(2.4)

Table 2.7 Statistics of miswriting

(a) <i>Miswriting characters (unit: character)</i>			
Total	Deletion	Insertion	Substitution
2280	1884	110	274
(b) <i>Miswriting blocks and its linguistic effect (unit: block)</i>			
Total	Hurting context	Favoring context	
824	274	281	

Reprinted from Ref. Su et al. (2007a), with kind permission from Springer Science + Business Media

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 2.7. From Table 2.7a, 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 is easier to miss characters than other miswriting cases. In Table 2.7b, 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 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. Suppose the recognition rate without SLM is 65 %, 80 % after SLM, and there is no rejection. It is 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 analyze the substituted blocks, we may infer some tips concerning nerve mechanism of Chinese handcopying (Fu et al. 2002) which is out of the scope of this book.

2.6 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.

- (1) 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 Sect. 2.4.4, using global projection method directly, only 51.92 % of them can be correctly separated. The failure lies in irregular text lines. As soon as single text line is concerned, irregularity 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.
- (2) 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 [Sect. 2.5.2](#), without modeling them, the recognition rate will suffer a bit. In this complex environment, more advanced techniques are needed.
 - (3) SLM in real situation. As we know, 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 them is writing and outlier (such as erasures). In addition, how to efficiently incorporate the SLM into the handwritten text recognition framework raises a new problem.
 - (4) Segmentation-free recognition. Current Chinese character recognition algorithms are all segmentation-based. As mentioned in [Sect. 2.1](#), 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.

2.7 Discussions and Conclusions

HIT-MW database inherits data sparseness from natural language, since texts are sampled from corpus. Character frequency of database is shown in [Table 2.8](#). 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

Table 2.8 Character frequency of HIT-MW database

Occurrences	# of Characters
≥1000	16
≥100	456
≥10	1469
≥5	1853

Reprinted from Ref. Su et al. ([2007a](#)), with kind permission from Springer Science + Business Media

our database and obtain complete f1GBset, we can incorporate character level databases (such as IAAS-4M, HCL2000) into our database.

The handwritten Chinese text database discussed in this chapter 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 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 recognition. Study on them may promote the real-world Chinese handwriting recognition greatly.

The database can be downloaded freely. The latest details are available at <https://sites.google.com/site/hitmwdb/>. In addition, the ground truth and the grayscale version of the database are also available upon request (Please contact hitmwdb@gmail.com).

References

- U. Bhattacharya, B.B. Chaudhuri, Databases for research on recognition of handwritten characters of Indian scripts, in *Proceedings of the 8th International Conference on Document Analysis and Recognition*, Seoul, pp. 789–793 (2005)
- R.G. Casey, E. Lecolinet, A survey of methods and strategies in character segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**(7), 690–706 (1996)
- Q. Fu, X. Ding, C. Liu, A new AdaBoost algorithm for large scale classification and its application to Chinese handwritten character recognition, in *ICFHR*, pp. 43–48 (2008)
- S. Fu, Y. Chen, S. Smith, S. Iversen, P.M. Matthews, Effects of word form on brain processing of written Chinese. *NeuroImage* **17**(3), 1538–1548 (2002)
- Y. Ge, Q. Huo, Z.D. Feng, Offline recognition of handwritten Chinese characters using Gabor features, CDHMM modeling and MCE training, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, pp. 1053–1056 (2002)
- 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)
- D. Guillevic, C.Y. Suen, Recognition of legal amounts on bank cheques. *Pattern Anal. Appl.* **1**(1), 28–41 (1998)
- I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, UNIPEN project of on-line data exchange and recognizer benchmarks, in *Proceedings of 12th ICPR*, pp. 29–33 (1994)
- W. Highleyman, An analog method for character recognition. *IRE Trans. Electron. Comput.* **EC-10**, 502–512 (1961)
- J. Hull, A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(5), 550–554 (1994)
- L. Jin, Study on handprinted Chinese character recognition. Ph.D. thesis, South China University of Technology, Guangzhou, 1996 (in Chinese)

- L. Jin, Y. Gao, G. Liu, Y. Li, K. Ding, SCUT-COUCH2009—a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation. *Int. J. Doc. Anal. Recogn.* **14**(1), 53–64 (2011)
- E. Kavallieratou, N. Liolios, E. Koutsogeorgos, N. Fakotakis, G. Kokkinakis, The GRUHD database of Greek unconstrained handwriting, in *Proceedings of the 6th International Conference on Document Analysis and Recognition*, Seattle, pp. 561–565 (2001)
- D.H. Kim, Y.S. Hwang, S.T. Park, E.J. Kim, S.H. Paek, S.Y. Bang, Handwritten Korean character image database PE92. *IEICE Trans. Inf. Syst.* **E79-D**(7), 943–950 (1996)
- G. Kim, V. Govindaraju, S.N. Srihari, An architecture for handwritten text recognition systems. *Int. J. Doc. Anal. Recogn.* **2**(1), 37–44 (1999)
- H.I. Koo, N.I. Cho, Text-line extraction in handwritten Chinese documents based on an energy minimization framework. *IEEE Trans. Image Process.* **21**(3), 1169–1175 (2012)
- Z. Liang, P. Shi, A metasynthetic approach for segmenting handwritten Chinese character strings. *Pattern Recogn. Lett.* **26**(10), 1498–1511 (2005)
- C.-L. Liu, M. Koga, H. Fujisawa, Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(11), 1425–1437 (2002)
- C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, CASIA online and offline Chinese handwriting databases. in *ICDAR* pp. 37–41 (2011)
- Y.J. Liu, J.W. Tai, J. Liu, 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)
- U. Marti, H. Bunke, The IAM-database: an English sentence database for off-line handwriting recognition. *Int. J. Doc. Anal. Recogn.* **5**(1), 39–46 (2002)
- U.V. Marti, H. Bunke, A full English sentence database for off-line handwriting recognition, in *Proceedings of the 5th International Conference on Document Analysis and Recognition*, Bangalore, pp. 705–708 (1999)
- Ministry of Education, *Educational Statistics Yearbook of 1998* (People's Education Press, Beijing, 1998). (in Chinese)
- S. Mori, K. Yamamoto, H. Yamada, T. Saito, On a handprinted Kyoiku-Kanji character data base. *Bull. Electrotech. Lab.* **43**(11–12), 752–773 (1979)
- J.H. Munson, Experiments in the recognition of hand-printed text, part I-character recognition, in *Proceedings of the Fall Joint Computer Conference*, San Francisco, pp. 1125–1125 (1968)
- M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L. Higashigawa, K. Akiyama, On-line handwritten character pattern database sampled in a sequence of sentences without any writing instructions, in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, Ulm, pp. 376–381 (1997)
- National Bureau of Statistics of China, *China Statistical Year-book 2004* (China Statistics Press, Beijing, 2005) (in Chinese)
- N. Otsu, A threshold selection method from gray-level histogram. *IEEE Trans. Syst. Man Cybern.* **SMC-9**(1), 62–66 (1979)
- J.S. Park, H.J. Kang, S.W. Lee, Automatic quality measurement of gray-scale handwriting based on extended average entropy. in *Proceedings of the 15th international conference on pattern recognition*, Barcelona (2000)
- T. Saito, H. Yamada, K. Yamamoto, On the data base ETL9 of handprinted characters in JIS Chinese characters and its analysis. *IEICE Trans.* **J68-D**(4), 757–764 (1985)
- K. Sayre, Machine recognition of handwritten words: a project report. *Pattern Recogn.* **5**(3), 213–228 (1973)
- A.W. Senior, A.J. Robinson, An off-line cursive handwriting recognition system. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3), 309–321 (1998)
- T.-H. Su, T.-W. Zhang, D.-J. Guan, Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text. *Int. J. Doc. Anal. Recogn.* **10**(1), 27–38 (2007a)

- T.-H. Su, T.-W. Zhang, H.-J. Huang, Y. Zhou, Skew detection for Chinese handwriting by horizontal stroke histogram, in *IEEE Proceedings of the 9th International Conference on Document Analysis and Recognition*, pp. 899–903 (2007b)
- T. Su, T. Zhang, H. Huang, G. Xue, Z. Zhao, Transformation-based hierarchical decision rules using genetic algorithms and its application to handwriting recognition domain, in *IEEE Congress on Evolutionary Computation*, pp. 951–956 (2008)
- C.Y. Suen, M. Berthod, S. Mori, Automatic recognition of handprinted characters—the state of the art. *Proc. IEEE* **68**(4), 469–487 (1980)
- C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam, Computer recognition of unconstrained handwritten numerals. *Proc. IEEE* **80**(7), 1162–1180 (1992)
- L.-T. Tu, Y.S. Lin, C.P. Yeh, I.S. Shyu, J.L. Wang, K.H. Joe, W.W. Lin, Recognition of handprinted Chinese characters by feature matching, in *Proceedings of the 1st National Workshop on Character Recognition*, pp. 166–175 (1991)
- C. Viard-Gaudin, P.M. Lallican, S. Knerr, P. Binter, The IRESTE On/Off (IRONOFF) dual handwriting database, in *Proceedings of the 5th International Conference on Document Analysis and Recognition*, Bangalore, pp. 455–458 (1999)
- A. Vinciarelli, S. Bengio, H. Bunke, Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(6), 709–720 (2004)
- Y. Wu, X. Ding, *Character Recognition: Theory, Method and Implementation* (Higher Education Press, Beijing, 1992) (in Chinese)
- M.E. Yacoubi, M. Gilloux, J.M. Bertille, 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)
- F. Yin, C.-L. Liu, Handwritten Chinese text line segmentation by clustering with distance metric learning. *Pattern Recogn.* **42**(12), 3146–3157 (2009)
- F. Yin, C.-L. Liu, A variational Bayes method for handwritten text line segmentation, in *ICDAR*, pp. 436–440 (2009b)
- F. Yin, Q.-F. Wang, C.-L. Liu, A tool for ground-truthing text lines and characters in off-line handwritten Chinese documents, in *ICDAR*, pp. 951–955 (2009)
- H. Zhang, J. Guo, Introduction to HCL2000 database, in *Proceedings of Sino-Japan Symposium on Intelligent Information Networks*, Beijing (2000)
- S. Zhou, Q. Chen, X. Wang, HIT-OR3C: an opening recognition corpus for Chinese characters, in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems* (2010)
- M. Zimmermann, H. Bunke, N-gram language models for offline handwritten text recognition, in *9th International Workshop on Frontiers in Handwriting Recognition*, Kokubunji, pp. 203–208 (2004)

Chinese Handwriting Recognition: An Algorithmic
Perspective

Su, T.

2013, XI, 124 p. 62 illus., 16 illus. in color., Softcover

ISBN: 978-3-642-31811-5