

Offline Chinese handwriting recognition: an assessment of current technology

Sargur N. Srihari✉, Xuanshen Yang, Gregory R. Ball

Center of Excellence for Document Analysis and Recognition (CEDAR), Department of Computer Science and Engineering,
University at Buffalo, State University of New York, Amherst, NY 14228, USA

© Higher Education Press and Springer-Verlag 2007

Abstract Offline Chinese handwriting recognition (OCHR) is a typically difficult pattern recognition problem. Many authors have presented various approaches to recognizing its different aspects. We present a survey and an assessment of relevant papers appearing in recent publications of relevant conferences and journals, including those appearing in ICDAR, SDIUT, IWFHR, ICPR, PAMI, PR, PRL, SPIE-DRR, and IJDAR. The methods are assessed in the sense that we document their technical approaches, strengths, and weaknesses, as well as the data sets on which they were reportedly tested and on which results were generated. We also identify a list of technology gaps with respect to Chinese handwriting recognition and identify technical approaches that show promise in these areas as well as identify the leading researchers for the applicable topics, discussing difficulties associated with any given approach.

Keywords Chinese recognition, OCR, document analysis, survey, assessment, line segmentation, OCR architecture

1 Introduction

Despite the widespread creation of documents in electronic form, the conversion of handwritten material on paper into electronic form continues to be a problem of great economic importance. While there exist satisfactory solutions to cleanly printed documents in many scripts, particularly in the Roman alphabet, solutions to handwriting still face many challenges. In terms of methods to recognize different languages and scripts, although there exist many commonalities in image processing and document pre-processing, e.g., noise removal, zoning, line segmentation, etc., recognition methods necessarily diverge depending on the particular script.

This paper is a survey and an assessment of the state-of-the-art in Chinese handwriting recognition. The approach

taken is one of studying recent papers on the topic published in the open literature. Particular attention has been paid to papers published in the proceedings of recent relevant conferences as well as journals where such papers are typically published. The conference proceedings covered in this assessment are: *International Conferences on Document Analysis and Recognition*, *International Workshops on Frontiers in Handwriting Recognition*, *Symposia on Document Image Understanding Technology*, *SPIE Conferences on Document Recognition and Retrieval* and *International Conferences on Pattern Recognition*. The journals covered are: *Pattern Recognition Journal*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Document Analysis and Recognition* and *Pattern Recognition Letters*.

The breadth and depth of any survey is necessarily subjective. The goal here was to assess the most promising approaches and to identify any gaps in the technology so that future research can be focused.

1.1 Chinese script overview

From the viewpoint of recognition, Chinese script consists of approximately 50,000 characters, of which only a few thousand are commonly used. For instance, 99.65% of the usage is covered by 3775 common characters, 99.99% by 6763 and 99.999% by 8500 characters. Sentences are written horizontally left to right from the top of the page to the bot-

故天将降大任于是人也，必先苦其心志，劳其筋骨，饿其体肤，空乏其身，行拂乱其所为，所以动心忍性，曾益其所不能。

Fig. 1 A page of simplified Chinese text

tom, as in English, called simplified (Fig. 1), or vertically from top to bottom, the lines going right to left across the page, called traditional (Fig. 2)

Traditional Chinese is less often used in handwriting today; it has been used more often historically, except in Taiwan and Hong Kong. There are many strokes in traditional Chinese characters, about 16.03 strokes per character. To reduce the complexity, from year 1956 to 1964, 2235 simplified Chinese characters were revised so as to substitute for corresponding traditional ones in mainland China. The average strokes reduced to 10.3 per character.

While the number of characters is high, each character is made from about 500 component sub-characters (called radicals) in predefined positions and written in a predefined order. While the stroke order can be exploited by on-line recognition algorithms, off-line recognition is particularly challenging since this information is lost. Due to the high number of characters, the average word length of a Chinese word is short consisting of 2-4 characters. Furthermore, the characters are always written in “print fashion”, not connected. Segmentation is often easier than in other languages—characters rarely need to be split apart, but it is sometimes challenging to determine whether two radicals are in fact two separate characters or two component parts of the same character. Still, the largest challenge is recognizing the large number of characters and the majority of research has been devoted to overcoming this difficulty. Furthermore, the great individual variation in writing characters (Fig. 3) contributes to the problem. Many methods have been developed for recognizing individual characters. Other works have been devoted to such topics as word or address recognition and language discrimination, such as determining if a piece of writing consists of traditional or simplified Chinese characters. Chinese recognition technology generally falls into three main categories—document preprocessing, character recognition, and word recognition.

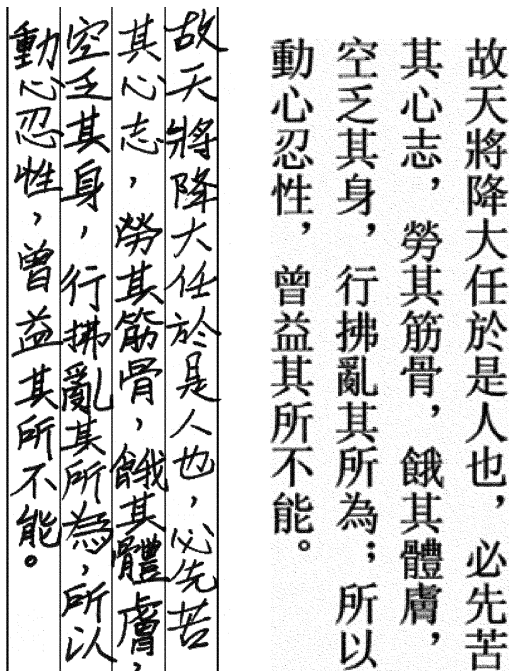


Fig. 2 A page of traditional Chinese text



Fig. 3 “Ying” (which means elite) written by 25 writers from Ref. [1]

1.2 Overview of the recognition process

While different algorithms vary in their implementation details, most follow a general path from document to UNICODE text. Figure 4 shows a general process of Chinese handwriting recognition.

A document is generally input as a scanned grayscale image. Preprocessing may convert the document to a black and white image (binarization), though some methods extract features directly from the gray-scale images (such as the gradient feature, described in Section 3). Preprocessing can also convert the initial image to some representation which is easier to process than the raw image, such as chain code or skeleton representation. Additional preprocessing steps such as noise reduction, interfering-lines removal, and smoothing can also be applied. This cleaned image is then passed to a segmentation algorithm.

A segmentation algorithm splits a large image into smaller regions of interest. For example, a page segmentation algorithm segments an unconstrained page of text into component lines. Lines are segmented into words, words into characters or subcharacters. The segmentation algorithm can then pass its output to a normalization unit, then to a recognition algorithm. The normalization unit aims to reduce shape variation within the same class by regulating size, position, and shape of each character. If the segmentation algorithm is completely distinct from the recognition, its performance can be a bottleneck for recognition. If the segmentation is incorrect, the lines, words, or characters might not be present in a single image, or multiple ones might be present in the same image. For this reason, some algorithms (such as segmentation-free algorithms) might make only partial use of a segmentation tool. They may use the

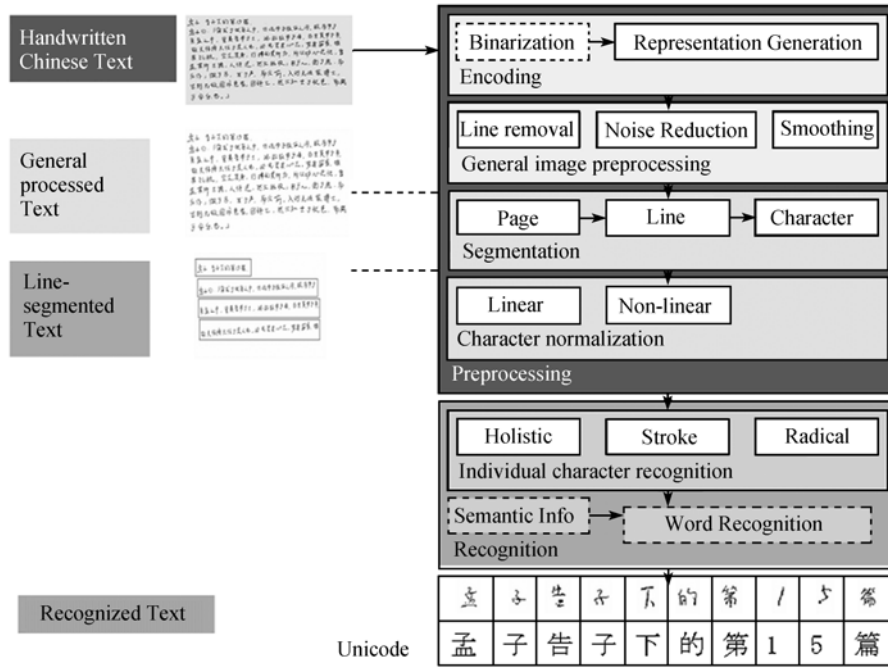


Fig. 4 Process of Chinese handwriting recognition

output of page segmentation, but not line, for example. Another approach is to use a segmentation algorithm as a tool to make suggestions rather than to dictate choices.

The recognition process usually works recursively as a closed loop shown in Fig. 5. The output of early round of classification may be used as input of later round to improve the accuracy of the early process, e.g., to improve segmentation or feature extraction accuracy.

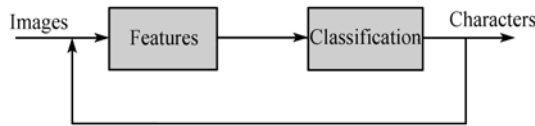


Fig. 5 Recursive process of recognition

The accuracy of recognition largely depends on the cursiveness of handwriting. Roughly, researchers rank Chinese handwriting into three categories: regular or hand-print scripts, fluent scripts, and cursive scripts. According to an invited talk in ICPR2006, the accuracy of recognizing each category reaches 98.1%, 82.1%, and 70.1%, respectively. Though how to decide the threshold of cursiveness to categorize Chinese handwriting characters is still not obvious or formalized, it is clear that researchers still face great difficulty to recognize cursive scripts. Fortunately, many groups have been devoted into this challenging problem.

1.3 Databases

Some large databases have been established along with the development of recognition technology. While some researchers use their own collected databases, publicly recognized databases are often used so as to facilitate performance comparison and technique improvement. Due to the strong

similarity of many Japanese and Chinese characters, Japanese character databases often have utility in the Chinese recognition setting. Following are some Chinese character databases used in literature.

ETL character databases: were collected at the Electrotechnical Laboratory under the cooperation with Japan Electronic Industry Development Association, universities and other research organizations. These databases ETL1-ETL9 contain about 1.2 million hand-written and machine-printed character images that include Japanese, Chinese, Latin and numeric characters for character recognition researches. Character images of the databases were obtained by observing OCR sheets or Kanji printed sheets with a scanner. All databases ETL1 - ETL9 are gray-valued image data. ETL8 and ETL9, binarized image data (ETL8B and ETL9B) are open to the public. ETL9 consists of 2 965 Chinese and 71 Hiragana, 200 samples per class written by 4000 people. ETL8 consists of 881 classes of handwritten Chinese and 75 classes of Hiragana, 160 samples per class written by 1 600 people. There are 60×60, 64×63, 72×76 and 128×127 pixel versions of character images. The character image files consist of more than one record, which has a character image and ID information with a correct code.

KAIST Hanja1 and Hanja2: was collected by the Korea Advanced Institute of Science and Technology. The Hanja1 database has 783 most frequently used classes. For each class, 200 samples, artificially collected from 200 writers, are contained. The Hanja2 database has 1 309 samples collected from real documents. The number of samples in the Hanja2 database varies with the class. The image quality of Hanja1 is quite clean, while the Hanja2 database is very noisy.

JEITA-HP: was originally collected by Hewlett-Pack-

ard Japan and later released by JEITA (Japan Electronics and Information Technology Association). It consists of two datasets: Dataset A (480 writers) and Dataset B (100 writers). Generally, Dataset B is written more neatly than Dataset A. The entire database consists of 3214 character classes (2 965 kanji, 82 hiragana, 10 numerals, 157 other characters (English alphabet, katakana and symbols)). The most frequent hiragana and numerals appear twice in each file. Each character pattern has a resolution of 64×64 pixels, which are encoded into 512 bytes.

HCL2000: was collected by Beijing University of Posts and Telecommunications for China 863 project. It contains 3 755 frequently used simplified Chinese characters written by 1 000 different people. The writers' information is incorporated in the database to facilitate testing on grouping writers with different backgrounds.

ITRI: was collected by the Industrial Technology Research Institute (Taiwan Province, China). It contains 5 401 handwritten Chinese classes with 200 samples per class.

4MSL: was collected by the Institute of Automation Chinese Academy of Sciences. It contains 4 060 Chinese characters, 1 000 samples per class.

Databases vary in writing styles, cursiveness, etc. Figures 6–8 are examples from some of these databases.

1.4 Commercial products

Chinese handwriting recognition techniques have been widely implemented into commercial products in recent years. On-line handwriting recognition is broadly used for the input of PDAs, cellphones, and tablets; off-line technology has been embedded into systems that automatically transfer optical documents into editable digital format such as MS Word. Hanwang Technology (www.hw99.com) is a leading company of China in the fields of research, development, and manufacture of handwriting recognition technologies and products. It cooperates with or provides technology to many software, PC, PDA, and cellphone compa-



Fig. 6 Samples of JEITA-HP database (writers A-492 and B-99)



Fig. 7 Samples of 4MSL database



Fig. 8 Samples of the KAIST databases [2]: (a) Hanja1, (b) Hanja2

Table 1 Some commercial products implementing CHR

Product	Company	Features and function
SuperPenx	Hanwang Technology	input handwriting Chinese and Japanese into Microsoft Word, Excel, PowerPoint, MSN, ICQ, WPS et al.
Smart Reader Master	Hanwang Technology	convert any type of paper documents into an editable and accurate formatted digital file, whether they are printed or handwritten
Smart Mouse	Beijing Wintone	supports all kinds of devices such as electronic pens, touch pad or cursor, etc. and the built-in handwriting recognition system fits mobile phones and PDAs.
ChinesePen	Beijing InfoQuick SinoPen	integrates many functions including Chinese characters input, learning Chinese, painting and entertainment by means of handwriting

nies such as Microsoft, Nokia, Lenovo, et al, and it's reported that it takes almost 95% of the market of handwriting input on cellphones in China. Many other companies in China also contribute, such as Beijing Wintone (www.wintone.com.cn) and Beijing InfoQuick SinoPen (www.sinopen.com.cn). Table 1 lists some of their products besides technique support or transfer to other companies.

2 Document preprocessing

Document preprocessing converts a scanned image into a form suitable for later steps. Preprocessing includes such actions as noise removal, smoothing, interfering-lines removal, binarization, representation code generation, segmentation, character normalization, slant or stroke width adjustment, etc. Individual writing variations and distortions are particularly disruptive to Chinese character recognition due to the complexity of the characters. To overcome this issue, normalization algorithms can make use of such features as stroke density variation, character area coverage, and centroid offset. While characters are usually not con-

nected as in script languages, segmentation can still be a significant issue. Many characters are composed of radicals that are themselves individual characters. While, ideally, characters should each cover the same approximate area, such precision is not always present in handwriting. This introduces ambiguity of adjacent characters-namely, whether two characters are present (the second one being composed of two primitives) or if three distinct characters are present. In this section, we focus on two preprocessing steps, segmentation and normalization, because they receive most attention in the preprocessing of Chinese handwriting.

2.1 Segmentation

The accuracy of segmenting Chinese characters, especially connected Chinese characters, is essential for the performance of a Chinese character recognition system. Because the key for Chinese recognition is individual character recognition, usually, the main effort is focused on segmenting pages into lines, then directly into characters. e.g., a word segmentation is generally ignored if semantic information (word lexicon) is not available. A general segmentation process of handwriting Chinese is shown in Fig.9.

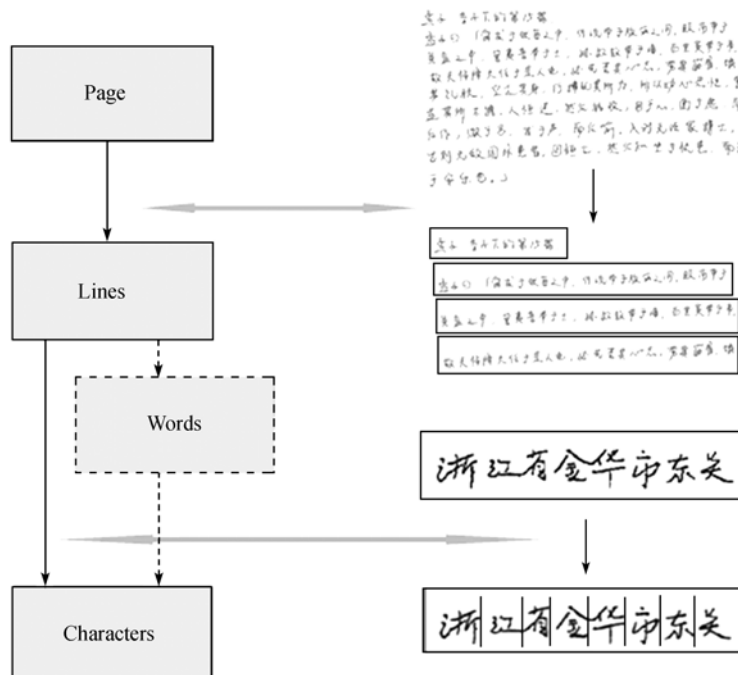


Fig. 9 Process of Chinese segmentation

Since in Chinese, the characters are always written in “print fashion”, or not connected, in some respects segmentation is easier than in other languages such as English. Two widely used techniques, vertical projection and connected component analysis, are able to handle the problem in some sense. On the other hand, Chinese characters usually consist of more than one radical and even some radicals themselves are characters, and handwriting strokes of adjacent characters may join. In general, non-linear paths are used to segment characters in such cases. Since the requirement for segmentation accuracy is inevitably very high and how to obtain non-linear paths is still an important issue, we need more powerful methods to handle the problem. Since Lu and Shridhar [3] reviewed different approaches for the segmentation of handwritten characters, many segmentation algorithms have been presented in literature. Table 2 lists some of the recent results on Chinese character segmentation. Based on what we have reviewed in literature, segmentation algorithms performed not very well if completely separated from recognition, and may become a bottleneck of the whole recognition system. The best results come from the methods that incorporate segmentation and recognition together,

using feedback and heuristics, even consulting semantic or context information.

2.2 Character normalization

The performance of handwritten character recognition is also largely dependent on character shape normalization, which aims to regulate the size, position, and shape of character images so as to reduce the shape variation between the images of the same class. Generally, mapping the input character image onto a standard image plane can make all the normalized images have the same dimensionality. Also, it is desirable to restore the deformation of characters. Linear normalization (LN) can scale the image despite its inner structure, and the nonlinear normalization (NLN) method based on line density equalization is popularly used now. The practical process of shape normalization may vary in detail for different systems. Figure 10 shows the process used by Dong et al.

Liu et al. have tested different normalization strategies, Fig. 11 shows some of their work in [11].

Table 2 Some results of segmentation

Publication	Method	Data set	Accuracy
2005 Xianghui wei etc. [4]	genetic algorithm	428 images, each of two Chinese characters	88.9
2005 Zhizhen Liang etc. [5]	metasynthetic	921 postal strings (7913 characters)	87.6
2001 Shuyan Zhao etc. [6,7]	two-stage	1000 strings containing 8011 characters	81.6
1999 Yi-Hong Tseng etc. [8]	recognition-based	125 text-line images containing 1132 characters	95.58
1998 Lin Yu Tseng etc. [9]	heuristic merging, DP	900 handwriting and 100 printing	95

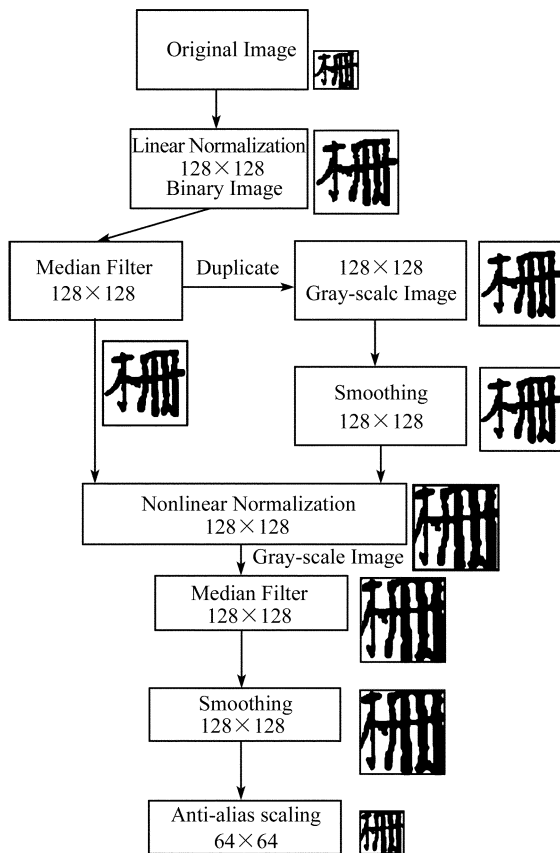


Fig. 10 Normalization process from Ref. [10]. The transformed image is shown beside each box

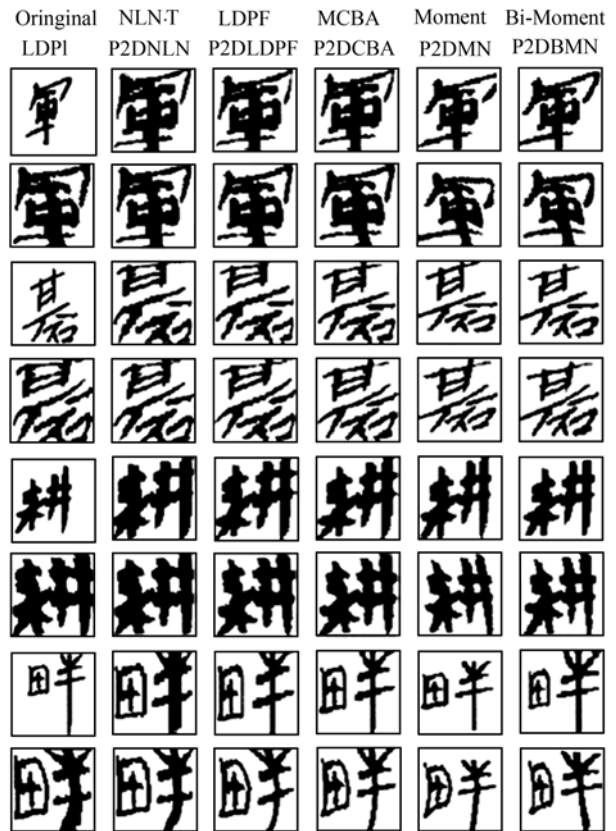


Fig. 11 Examples of character shape normalization from Liu’s. Each group of two rows include an input image and 11 normalized images generated by different methods

2.3 Bibliographic remarks

Yi-Hong Tseng et al. [12] presented a method for removing interfering-lines from interfered-characters. Black-run detection and character stroke recovery are the two main steps used in the process. It's shown that the recognition accuracy could be highly improved with the process.

Chiu et al. [13] proposed a feature-preserved thinning algorithm for handwritten Chinese characters. The proposed approach is free from the Y-shape distortion, L-shape distortion, hairy problem, shortening problem, and hole problem.

2.3.1 Segmentation

Wei et al. [4] proposed a method for segmenting connected Chinese characters based on a genetic algorithm. They achieved a segmentation accuracy of 88.9% in a test set without using special heuristic rules.

Liang et al. [5] applied the Viterbi algorithm to search segmentation paths and used several rules to remove redundant paths, then adopted a background thinning method to obtain non-linear segmentation paths. They tested on 921 postal strings (7 913 characters), and got a correct rate of 87.6%.

Zhao et al. [6, 7] proposed a two-stage approach including coarse and fine segmentation. They used fuzzy decision rules generated from examples to evaluate the coarse and fine segmentation paths. Experimental results on 1000 independent unconstrained handwritten Chinese character strings containing a total of 8 011 Chinese characters showed that their two-stage approach can successfully segment 81.6% of characters from the strings.

Yi-Hong Tseng et al. [8] presented a means of extracting text-lines from text blocks and segmenting characters from text-lines. A probabilistic Viterbi algorithm is used to derive all possible segmentation paths which may be linear or non-linear. They tested 125 text-line images from seven documents, containing 1 132 handwritten Chinese characters, and the average segmentation rate in the experiments is 95.58%.

Lin Yu Tseng et al. [9] presented a segmentation method based on heuristic merging of stroke bounding boxes and dynamic programming. They used strokes to build stroke bounding boxes first. Then, the knowledge-based merging operations are used to merge those stroke bounding boxes and, finally, a dynamic programming method is applied to find the best segmentation boundaries. They tested on 900 handwritten characters written by 50 people and 100 machine-printed characters, and got over 95% segmentation accuracy.

Gao et al. [14] proposed a segmentation by recognition algorithm using simple structural analysis of Chinese characters, a split and merge strategy for presegmentation, and level building algorithm for final decision.

2.3.2 Normalization

Wang [15] reviewed five shape normalization schemes in a unified mathematical framework, and discussed various feature extraction methodologies. Experimental results re-

veal that it can be beneficial to perform nonlinear shape normalization to improve the recognition performance on a large scale handwritten Chinese character set. Compared with a linear normalization method, which solely standardizes the size of a character image, four nonlinear normalization schemes produced better recognition performance. Furthermore, it was found that certain normalization schemes match better with certain feature sets. It provides some guidelines for selecting a normalization scheme and a feature set.

Cheng-Lin Liu et al. [16] described. substitutes to nonlinear normalization (NLN) in handwritten Chinese character recognition (HCCR). The alternative methods (moment normalization, center-bound, and bi-moment method) generate more natural normalized shapes and are less complicated than NLN. In Ref. [17], they proposed a new global transformation method, named modified centroid-boundary alignment (MCBA) method, for HCCR. It adds a simple trigonometric (sine) function to quadratic function to adjust the inner density. Experiments on the ETL9B and JEITA-HP databases show that the MCBA method yields comparably high accuracies to the NLN and bi-moment methods and shows complementarity. In Ref. [11], they proposed a pseudo 2D normalization method using line density projection interpolation (LDPI), which partitions the line density map into soft strips and generates 2D coordinate mapping function by interpolating the 1D coordinate functions that are obtained by equalizing the line density projections of these strips. They got a significant improvement: for example, using the feature ncf-c, the accuracy on the ETL9B test set was improved from 99.03% by NLN-T to 99.19% by P2DNLN, and the accuracy of HEITA-HP test set was improved from 97.93% by NLN-T to 98.27% by P2DNLN. In Ref. [18], they further compared different normalization methods with better implementation of features, and proposed that the best normalization-feature combination on unconstrained characters is P2DBMN with NCCF.

3 Recognition

Character recognition is perhaps the most scrutinized problem. The approaches are wide-ranging, but from a structural aspect, there are three categories: radical-based, stroke-based, and holistic approaches (Fig. 12).

The radical-based approach attempts to decompose the character into radicals and categorizes the character based on the component parts and their placement [1, 19–24]. Instead of attempting to recognize the radicals directly, stroke-based approaches try to break the character into component parts as strokes [2, 25–36], then recognize the character according to the stroke number, order, and position. Holistic approaches ignore such components entirely, with the idea that they may be too hard to individually identify-much as recognizing individual characters in the middle of a word is very difficult in script based languages. They try to holistically recognize the character [10, 37–53]. Holistic

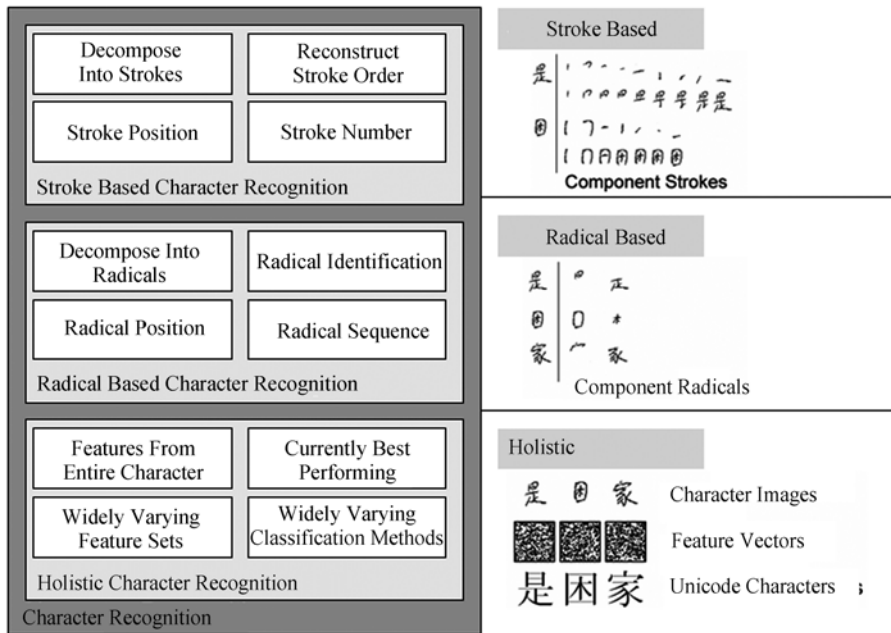


Fig. 12 Character recognition

approaches are favored by their generally better performance; gradient features and directional element features (DEF) have shown great ability to capture holistic characteristics of different characters in practice.

During recognition, a coarse classifier is often used to prune the search candidates of fine classifier. It accelerates the recognition of a system. Recently, Feng-Jun Guo et al. [54] introduced an efficient clustering based coarse classifier for Chinese handwriting recognition systems. They define a candidate-cluster-number for each character that indicates the within-class diversity in the feature space, and then they use a candidate-refining module to reduce the size of the candidate set. Also, in their paper, they cited some previous work on coarse classifiers.

3.1 Holistic approaches

Since Chinese characters are highly complex, there is a high number of potential feature sets. Due to the structure of a Chinese character, several features naturally present themselves. After choosing such a set of features, initial character recognitions vary. Statistical approaches, structural approaches, hidden Markov models, neural networks, support vector machines, Markov random fields and different discriminant functions have been applied. Several groups have achieved very high recognition rate in their experiments on large, realistic data sets [10, 40, 41].

Hailong Liu et al. [41] use gradient features and a quadratic classifier with multiple discrimination schemes for Chinese handwriting recognition. Gradient features are extracted from a gray scale image (with gray scale values ranging from 0 to 255). 3×3 Sobel operators are used to get the horizontal and vertical grayscale gradient at each image

pixel. They then define L directions with an equal interval $2\pi/L$, and decompose the gradient vector into its two nearest directions in a parallelogram manner, as illustrated in Fig. 13.

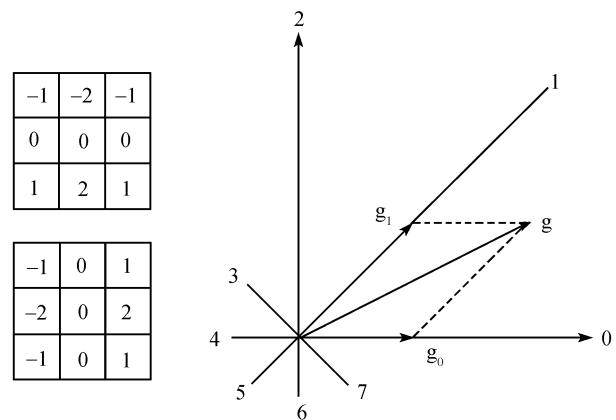


Fig. 13 Gradient feature extraction using Sobel operators and gradient vector decomposition (eight directions for illustration)

In the classification stage, the performance of the modified quadratic discriminant function (MQDF) classifier is enhanced by multiple discrimination schemes, including minimum classification error (MCE) training on the classifier parameters and modified distance representation for similar characters discrimination by compound Mahalanobis function (CMF). The system achieves very high accuracy: 99.54% on MNIST, 99.33% on ETL9B, and 98.56% on HCL2000. For a d -dimensional feature vector x , the QDF distance can be represented as follows,

$$g_i(x) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \log |\Sigma_i|, \quad i = 1, 2, \dots, C$$

where C is the number of character classes. μ_i and Σ_i denote the mean vector and the covariance matrix of class Σ_i . Using orthogonal decomposition on Σ_i , and replacing the minor eigenvalues with a constant Σ^2 to compensate for the estimation error caused by the small training set, the MQDF distance is derived as

$$g_i(x) = \frac{1}{\sigma^2} \left\{ \|x - \mu_i\|^2 - \sum_{j=1}^k \left(1 - \frac{\sigma^2}{\lambda_{ij}} \right) [\varphi_{ij}^T(x - \mu_i)]^2 \right\} + \sum_{j=1}^k \log \lambda_{ij} + (d - k) \log \sigma^2, \quad i = 1, 2, \dots, C$$

where λ_{ij} and φ_{ij} denote the j -th eigenvalue (in descending order) and the corresponding eigenvector of Σ_i . k ($k < d$) denotes the number of dominant principal axes. The classification rule then becomes

$$(x) \arg \min_i g_i(x)$$

Jian-xiong Dong et al. [10] tested support vector machines (SVM) on a large data set for the first time. Their system achieved a high recognition rate of 99% on ETL9B. The following is the procedure they used for feature extraction. We cite it here to illustrate how holistic features are extracted.

1. Standardize the gray-scale normalized image such that its mean and maximum values are 0 and 1.0, respectively.

2. Center a 64×64 normalized image in an 80×80 box in order to efficiently utilize the information in the four peripheral areas.

3. Apply Robert edge operator to calculate image gradient strengths and directions.

4. Divide the image into 9×9 subareas of size 16×16 , where each subarea overlaps the eight pixels of the adjacent subareas. Each subarea is divided into four parts: A, B, C and D (see Fig. 14).

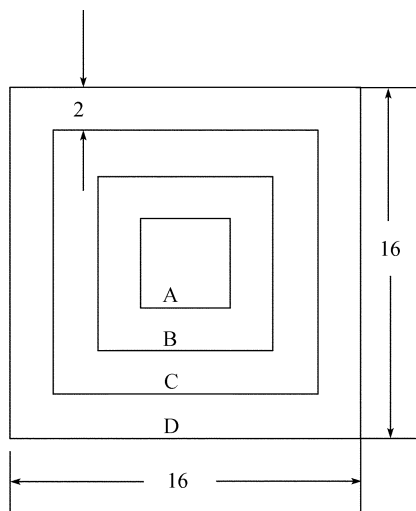


Fig. 14 16×16 subarea

5. Accumulate the strengths of gradients with each of 32

quantized gradient directions in each area of A, B, C, D. In order to reduce side-effects of position variations, a mask (4, 3, 2, 1) is used to down-sample the accumulated gradient strengths of each direction. As a result, a 32 dimensional feature vector is obtained.

6. Reduce the directional resolution from 32 to 16 by down-sampling using one-dimensional mask (1, 4, 6, 4, 1) after gradient strengths on 32 quantized directions are generated.

7. Generate a feature vector of size 1296 (9 horizontal, 9 vertical and 16 directional resolutions).

8. Apply the variable transformation $x^{0.4}$ to each component of the feature vector such that the distribution of each component is normal-like.

9. Scale the feature vector by a constant factor such that values of feature components range from 0 to 1.0.

Nei Kato et al. [40] extract DEF, then use city block distance with deviation (CBDD) and asymmetric Mahalanobis distance (AMD) for rough and fine classification. The DEF extraction includes three steps: contour extraction, dot orientation, and vector construction. The purpose of rough classification is to select a few candidates from the large number of categories as rapidly as possible, and CBDD is applied. The candidates selected are usually characters with similar structure, and then they are fed to fine classification by AMD. The recognition rate of this system on ETL9B reaches 99.42%. Let $v = (v_1, v_2, \dots, v_n)$ be an n -dimensional input vector, and $\mu = (\mu_1, \mu_2, \dots, \mu_n)$ be the standard vector of a category. The CBDD is defined as:

$$d_{CBDD}(v) = \sum_{j=1}^n \max\{0, |v_j - \mu_j| - \theta \cdot s_j\}$$

where s_j denotes the standard deviation of j th element, and θ is a constant. And the AMD is defined as:

$$d_{AMD}(v) = \sum_{j=1}^n \frac{1}{\hat{\sigma}_j^2 + b} (v - \hat{\mu}, \phi_j)^2$$

where b is a bias. Readers are referred to their papers cited above for more details.

Hidden Markov Model (HMM) have also been widely applied, in a variety of ways, to the problem of offline handwriting recognition. Bing Feng et al. [42] first segment each character image into a number of local regions, which is a line (row or column) of the character image, then extract feature vectors of these regions and assign each vector a value (a code) used to get the observations for the HMM. The states of the HMM are to reflect the characteristic space structures of the character. The training of HMM is an iterative procedure that starts from the initial models and ends when getting the local optimal ones. The initial state probability distribution π and state transition probability distribution A can be determined according to the characteristic of left-to-right HMM and the meaning of the state in representing handwritten character. The observation symbol probability distribution B can be decided by counting the appearance of each observation of the parts on training samples or just assigning an equal distribution. Yong Ge et

al. [37–39] use the Gaussian mixture continuous-density HMMs (CDHMMs) for modeling Chinese characters. Given an $N \times N$ binary character image, two feature vector sequences, $Z^H = (Z_0^H, Z_1^H, \dots, Z_{N-1}^H)$ and $Z^V = (Z_0^V, Z_1^V, \dots, Z_{N-1}^V)$, which characterize how the character image changes along the horizontal and vertical directions, are extracted independently. Consequently, two CDHMMs, λ^H and λ^V , are used to model Z^H and Z^V , respectively. Each CDHMM is a J -state left-to-right model without state skipping. The state observation probability density function is assumed to be a mixture of multivariate Gaussian PDFs. The CDHMM parameters can be estimated from the relevant sets of training data by using the standard maximum likelihood (ML) or minimum classification error (MCE) training techniques.

Table 3 lists more recent work on holistic-based recognition of handwritten Chinese characters.

3.2 Radical-Based Approaches

Radical-based approaches decompose a large set of Chinese characters into a smaller set of radicals. Then, the problem is converted to one of radical extraction and optimization of combination of the radical sequences. Figure 15 illustrates how four complex Chinese characters can be represented by a hierarchical composition of radicals. If the redundant information is ignored, the four characters can be classified by recognizing the last row of radicals, which should be intuitively easier than recognizing whole characters.

Daming Shi et al. [1, 19–23] have done successful work on radical-based recognition for off-line handwritten Chinese characters. They published a series of papers that elaborated an approach based on nonlinear active shape

modeling. The first step is to extract character skeletons by an image thinning algorithm, then landmark points are labeled semi-automatically, and principal component analysis is applied to obtain the main shape parameters that capture the radical variations. They use chamfer distance minimization to match radicals within a character, and during recognition the dynamic tunneling algorithm is used to search for optimal shape parameters. They finally recognize complete, hierarchical-composed characters by using the Viterbi algorithm, considering the composition as a Markov process. They did experiments on 200 radical categories covering 2, 154 loosely-constrained character classes from 200 writers and the matching rate obtained was 96.5% for radicals, resulting in a character recognition rate of 93.5%.

3.3 Stroke-Based Approaches

Stroke-based approaches first decompose Chinese characters into several kinds of strokes, four directional or eight directional (Fig. 16) for example. Then the problem is converted to how to recognize the shape, order, and position of the small set of strokes, in which MRF [32], HMMRF [33], statistical structure modeling [2] et al. are widely used. Figure 17 is an example of four directional stroke segments de/composition from [32].

For stroke extraction, several different strategies have been presented in literature, such as a region-based strategy which utilizes run-length information [29], a boundary-based strategy which utilizes contour information, a pixel-based strategy which detects feature-points on skeletons [25], and a filtering-based strategy which uses directional filters such as Gabor, SOGD filters to extract stroke segments [26, 27, 32]. Table 4 lists some results of stroke extraction in literature.

Table 3 Some results of holistic recognition

Year	Authors	Features	Classifier	Data set	Recognition percentage
2006	Fu Chang [55]	vector feature	Multi-Trees, Sub-Vec, SVM	ETL9B	96.59
2005	Hailong Liu etc. [41]	gradient feature	Multi-scheme enhanced MQDF	MNIST, ETL9B, HCL2000	99.54, 99.33, 98.56
2005	Jian-xiong Dong etc. [10]	directional feature	SVM	ETL9B	99.0
2002	Rui Zhang etc. [43]	directional features	MCE enhanced MQDF	3775 Chinese characters	97.93 on regular, 91.19 on cursive
2002	Yong Ge etc. [37–39]	Gabor features and their spatial derivatives	CDHMM	4616 characters in GB2312–80	96.34
2002	Bing Feng etc. [42]	Five features	HMM	3775 Chinese	94
2001	Yan Xiong etc. [56]	Five global features	Contextual stochastic model	50 pairs of similar Chinese	94
2000	Mingrui Wu etc. [44]	directional line element feature	Single CSN network	700 Chinese	95
2000	C.H. Wang etc. [45]	three kinds of statistical feature	MLP network	3775 Chinese in 4MSL	92.37
2000	B.H. Xiao etc. [48]	Three groups of features: CDF7, CDF8, SDF	Adaptive combination of classifiers	3775 Chinese in 4MSL	93
2000	Jiayong Zhang etc. [49]	Multi-scale directional feature	NSMDM	3775 Chinese	99.3 on regular, 88.4 on cursive
1999	Nei Kato etc. [40]	directional element feature	CBDD and AMD	ETL9B	99.42
1998	Y. Mizukami [53]	directional feature	displacement extraction	ETL8B	96

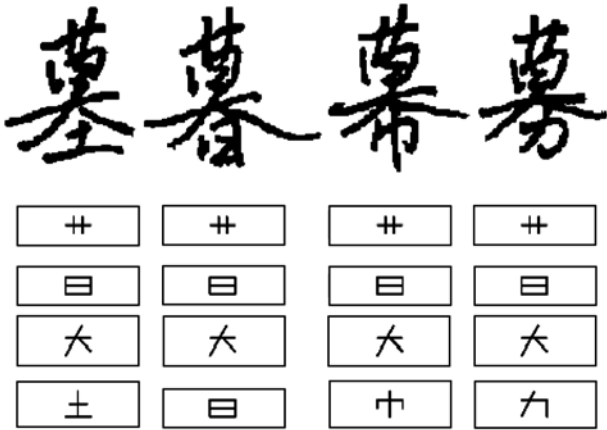
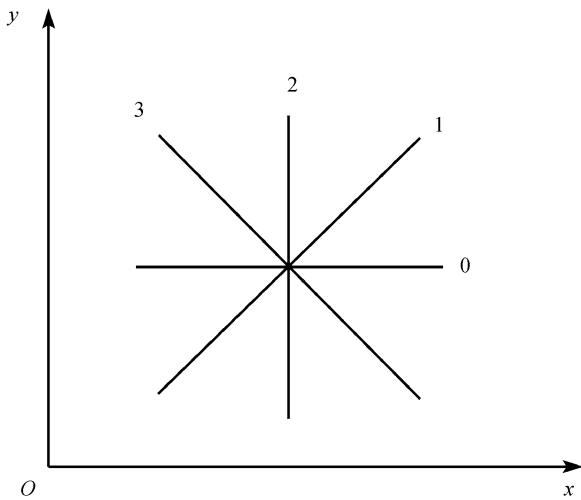
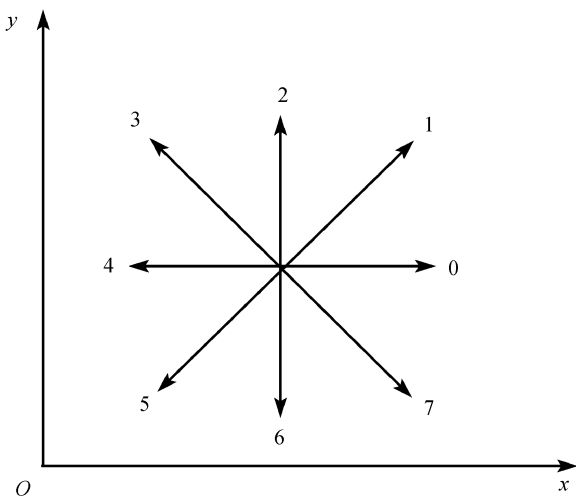


Fig. 15 Examples (from Ref. [1]) of similar characters (tomb, dusk, curtain and subscription) decomposed into radicals



(a)



(b)

Fig. 16 Two popular stroke direction definitions

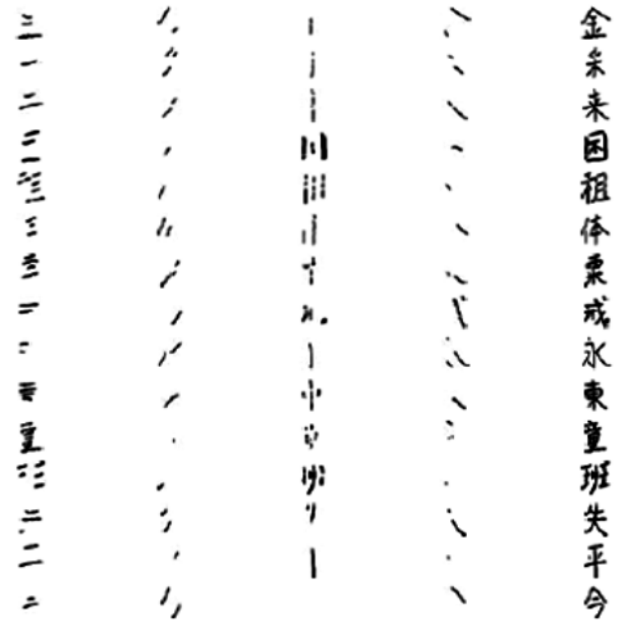


Fig. 17 From Ref. [32]: Four directional stroke segments de/composition

During recognition, the key is how to model each stroke and characterize relationships between strokes in a character.

Kim et al. proposed a statistical structure modeling method in [2]. They represent a character by a set of model strokes (manually designed structure), which are composed of a poly-line connecting K feature points (Fig. 18). Accordingly, a model stroke s_i is represented by the joint distribution of its feature points as $Pr(s_i) \equiv Pr(p_{i1}, p_{i2}, \dots, p_{ik})$. They assume the feature points have a normal distribution which can be described by the mean and the covariance matrix. Consequently, the structure composed of multiple strokes is represented by the joint distribution of the component strokes. For example, a structure S composed of two strokes s_i and s_j can be represented as $Pr(S) \equiv Pr(s_i, s_j)$. Readers are referred to [2] for more details.

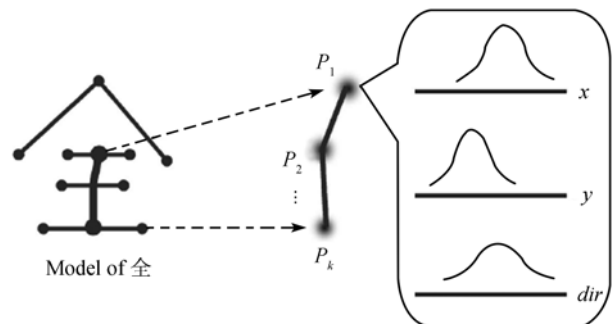


Fig. 18 Kim's statistical stroke model

Jia Zeng et al. proposed a statistical-structural scheme based on Markov random fields (MRFs). They represent the stroke segments by the probability distribution of their locations, directions and lengths. The states of MRFs encode these information from training samples and different character structures are represented by different state relation

Table 4 Some results of stroke extraction

Publication	Method	Data Set	Accuracy
2003 2004 Yih-Ming Su etc. [26,27]	Filtering-based: Gabor, SOGD	hand-printed in ETL8B and some self-collected	Good
2002 Feng Lin etc. [25]	Pixel-based: feature-points on skeletons	3039 characters, over 18000 strokes	99
2000 Ruini Cao etc. [28]	degree information, stroke continuation properety	111 printed characters, 849 strokes	93
2000 Kuo-Chin Fan etc. [29]	Region-based: run-length	2500 Hei fong characters	94.2
1999 Jin Wook Kim etc. [31]	mathematical morphology	1800 basic handprinted Chinese	97

Table 5 Some results of stroke-based recognition

Publication	Method	Data Set	Rec Rate
2005 Jia Zeng etc. [32]	MRF	9 pairs of similar Chinese in ETL9B	92 to 100
2003 In-Jung Kim etc. [2]	Statistical Character Structure Modeling	KAIST Hanja1 and Hanja2	98.45
2001 Cheng-Lin Liu etc. [36]	Model-based stroke extraction and matching	783 characters	97
2000 Qing Wang etc. [33]	HMMRF	470 Chinese from 3775	88

ships. Through encouraging and penalizing different stroke relationships by assigning proper clique potentials to spatial neighboring states, the structure information is described by a neighborhood system and multi-state clique potentials and the statistical uncertainty is modeled by single-state potentials for probability distributions of individual stroke segment. Then, the recognition problem can be formulated into the MAP-MRF framework to find an MRF that produces the maximum a posteriori for unknown input data. More details can be found in their paper [32].

Table 5 list some recognition results of stroke-based approaches.

3.4 Word recognition

Chinese words are generally short. Word recognition actually goes hand-in-hand with character recognition; attempting to recognize words can actually help recognize characters as well. Several approaches use association rules to improve Chinese handwritten character recognition. Word recognition has the additional difficulty that in Chinese handwriting the words are not separated by spaces. Since this is the case, word boundaries can only be deduced by the semantic relationship of characters. Chinese has other unique issues for word recognition as well, such as storing a lexicon for easy access. In languages such as English, this is fairly straightforward due to the few number of potential characters. In Chinese, however, with many possibilities at each transition, the best data structures to use are not at all obvious.

Most word recognition research in literature has focused on address [57–59] or bank check [60] recognition. Chunheng Wang [57] proposed a handwritten Chinese address recognition (HCAR) system, and got a total recognition rate of 96.45% on 600 address images. Their system takes advantage of Chinese address knowledge, and applies key character extraction and holistic word matching to solve the problem. Hanshen Tang et al. [60] presented the spiral recognition methodology with its application in unconstrained handwritten Chinese legal amount recognition. On a database consisting of a training set of 47 800 real bank checks

and a test set of 12,000, the recognition rate at the character level was 93.5%, and the recognition rate at the legal amount level was 60%. Combined with the recognition of the courtesy amount, the overall error rate is less than 1%.

3.5 Classifier combination

One effective approach to improving the performance of handwriting recognition is to combine multiple classifiers. However, it's not a trivial problem to collect votes from different classifiers while adopting proper voting weight strategy.

Xiaofan Lin et al. [61] proposed a classifier combination method. It consists of three main components: adaptive confidence transform (ACT), consensus theoretic combination and reliability-based speedup scheme. Their experimental results showed a significant reduction of error rates.

K.Y. Hung etc. [62] proposed an application of multiple classifiers for the detection of recognition errors. Their results showed that the recall rate of recognition errors, the precision rate of recognition error detection and the savings in manual effort were better than the corresponding performance using a single classifier or a simple threshold detection scheme.

3.6 Bibliographic remarks

3.6.1 Holistic character recognition

Tianlei Wu et al. [63] presented a feature extraction method for handwritten Chinese character recognition. To compensate for border phenomena, overlap and fuzzy membership are introduced into elastic meshing by mapping the input image to a virtual normalized image on which these attributes are predefined. Furthermore, the global overlapped elastic meshing (GOEM) is extended to a hierarchical framework to adapt to bad alignments and local distortions in handwritten Chinese characters.

Yong Ge et al. [37, 39] used four Gabor features and their spatial derivatives to parameterize the whole character image in both horizontal and vertical directions. They then use

linear discriminant analysis for data reduction and two continuous-density hidden Markov models (CDHMMs) to model the characters in the two (horizontal and vertical) directions. Gabor filters are spaced every 8 pixels, and the CDHMMs are trained using a minimum classification error criterion. They support 4616 characters, which include 4516 simplified Chinese characters in GB2312-80, 62 alphanumeric characters, 38 punctuation marks and symbols. By using 1 384 800 character samples to train the recognizer, an averaged character recognition accuracy of 96.34% was achieved on a testing set of 1 025 535 character samples.

Hailong Liu et al. [41] presented a handwritten character recognition system. High-resolution gradient feature, extracted from the character image, then minimum classification error training, together with similar character discrimination by compound Mahalanobis function are used to improve the recognition performance of the baseline modified quadratic discriminant function. For handwritten digit recognition, they use the MNIST database, which consists of 60 000 training images and 10, 000 testing images, all size-normalized and centered in a 28x28 box. For handwritten Chinese character recognition, they use ETL9B and HCL2000 databases. ETL9B contains binary images of 3, 036 characters, including 2 965 Chinese characters and 71 hiragana characters, and each of them with 200 binary images. They use the first 20 and last 20 images of each character for testing, and the remaining images are used as training samples. HCL2000 contains 3 755 frequently used simplified Chinese characters written by 1 000 different people. They use 700 sets for training and the remaining 300 sets for testing. The highest recognition accuracy they achieved on MNIST test set was 99.54%, and 99.33% on the ETL9B test set. On HCL2000, the 98.56% testing recognition accuracy they achieved is the best in literature to their knowledge. Recently, they [64] summarized their efforts on offline handwritten script recognition using a segmentation-driven approach.

Lei Huang et al. [65] proposed a scheme for multiresolution recognition of offline handwritten Chinese characters using wavelet transform. The wavelet transform enables an invariant interpretation of character image at different basic levels and presents a multiresolution analysis in the form of coefficient matrices. Thus, it makes the multiresolution recognition of character possible and gives a new view to handwritten Chinese character recognition. Their experiments were carried out on 3 755 classes of handwritten Chinese characters. There were 50 samples in each class, 187, 750 in total, used for testing, and 600 samples each, 2 253 000 in total, used for training. Using D-4 wavelet function, a recognition accuracy rate of 80.56% was achieved.

Yan Xiong et al. [56] studied using a discrete contextual stochastic model for handwritten Chinese character recognition. They tested on a vocabulary consisting of 50 pairs of highly similar handwritten Chinese characters, and achieved a recognition rate around 98 percent on the training set (closed-test) and about 94 percent on the testing set (open-test).

Bing Feng et al. [42] applied HMM to recognition of handwritten Chinese. The image is segmented into a number of local regions. They used 3755 Chinese characters, 1300 sets for training and 100 for testing. They got about 95% recognition rate on training, and about 94% on testing.

Rui Zhang et al. [43] improved the MQDF (modified quadratic discriminant function) performance by the MCE (minimum classification error) based discriminative training method. The initial parameters of the MQDF were estimated by the ML estimator, and then the parameters were revised by the MCE training. The MCE-based MQDF achieves higher performance for both numbers and Chinese characters, and keeps the same computational complexity. Their recognition accuracy reached 97.93% on regular and 91.19% on cursive characters.

Mingrui Wu et al. [44] proposed a geometrical approach for building neural networks. It is then easy to construct an efficient neural classifier to solve the handwritten Chinese character recognition problem as well as other pattern recognition problems of large scale. In the experiment, 700 Chinese characters corresponding to the 700 smallest region-position codes are chosen. For each character there are 130 samples among which 70 samples are randomly selected for training and the remaining 60 for testing. The recognition rate was over 95%.

C.H. Wang et al. [45] proposed an integration scheme with multi-layer perceptron (MLP) networks to solve the handwritten Chinese character recognition problem. In the experiments, the 3 755 categories in the first level of national standard GB2312-80 are considered, and the 4-M Samples Library (4MSL) is selected as the database. There are 100 samples for training and 100 for testing, for each character. They achieved a 92.73% recognition rate.

Chris K.Y. Tsang et al. [47] developed an SDM (Structural Deformable Model) which explicitly takes structural information into its modeling process and is able to deform in a well-controlled manner through an ability of global-to-local deformation. As demonstrated by the experiments on Chinese character recognition, SDM is a very attractive post-recognizer. SDM is developed in [48].

B.H. Xiao et al. [48] proposed an adaptive classifier combination approach, motivated by the idea of metasynthesis. Compared with previous integration methods, parameters of the proposed combination approach are dynamically acquired by a coefficient predictor based on a neural network and vary with the input pattern. It is also shown that many existing integration schemes can be considered as special cases of the proposed method. They tested on the 3 755 most frequently used Chinese characters. For each class, 300 samples are picked up from the 4-M Samples Library (4MSL), where 200 were for training and 100 for testing. They got about 97% training accuracy and 93% testing.

Jiayong Zhang et al. [49] proposed a multi-scale directional feature extraction method called MSDFE, which combines feature detection and compression into an integrated optimization process. They also introduced a structure into the Mahalanobis distance classifier and proposed

the NSMDM (Nested-subset Mahalanobis Distance Machine). The system using the methods proposed achieves a high accuracy of 99.3% on regular and an acceptable one of 88.4% on cursive with severe structural distortions and stroke connections.

T. Shioyama et al. [50] proposed an algorithm for recognition of hand-printed Chinese characters, which absorbs the local variation in hand printing by the wavelet transform. The algorithm was applied to hand-printed Chinese characters in the ETL9 database. The algorithm attains a recognition rate of 97.33%. In another paper [52], they used 2D Fourier power spectra instead of wavelet transform, and the recognition rate reached 97.6%.

Din-Chang Tseng et al. [51] proposed an invariant handwritten Chinese character recognition system. Characters can be in arbitrary location, scale, and orientation. Five invariant features were employed: number of strokes, number of multi-fork points, number of total black pixels, number of connected components, and ring data.

Nei Kato et al. [40] proposed city block distance with deviation (CBDD) and asymmetric Mahalanobis distance (AMD) for rough classification and fine classification in recognition. Before extracting the directional element feature (DEF) from each character image, transformation based on partial inclination detection (TPID) is used to reduce undesired effects of degraded images. Experiments were carried out with the ETL9B. Every 20 sets out of the 200 sets of the ETL9B is considered as a group, thus 10 groups were made in total. In rotation, nine groups were used as the training data, and the excepted group was employed as test data. The average recognition rate reached to 99.42%.

Jian-xiong Dong et al. [10] described several techniques improving a Chinese character recognition system: enhanced nonlinear normalization, feature extraction and tuning kernel parameters of support vector machine. They carried experiments on the ETL9B database, using directional features, and achieved a high recognition rate of 99.0%.

Y. Mizukami [53] proposed a recognition system using displacement extraction based on directional features. In the system, after extracting the features from an input image, the displacement is extracted by the minimization of an energy functional, which consists of the Euclidean distance and the smoothness of the extracted displacement. Testing on ETL-8B containing 881 categories of handwritten Chinese characters, a recognition rate over 96% was achieved, an improvement of 3.93% over simple matching.

3.6.2 Radical-based character recognition

An-Bang Wang et al. [24] proposed a recursive hierarchical scheme for radical extraction which combined structural method and statistical method. It included three modules: the character pattern detection module, straight cut line detection module, and stroke clustering module. The experiments were conducted on 1056 constrained handwritten characters. The successful rate of radical extraction is 97.1%. In [66] they proposed a radical-based OCR system. A recursive hierarchical scheme was developed to perform radical extraction first. Then, character features and radical features

were extracted for matching. Finally, a hierarchical radical matching scheme was devised to identify the radicals embedded in an input Chinese character and recognize the input character accordingly. Experiments for radical extraction were conducted on 1856 characters. The successful rate of radical extraction is 92.5%. Experiments for the matching process were conducted on two sets: a training set and testing set, each set including 900 characters. The overall recognition rate was 98.2% and 80.9% (for the training set and testing set, respectively).

Daming Shi et al. [1, 19–23] proposed an approach to radical recognition for offline handwritten Chinese characters, based on nonlinear active shape modeling, called active radical modeling (ARM). Experiments for (writer-independent) radical recognition were conducted on 200 radical categories covering 2 154 loosely-constrained character classes from 200 writers, using kernel PCA to model variations around mean landmark values. The matching rate obtained was 96.5% correct radicals. Accordingly, they proposed a subsequent stage of character recognition based on Viterbi decoding with respect to a lexicon, resulting in a character recognition rate of 93.5%.

3.6.3 Stroke-based character recognition

Feng Lin et al. [25] presented a scheme that can rapidly and accurately extract the strokes from the thinned Chinese character images. By removing two types of bug pixels in a fork region, they directly use the Rutoviz number to detect all fork point in the character skeleton. Then using a bi-directional graph, they connect the stroke segments at a fork point with high accuracy. On a large data set with over 18 000 character strokes (3 039 Chinese characters from 10 writers), they achieved over 99% accuracy for stroke extraction.

Yih-Ming Su et al. presented a directional stroke extraction approach based on a directional filtering technique for Chinese character recognition. In [27], first, a set of Gabor filters is used to break down an image of a character into different directional features. Next, a new iterative thresholding technique that minimizes the reconstruction error is proposed to recover stroke shape. Finally, a refinement process based on measuring the degree of stroke overlap is used to remove redundant stroke pieces. In [26], the proposed filtering technique uses a set of the second-order Gaussian derivative (SOGD) filters to decompose a character into a number of stroke segments. Moreover, a Gaussian function is used to extract the stroke segments along arbitrary orientations. The system is applied to a database set of 956 hand-printed character categories from ETL8b, and 200 hand-printed Chinese character categories from the authors' collection, where each category contained 100 sample characters. Out of 100 sample characters, 50 were used as training data, and the remainder was used as test data. The recognition performance from the application of the decomposition process can be improved about 17.31% compared to the whole-based approach, in the test character set.

Ruini Cao et al. [28] proposed a model of stroke extrac-

tion. They use the degree information and the stroke continuation property to tackle two major problems: to extract the primary strokes and the segmentation ambiguities at intersection points. The proposed model can be used to extract strokes from both printed and handwritten character images. The test was carried on 111 printed Chinese characters. There were about 849 strokes in total. The model segmented 806 strokes correctly. The accuracy was 93%. There were about 98 characters whose strokes were all segmented correctly. The correct rate is 88%.

Kuo-Chin Fan et al. [29] proposed a run length-based stroke extraction approach without using the thinning method to avoid distortions around junction areas generated by the method. They applied the proposed method to 2500 Hei fong characters. The rate for correctly extracting the strokes of all characters is 94.2%.

Jin Wook Kim et al. [31] proposed a method of decomposing a Chinese character into a set of strokes, based on the concepts of mathematical morphology.

Jia Zeng et al. [32] used 2-D Gabor filters to extract directional stroke segments from images of Chinese characters, where each stroke segment is associated with a state in Markov random field models. Extensive experiments on similar characters have been carried out on the ETL9B Japanese database. In this context, 2 965 different Chinese characters (Kanji) and 71 kinds of Japanese characters (hiragana), called the first class of Japanese Industrial Standard (JIS), are included. The characters have been written by 4 000 writers. There are 200 samples of each character, so that 607 200 samples are included in the ETL9B. Nine pairs of highly similar Chinese characters are used as the recognition vocabulary to study the effectiveness of Markov random fields for Chinese characters modeling. Each character is written by 200 writers with 150 of them used for training and the remaining 50 samples for testing. Two mixtures are used in the MRF model. The recognition rate ranges from 92% to 100% for each of the 18 characters.

In-Jung Kim et al. [2] proposed a statistical character structure modeling method. It represents each stroke by the distribution of the feature points. The character structure is represented by the joint distribution of the component strokes. Based on the character representation, a stroke neighbor selection method is also proposed. The proposed character modeling method was applied to a handwritten Chinese character recognizer. They did experiments on KAIST Hanja1 and Hanja2 databases. The Hanja1 database has 783 most frequently used classes. For each class, 200 samples, artificially collected from 200 writers, are contained. The Hanja2 database has 1 309 samples collected from real documents. The overall recognition rate was 98.45%.

Qing Wang et al. [33] presented a Hidden Markov Mesh Random Field (HMMRF) based approach using statistical observation sequences embedded in the strokes of a character. In their experiments, 470 characters (from the 3 755 most frequently used Chinese) were used, with each character having 100 samples among which 60 samples were used for the training and 40 samples for testing. They got about

93% accuracy on training and 88% on testing.

Zen Chen et al. [34] proposed a set of rules for stroke ordering for producing a unique stroke sequence for Chinese characters. For most of the everyday Chinese characters, the sequence of strokes with their stroke type identity can tell which character is written. The proposed method can pre-classify Chinese characters well.

In-Jung Kim et al. [35] suggested that the structural information should be utilized without suffering from the faults of stroke extraction. Then they proposed a stroke-guided pixel matching method to satisfy the requirement.

Cheng-Lin Liu et al. [36] proposed a model-based structural matching method. In the model, the reference character of each category is described in an attributed relational graph (ARG). The input character is described with feature points and line segments. The structural matching is accomplished in two stages: candidate stroke extraction and consistent matching. They tested on 783 Chinese characters, and got over 97% recognition accuracy.

3.6.4 Word recognition

Qing Fu et al. [59] presented an efficient method of Chinese handwritten address character string segmentation and recognition. First, an address string image is presegmented into several radicals using stroke extraction and stroke merging. Next, the radical series obtained by pre-segmentation merge into different character image series according to different merging paths. After that, the optimal merging path is selected using recognition and semantic information. The recognition information is given by the character classifier. The semantic information is obtained from an address database that contains 180, 000 address items. Finally, the optimal recognition results of the character image series that are combined by radical series according to the optimal merging paths are obtained. In experiments on 897 mail images, the proposed method achieves a correct rate of 85%.

Zhi Han et al. [58] proposed a two-stage handwritten Chinese character segmentation approach in mail address recognition. On more than 500 real envelope images, the correct sorting rate of address recognition was up to 79.46% and that of address and postcode integrated recognition up to 96.26%.

Yuan-Xiang Li et al. [67,68] investigated several language models in the contextual post-processing of Chinese script recognition. They gave the proposal in choosing a suitable language model according to the requirement of a practical recognition system. In [67], they presented two methods based on a confusion matrix to recall the correct characters.

Chunheng Wang et al. [57] proposed a handwritten Chinese address recognition (HCAR) system. The proposed approach takes advantage of Chinese address knowledge, and applies key character extraction and holistic word matching to solving the problem. Unlike the conventional approach, the proposed approach can avoid character segmentation errors. The entire test dataset includes 600 test address images, which were categorized into three subsets according to writing quality, good, normal and bad. Each

subset contains 200 images respectively. The recognition rate is 100% for good, 98.31% for normal and 91.04% for bad. The total rate is 96.45%, which is much better than the 62.60% obtained using the conventional approach.

Hanshen Tang et al. [60] presented the spiral recognition methodology with its application in unconstrained hand-written Chinese legal amount recognition. It first describes the failed application of a neural network--a hidden Markov model hybrid recognizer on Chinese bank check legal amount recognition--then shows that the spiral methodology enables the system to increase its recognition power during the training iterations. They experimented on the database of a training set of 47.8 thousand real bank checks, and a test set of 12,000. The recognition rate at the character level is 93.5%, and the recognition rate at the legal amount level is 60%. Combined with the recognition of courtesy amount, the overall error rate is less than 1%.

3.7 Technology gaps

To advance offline Chinese handwriting recognition, many issues still need to be solved. Based on current techniques and the application expectation, the following are some of the most interesting and challenging topics. By solving all or some of them, many applications can be realized and related commercial products developed.

3.7.1 Cursive character recognition

In Section 3, we reviewed the work on individual character recognition as well as the corresponding researchers. Although some of them have achieved very promising results on regular or hand-print scripts, recognition of cursive script is still unacceptable. The proposed approaches can neither capture features distinctively nor extract sub-structures (radicals or strokes) precisely for cursive characters, which then makes accurate recognition impossible.

Also, segmentation and recognition are very hard for low quality document images with interference by noise such as lines, stamps or images with great distortion. Tseng et al. [12] presented a method for removing interfering-lines from interfered-characters. However, a lot more need to be done in order to handle a large volume of "bad" images.

3.7.2 Word spotting

In Chinese handwriting, words are not separated by spaces and word spotting can only rely on the semantic relationship of characters. However, it is a big issue to store a Chinese lexicon for easy access. The reason is there are many characters in Chinese, and with many possibilities at each transition, the best data structures to use are not at all obvious.

In Section 3.4, we reviewed some work on word recognition. Most researchers have focused on address or bank check recognition because it is relatively easier to build such a word lexicon. Similarly, other specialty word lexicons can be established in which the number of words are practically countable and most words are used frequently. For a long-term purpose, to finally handle word spotting, a fine

strategy to structure a lexicon is required.

3.7.3 Document decomposition

A real world document image may consist of text (hand-written or machine-printed), line drawings, tables, diagrams, pictures, icons, etc. To efficiently recognize the whole image, which means trying to utilize state-of-the-art on the individual component of the recognition process, It is necessary to decompose a document into its component parts. Basically, deriving the blocks with certain interest can take advantage of the fact that the structural elements of a document are generally laid down in rectangular blocks aligned. Knowledge of visual, spatial and linguistics can also be used. Alternatively, for particular tasks, the blocks may be predefined located, such as images on name cards. However, for "bad" script, the information may be not available, and it can be even harder if blocks are overlapped.

3.7.4 Hand-drawn diagram recognition

Hand-drawn diagram recognition in Chinese is also similar to that in other languages, such as in English, so few researchers deal with it specifically in Chinese handwriting recognition. However, it would be helpful if we deal with the problem by considering the language along with hand-drawn diagrams. For example, in a Chinese handwriting document, the hand drawn diagram recognition can utilize the recognition of Chinese characters along or inside the diagram.

3.7.5 Classifier combination

In Section 3.5, we listed two groups of works on classifier combination and showed some development in this area. However, it is still not obvious how a combination strategy can fully or mostly utilize the power of sub-classifiers, and how to deal with the trade-off between combination and efficiency. Most importantly, it is difficult to explore the power of each sub-classifier according to different groups of characters and to balance the weight between them.

3.7.6 Phone number, date and address recognition

Phone number and date recognition have not received much attention in Chinese handwriting recognition, partly because the problem is similar to the one in English. Some researchers have tried to recognize Chinese along with numbers, such as in Ref. [41] and some applications in bank check recognition.

Many researchers have concentrated on street address recognition for word recognition (Section 3.4). In Chinese handwriting, the big issue is how to accurately segment address names according to an appropriate lexicon. Many approaches try to combine postal code and address recognition together to guarantee the accuracy. The methods to purely recognize the address is still difficult, especially for cursive handwriting. Also, mix-language address recognition should be paid attention to, as international communication advances.

3.7.7 Others

More technology gaps may exist in aspects of language-independent [69] and segmentation-free recognition, and handwriting retrieval, which involves searching a repository of handwritten documents for those most similar to a given query image.

4 Conclusions

Offline handwriting Chinese recognition (CHR) is a difficult pattern recognition problem because the number of characters (classes) is very high. Many characters have very complex structures and some of them look very similar, with accurate segmentation very difficult under many situations.

Generally, the recognition process follows a similar path: document pages are first fed to a general image preprocessing unit to enhance the image quality and get a proper representation, then segmentation and normalization algorithms are applied to get normalized individual characters/words that are ready for input by classifiers. During classification, different approaches have been presented: holistic approaches extract features from the entire character then recognize them by match of the feature vectors; radical-based approaches first extract predefined radicals, then try to optimize the combination of radical sequences; stroke-based approaches decompose characters into strokes then recognize them according to the shape, order, and position of the set of strokes. Most research on CHR in literature has been concentrated on the following aspects: segmentation and normalization algorithms, feature extraction strategies, and design of models and classifiers.

In Chinese handwriting, characters are usually separated, but no additional space is between words, and it is still unclear as to how to model a lexicon for Chinese words. So usually, segmentation algorithms aim to output individual characters. However, previous work shows that if completely separated from recognition, segmentation itself cannot give very promising results, and may become a bottleneck of the whole system. It is often beneficial to consult semantic information while segmenting.

During recognition, currently, radical-based and stroke-based approaches have a lower rate than holistic ones. The reason is that the extraction of radicals and strokes is not easy itself, especially considering radicals and strokes are connected within characters and may cause segmentation errors. However, the characteristics of Chinese characters are mostly based on strokes, therefore it's recommended that when extracting holistic features, to focus on stroke aspects as well. The most promising features in the literature are gradient and directional features that aim to extract features based on stroke orientation and position.

Currently, word recognition methods are constrained in such areas as address or check recognition because it is relatively easier to build such a word lexicon and there are market application needs. Word recognition can be used as post-recognition to adjust previous character recognition, but

the semantic information is not easy to collect and the database is hard to establish. Much can be gained from on-line handwriting recognition techniques that have been widely implemented into commercial products, such as PCs, PDAs, cellphones and other popular software.

References

1. Shi D M, Damper R I, Gunn S R. Offline handwritten Chinese character recognition by radical decomposition. *ACM Transactions on Asian Language Information Processing*. 2003, 1: 27–48
2. Kim I J, Kim J H. Statistical character structure modeling and its application to handwritten Chinese character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2003, 25: 1422–1436
3. Lu Y, Shridhar M. Character segmentation in handwritten words—An overview. *Pattern Recognition*. 1996, 29: 77–96
4. Wei X H, Ma S P, Jin Y J. Segmentation of connected Chinese characters based on genetic algorithm. In: *ICDAR'05: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, Seoul, Korea, Vol 1. IEEE Computer Society, 2005, 645–649
5. Liang Z Z, Shi P F. A metasynthetic approach for segmenting handwritten Chinese character strings. *Pattern Recognition Letters*. 2005, 26: 1498–1511
6. Zhao S Y, Chi Z R, Shi P F, et al. Two-stage segmentation of unconstrained handwritten Chinese characters. *Pattern Recognition*, 2003, 36: 145–156
7. Zhao S Y, Chi Z R, Shi P F, et al. Handwritten Chinese character segmentation using a two-stage approach. In: *Proceedings of the sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, Seattle, WA, Vol 1. IEEE Computer Society, 2001, 179–183
8. Tseng Y H, Lee H J. Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. *Pattern Recognition Letters*. 1999, 20: 791–806
9. Tseng L Y, Chen R C. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. *Pattern Recognition Letters*. 1998, 19: 963–973
10. Dong J X, Krzyzak A, Suen C Y. An improved handwritten Chinese character recognition system using support vector machine. *Pattern Recognition Letters*. 2005, 26: 1849–1856
11. Liu C L, Marukawa K M. Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition. *Pattern Recognition*. 2005, 38: 2242–2255
12. Tseng Y H, Lee H J. Interfered-character recognition by removing interfering-lines and adjusting feature weights. In: *Proceedings of Fourteenth International Conference on Pattern Recognition*. 1998, Vol 2. 1865–1867
13. Chiu H P, Tseng D C. A feature-preserved thinning algorithm for handwritten Chinese characters. In: *Proceedings of 13th International Conference on Pattern Recognition*, 1996, Vol 3. 235–239
14. Gao J, Ding X q, Wu Y S. A segmentation algorithm for handwritten Chinese character strings. In: *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99)*, India, Vol 1. IEEE Computer Society. 1999, 633–636
15. Wang Q, Chi Z R, Feng D D, et al. Match between normalization

- schemes and feature sets for handwritten Chinese character recognition. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition(ICDAR'01), Seattle, Vol 1. IEEE Computer Society, 2001, 551-555
16. Liu C L, Sako H, Fujisawa H. Handwritten Chinese character recognition: alternatives to nonlinear normalization. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition(ICDAR'03), Scotland, Vol 1. IEEE Computer Society, 2003, 524-528
 17. Liu C L, Marukawa K. Global shape normalization for handwritten Chinese character recognition: a new method. In: Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, Vol 1. 2004, 300-305
 18. Liu C L. Handwriting Chinese character recognition: Effects of shape normalization and feature extraction. In: Summit on Arabic and Chinese Handwriting, College Park, USA. 2006, 13
 19. Shi D, Gunn S R, Damper R I. Handwritten Chinese radical recognition using nonlinear active shape models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25: 277-280
 20. Shi D M, Gunn S R, Damper R I. Handwritten Chinese character recognition using nonlinear active shape models and the Viterbi algorithm. Pattern Recognition Letters, 2002, 23: 1853-1862
 21. Ng G S, Shi D, Gunn S R, et al. Nonlinear active handwriting models and their applications to handwritten Chinese radical recognition. In: ICDAR'03: Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, Scotland, Vol 1. IEEE Computer Society, 2003, 534-538
 22. Shi D, Gunn S R, Damper R I. A radical approach to handwritten Chinese character recognition using active handwriting models. In: Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1. 2001, 670
 23. Shi D, Gunn S R, Damper R I. Active radical modelling for handwritten Chinese characters. In: ICDAR'01: Proceedings of the sixth International Conference on Document Analysis and Recognition, Seattle, WA, Vol 1. IEEE Computer Society, 2001, 236-240
 24. Wang A B, Fan K C, Wu W H. A recursive hierarchical scheme for radical extraction of handwritten Chinese characters. In: Proceedings of 13th International Conference on Pattern Recognition, Vol 3. 1996, 240-244
 25. Lin F, Tang X O. Off-line handwritten Chinese character stroke extraction. In: Proceedings of 16th International Conference on Pattern Recognition, Vol 3. 2002, 249-252
 26. Su Y M, Wang J F. Decomposing Chinese characters into stroke segments using SOGD filters and orientation normalization. In: Proceedings of the 17th International Conference on Pattern Recognition, Vol 2. 2004, 351-354
 27. Su Y M, Wang J F. A novel stroke extraction method for Chinese characters using Gabor filters. Pattern Recognition, 2003, 36: 635-647
 28. Cao R N, Tan C L. A model of stroke extraction from Chinese character images. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 4. 2000, 368-371
 29. Fan K C, Wu W H. A run-length coding based approach to stroke extraction of Chinese characters. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 565-568
 30. Chiu H P, Tseng D C. A novel stroke-based feature extraction for handwritten Chinese character recognition. Pattern Recognition, 1999, 32: 1947-1959
 31. Kim J W, Kim K I, Choi B J, et al. Decomposition of Chinese character into strokes using mathematical morphology. Pattern Recognition Letters, 1999, 20: 285-292
 32. Zeng J, Liu Z Q. Markov random fields for handwritten Chinese character recognition. In: ICDAR'05: Proceedings of the Ninth International Conference on Document Analysis and Recognition, Korea, Vol 1. IEEE Computer Society, 2005, 101-105
 33. Wang Q, Chi Z r, Feng D D, et al. Hidden Markov random field based approach for off-line handwritten Chinese character recognition. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 347-350
 34. Chen Z, Lee Ch W, Cheng R H. Handwritten Chinese character analysis and preclassification using stroke structural sequence. In: Proceedings of 13th International Conference on Pattern Recognition, Vol 3. 1996, 89-93
 35. Kim I J, Liu C L, Kim J H. Stroke-guided pixel matching for handwritten Chinese character recognition. In: ICDAR'99: Proceedings of the Fifth International Conference on Document Analysis and Recognition, India, Vol 1. IEEE Computer Society, 1999, 665-668
 36. Liu C L, Kim I J, Kim J H. Model-based stroke extraction and matching for handwritten Chinese character recognition. Pattern Recognition, 2001, 34: 2339-2352
 37. Ge Y, Huo Q. A comparative study of several modeling approaches for large vocabulary offline recognition of handwritten Chinese characters. In: Proceedings of 16th International Conference on Pattern Recognition, Vol 3. 2002, 85-88
 38. Ge Y, Huo Q. A study on the use of CDHMM for large vocabulary off-line recognition of handwritten Chinese characters. In: Proceedings of Eighth International Workshop on Frontiers in Handwriting Recognition, Canada, Vol 1. 2002, 334-338
 39. Ge Y, Huo Q, Feng Z D. Offline recognition of handwritten Chinese characters using Gabor features, CDHMM modeling and MCE training. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol 1. 2002, I-1053-I-1056
 40. Kato N, Suzuki M, Omachi S, et al. A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21: 258-262
 41. Liu H L, Ding X Q. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, Korea, Vol 1. IEEE Computer Society, 2005, 19-23
 42. Feng B, Ding X Q, Wu Y S. Chinese handwriting recognition using hidden Markov models. In: Proceedings of 16th International Conference on Pattern Recognition, Vol 3. 2002, 212-215
 43. Zhang R, Ding X Q. Minimum classification error training for handwritten character recognition. In: Proceedings of 16th International Conference on Pattern Recognition, Vol 1. 2002, 580-583
 44. Wu M R, Zhang B, Zhang L. A neural network based classifier for handwritten Chinese character recognition. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 561-564
 45. Wang C H, Xiao B H, Dai R W. A new integration scheme with multilayer perceptron networks for handwritten Chinese character recognition. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 961-964
 46. Tsang C K Y, Chung F L. Development of a structural deformable model for handwriting recognition. In: Proceedings of Fourteenth

- International Conference on Pattern Recognition, Vol 2. 1998, 1130-1133
47. Tsang C K Y, Chung F L. A structural deformable model with application to post-recognition of handwriting. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 129-132
 48. Xiao B H, Wang C H, Dai R W. Adaptive combination of classifiers and its application to handwritten Chinese character recognition. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 327-330
 49. Zhang J Y, Ding X Q, Liu C S. Multi-scale feature extraction and nested-subset classifier design for high accuracy handwritten character recognition. In: Proceedings of 15th International Conference on Pattern Recognition, Vol 2. 2000, 581-584
 50. Shioyama T, Wu H Y, Nojima T. Recognition algorithm based on wavelet transform for handprinted Chinese characters. In: Proceedings of Fourteenth International Conference on Pattern Recognition, Vol 1. 1998, 229-232
 51. Tseng D C, Chiu H P. Fuzzy ring data for invariant handwritten Chinese character recognition. In: Proceedings of 13th International Conference on Pattern Recognition, Vol 3. 1996, 94-98
 52. Shioyama T, Hamanaka J. Recognition algorithm for handprinted Chinese characters by 2D-FFT. In: Proceedings of 13th International Conference on Pattern Recognition, Vol 3. 1996, 225-229
 53. Mizukami Y. A handwritten Chinese character recognition system using hierarchical displacement extraction based on directional features. *Pattern Recognition Letters*, 1998, 19: 595-604
 54. Guo F J, Zhen L X, Ge Y, et al. An efficient candidate set size reduction method for coarse-classifier of Chinese handwriting recognition. In: Summit on Arabic and Chinese Handwriting, College Park, USA. 2006, 41-46
 55. Fu C. Techniques for solving the large-scale classification problem in Chinese handwriting recognition. In: Summit on Arabic and Chinese Handwriting, College Park, USA. 2006, 87-92
 56. Xiong Y, Huo Q, Chan C K. A discrete contextual stochastic model for the off-line recognition of handwritten Chinese characters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23: 774-782
 57. Wang C H, Hotta Y, Suwa M, et al. Handwritten Chinese address recognition. In: Proceedings of Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, Vol 1. 2004, 539-544
 58. Han Z, Liu C P, Yin X C. A two-stage handwritten character segmentation approach in mail address recognition. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, Korea, Vol 1. IEEE Computer Society, 2005, 111-115
 59. Fu Q, Ding X Q, Liu C S, et al. A hidden Markov model based segmentation and recognition algorithm for Chinese handwritten address character strings. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, Seoul, Korea, Vol 2. IEEE Computer Society, 2005, 590-594
 60. Tang H S, Augustin E, Suen C Y, et al. Spiral recognition methodology and its application for recognition of Chinese bank checks. In: Proceedings of Ninth International Workshop on Frontiers in Handwriting Recognition, 2004, Tokyo, Japan, Vol 1. 2004, 263-268
 61. Lin X f, Ding X q, Chen M, et al. Adaptive confidence transform based classifier combination for Chinese character recognition. *Pattern Recognition Letters*, 1998, 19: 975-988
 62. Hung K Y, Luk R W P, Yeung D S, et al. A multiple classifier approach to detect Chinese character recognition errors. *Pattern Recognition*, 2005, 38: 723-738
 63. Wu T L, Ma S P. Feature extraction by hierarchical overlapped elastic meshing for handwritten Chinese character recognition. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, Scotland, Vol 1. IEEE Computer Society, 2003, 529-533
 64. Ding X Q, Liu H L. Segmentation-driven offline handwritten Chinese and Arabic script recognition. In: Summit on Arabic and Chinese Handwriting, 2006, College Park, USA. 2006, 61-73
 65. Huang L, Huang X. Multiresolution recognition of offline handwritten Chinese characters with wavelet transform. In: Proceedings of the sixth International Conference on Document Analysis and Recognition, Seattle, WA, Vol 1. IEEE Computer Society, 2001, 631-634
 66. Wang A B, Fan K C. Optical recognition of handwritten Chinese characters by hierarchical radical matching method. *Pattern Recognition*, 2001, 34: 15-35
 67. Li Y X, Tan C L. An empirical study of statistical language models for contextual post-processing of Chinese script recognition. In: Proceedings of Ninth International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, Vol 1. 2004, 257-262
 68. Li Yuan-Xiang, Tan Chew Lim. Influence of language models and candidate set size on contextual post-processing for Chinese script recognition. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004, Vol 2. 2004, 537-540
 69. Natarajan P, Saleem S, Prasad R, et al. Multi-lingual offline handwriting recognition using Markov models. In: Summit on Arabic and Chinese Handwriting, 2006, College Park, USA. 2006, 177-187