A Bayesian-based probabilistic model for unconstrained handwritten offline Chinese text line recognition

Nanxi Li, Lianwen Jin
School of Electronic and Information Engineering
South China University of Technology
Guangzhou 510641, China
pumpkinLNX@gmail.com, lianwen.jin@gmail.com

***Abstract*—A Bayesian-based probabilistic model is presented for unconstrained handwritten offline Chinese text line recognition. After pre-segmentation of a text line, plenty of invalid characters are produced which heavily interfere in the process of text line recognition. The proposed probabilistic model can incorporate isolated character recognition, character sample verification, and n-gram language model in a simple way, leading to more reliable recognition of a text line. When testing on HIT-MW database, experiments show that the proposed method can achieve character-level recognition accuracies of 63.19% without language model and 73.97% with bi-gram language model, respectively, outperforming the most recent results testing on the same dataset.**

***Keywords*—Chinese text line recognition, handwritten Chinese character recognition, verification, confidence measurement, invalid character**

I. Introduction

Unconstrained handwritten offline Chinese text line recognition is one of the most challenging problems in current Chinese character recognition domain. Since in unconstrained handwriting, the shape variances of an isolated character are innumerable and the spatial relationships between neighboring characters are complex, the automatic recognition of the text line is rather difficult. So far, the accuracy of computer recognition is not comparable with that of human reading [1-4].

The methods of unconstrained handwritten offline Chinese text line recognition fall into two categories: segmentation-based recognition [1-2], and segmentation-free recognition [3]. Although in unconstrained handwritten English text line recognition, segmentation-free recognition methods usually perform better than segmentation-based ones, they are quite less used in unconstrained handwritten Chinese text line recognition. This is because the currently available unconstrained handwritten offline Chinese text databases are rather few, which cannot guarantee sufficient Chinese character training samples for segmentation-free recognition methods [3].

On the other hand, in segmentation-based recognition methods, the interference from invalid characters is not trivial. An invalid character is usually assigned a valid class label with high confidence by isolated character recognition, leading to incorrect recognition of a text line. Geometrical information of a text line is usually adopted to help evaluate the validity of a character sample, and it often appears in the following two forms: segmentation score estimation [4-5], and character verification [6-7]. In segmentation score estimation, the measurement of the degree to which a character sample fits the geometrical features of a text line is empirically assumed. Although simple in computation, the empirical assumption is not enough to describe unconstrained handwritten text lines. A more reliable method is character verification, which utilizes a verifier for each character class and even for each pair of character classes to estimate the fitness of a character sample to a valid class. In this way, however, thousands of verifiers are needed in unconstrained handwritten Chinese text line recognition, which is computationally prohibitive.

In this paper, we present a Bayesian-based probabilistic model for unconstrained handwritten offline Chinese text line recognition. In a simple and reliable way, the proposed method incorporates geometrical features of a text line with isolated character recognition, linguistic information, and so on to evaluate possible segmentation hypotheses of a text line. Experiments show that when testing on HIT-MW database [8], character-level recognition accuracies can reach 63.19% without language model and 73.97% with bi-gram language model, respectively, better than the most recent results.

II. System Overview of Segmentation-Based Recognition Methods

The block diagram of segmentation-based recognition methods is illustrated in Fig. 1.

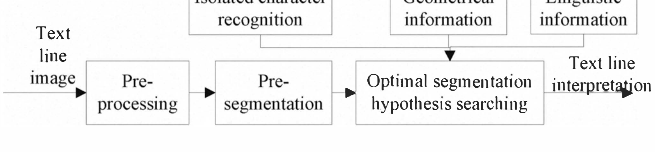
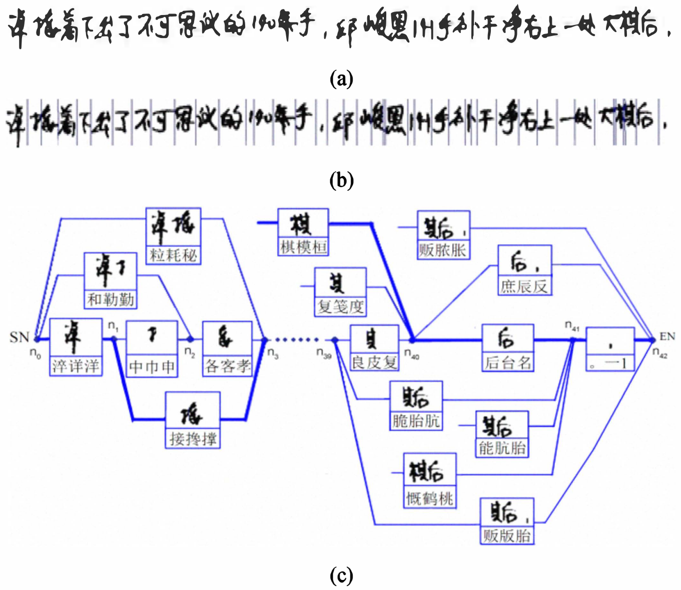


Figure 1. Block diagram of segmentation-based recognition methods.

After pre-segmentation of a text line, a segmentation candidate lattice can be constructed, as is shown in Fig. 2. Each node in the lattice corresponds to a segmentation path between

978-1-4244-6588-0/10/$25.00 ©2010 IEEE

neighboring segments in the pre-segmentation result, and an edge liking two nodes in the lattice represents a character sample. Clearly, plenty of invalid characters are contained in the segmentation candidate lattice, which are assigned valid class labels by isolated character recognition.


(a)

(b)

(c)

Figure 2. (a) An input text line, (b) pre-segmentation result of the text line, and (c) a segmentation candidate lattice constructed from the pre-segmentation result, where each edge shows a character sample and its top 3 recognition candidates in isolated character recognition.

A segmentation hypothesis is a path in a segmentation candidate lattice which starts from node SN and terminates at node EN (see Fig. 2(c)). Here the nodes SN and EN represent the beginning and the end of a text line, respectively. Then text line recognition becomes the problem of searching for the optimal segmentation hypothesis in the segmentation candidate lattice, which can be formulated as follows:

$$C^* = \arg\max_{C}\{\log P(C \mid E)\}, \tag{1}$$

where $C=\{c_1,c_2,...,c_K\}$ is the character label sequence in a possible segmentation hypothesis, $E=\{e_1,e_2,...,e_M\}$ is the segment sequence produced by text line pre-segmentation, and $\log P(C \mid E)$ is the log-likelihood of interpreting the text line as C given E.

III. EVALUATION OF SEGMENTATION HYPOTHESES

For the character image sequence $I=\{i_1,i_2,...,i_K\}$ in a possible segmentation hypothesis, assume that it is recognized as the character label sequence $C=\{c_1,c_2,...,c_K\}$, and that a validity sequence $V=\{v_1,v_2,...,v_K\}$ exists to describe the validity of I, where $v_k \in \{0,1\}(k=0,1,...,K)$, $v_k=1$ denotes that i_k is a valid character sample and $v_k=0$ means that i_k is an invalid one. Then the log-likelihood $\log P(C \mid E)$ in (1) can be rewritten as

$$\log P(C \mid E) = \log \sum_{I'} \sum_{V'} P(C,I',V' \mid E) \approx \log P(C,I,V_0 \mid E) \tag{2}$$

where $V_0=\{v_k=1,k=1,2,...,K\}$ is the sequence denoting that each images $i_k(k=1,2,...,K)$ is a valid character sample, so that the joint probability $P(C,I,V_0 \mid E)$ is dominant over any other probability $P(C,I',V' \mid E)(I' \neq I, or V' \neq V_0)$.

Suppose that the probability $P(I' \mid E)$ is under uniform distribution given E, by omitting the constant term, we have

$$\begin{aligned}
&\log P(C,I,V_0 \mid E) \\
&= \log P(C \mid I,V_0,E) + \log P(V_0 \mid I,E) \\
&= \sum_{k=1}^{K} \log P(c_k \mid c_1,...,c_{k-1},i_1,i_2,...,i_K,1,1,...,1,e_1,e_2,...,e_M) \\
&+ \sum_{k=1}^{K} \log P(v_k=1 \mid v_1=1,...,v_{k-1}=1,i_1,i_2,...,i_K,e_1,e_2,...,e_M)
\end{aligned} \tag{3}$$

where C, I, V_0 and E are the same as in (2).

In segmentation-based recognition methods, the assignment of character label c_k $(k=1,2,...,K)$ relies only on its previous character labels $\{c_1,...,c_{k-1}\}$ and the character image i_k which is presumed to be a valid character sample. Commonly, $\{c_1,...,c_{k-1}\}$ and i_k are conditionally independent for recognizing c_k. Meanwhile, we assume that the character image validity v_k largely depends on the image i_k and its two nearest neighbors i_{k-1} and i_{k+1}. Then by combining (2) and (3), the log-likelihood $\log P(C \mid E)$ in (1) becomes

$$\begin{aligned}
&\log P(C \mid E) \\
&\approx \sum_{k=1}^{K} \log P(c_k \mid c_1,...,c_{k-1}) + \sum_{k=1}^{K} \log P(c_k \mid i_k, v_k=1) \\
&- \sum_{k=1}^{K} \log P(c_k) + \sum_{k=1}^{K} \log P(v_k=1 \mid i_{k-1},i_k,i_{k+1})
\end{aligned} \tag{4}$$

where the probabilities $P(c_k \mid c_1,...,c_{k-1})$ and $P(c_k)$ relate to n-gram language model, the probability $P(c_k \mid i_k, v_k=1)$ is the posterior probability of isolated character recognition, and the probability $P(v_k=1 \mid i_{k-1},i_k,i_{k+1})$ describes the degree to which the character image i_k is a valid character sample.

In the following, we'll introduce the calculation of two probabilities: $P(c_k \mid i_k, v_k=1)$ and $P(v_k=1 \mid i_{k-1},i_k,i_{k+1})$.

A. *Posterior Probability of Isolated Character Recognition*

MQDF classifier [9] has shown good performance in isolated Chinese character recognition. Assume that the class conditional probability $P(i_k, v_k = 1 \mid c_k)$ is under Gaussian distribution for each character class c_k and for each valid character image i_k in class c_k, and that the probability $P(c_k)$ is the prior probability of class c_k. Then the posterior probability $P(c_k \mid i_k, v_k = 1)$ can be calculated as

$$P(c_k \mid i_k, v_k = 1) = \frac{P(i_k, v_k = 1 \mid c_k) * P(c_k)}{\sum_l P(i_k, v_k = 1 \mid c_l) * P(c_l)}, \qquad (5)$$

where the conditional probability $P(i_k, v_k = 1 \mid c_k)$ is negative exponential to the output of an MQDF classifier.

B. *Probability of Character Image Validity*

Assume that a character image i_k belongs to one and only one of the following classes in Chinese text line recognition:

$$\begin{cases} \omega_0 : \mathit{Chinese\ character} \\ \omega_1 : \mathit{digit} \\ \omega_2 : \mathit{punctuation} \\ \omega_3 : \mathit{over-segmented\ character} \\ \omega_4 : \mathit{under-segmented\ character} \end{cases} . \qquad (6)$$

Then the probability $P(v_k = 1 \mid i_{k-1}, i_k, i_{k+1})$ can be rewritten as

$$\begin{aligned} & P(v_k = 1 \mid i_{k-1}, i_k, i_{k+1}) \\ & = \sum_{l=0}^{4} P(v_k = 1, i_k \in \omega_l \mid i_{k-1}, i_k, i_{k+1}) \\ & = P(v_k = 1, \omega_{i_k} \mid i_{k-1}, i_k, i_{k+1}) \\ & = \begin{cases} P(\omega_{i_k} \mid i_{k-1}, i_k, i_{k+1}), & \mathit{if}\ \omega_{i_k} \in \{\omega_0, \omega_1, \omega_2\} \\ 0, & \mathit{if}\ \omega_{i_k} \in \{\omega_3, \omega_4\} \end{cases} \end{aligned} , \qquad (7)$$

where ω_{i_k} is the class in (6) that i_k should belong to.

According to (7), the problem of calculating the probability that i_k is a valid character sample turns into that of assigning one of the classes in (6) to i_k. Suppose that the following features of $\{i_{k-1}, i_k, i_{k+1}\}$ are useful for deciding ω_{i_k}: the geometrical information including the original width w and height h of i_k, the horizontal center distance d_1 between two neighbors i_{k-1} and i_k, and the horizontal center distance d_2 between two neighbors i_k and i_{k+1}; the recognition information d_r and p_r which stand for the recognition distance and the posterior probability of isolated character recognition of i_k. Then (7) becomes

$$\begin{aligned} & P(v_k = 1 \mid i_{k-1}, i_k, i_{k+1}) \\ & = \begin{cases} P(\omega_{i_k} \mid f_v), & \mathit{if}\ \omega_{i_k} \in \{\omega_0, \omega_1, \omega_2\}, \\ 0, & \mathit{if}\ \omega_{i_k} \in \{\omega_3, \omega_4\} \end{cases} \end{aligned} \qquad (8)$$

where $f_v = [w, h, \min(d_1, d_2), \max(d_1, d_2), d_r, p_r]$. Given Gaussian distribution of the class conditional probability $P(f_v \mid \omega_{i_k})$, the uniform distribution of the prior probability $P(\omega_l)(l = 0,\ldots,4)$, the posterior probability $P(\omega_{i_k} \mid f_v)$ can be calculated in a similar way to (5) using another MQDF classifier.

C. *Collection of Training Samples*

The character samples for training the second MQDF classifier associated with probability $P(v_k = 1 \mid i_{k-1}, i_k, i_{k+1})$ can be collected as follows:

(1) Firstly, pre-segment an input text line and construct a segmentation candidate lattice. Check at the position of each ground truth segmentation path in the text line (manually labeled, see Fig. 3(a)) whether there is a corresponding node in the segmentation candidate lattice. Mark a correct sign if the node exists and a missing sign otherwise.

(2) Secondly, for each edge in the segmentation candidate lattice, if the two nodes it links are both marked with a correct sign, and any other node marked with correct sign or missing sign is absent from the edge, then it is regarded as a positive sample. Otherwise, it is regarded as a negative sample.

(3) Lastly, the class label $\omega_l\ (l = 0,\ldots,4)$ of a positive sample is assigned according to the ground truth content of the text line (known beforehand), whereas for a negative sample it is assigned according to whether the edge contains any node marked with a correct or missing sign.

The above process is illustrated in Fig. 3. Specifically, the character sample between node n_{14} and node n_{15} is assigned a class label ω_3, because according to the ground truth content of the text line, this sample corresponds to an erasure.

D. *Description of the Evaluation Procedure*

The block diagram of evaluating a possible segmentation hypothesis is shown in Fig. 4 according to the evaluation criterion given in (4). The probabilities $P(c_k \mid i_k, v_k = 1)$ and $P(v_k = 1 \mid i_{k-1}, i_k, i_{k+1})$ are calculated in (5) and (8), respectively, and the probabilities $P(c_k \mid c_1, \ldots, c_{k-1})$ and $P(c_k)$ are given by n-gram language model.

Beam search algorithm [10] is one of the most frequently used algorithms for solving the optimization problem in (1) given the evaluation criterion for $\log P(C \mid E)$. However, since the search results tend to prefer short character label

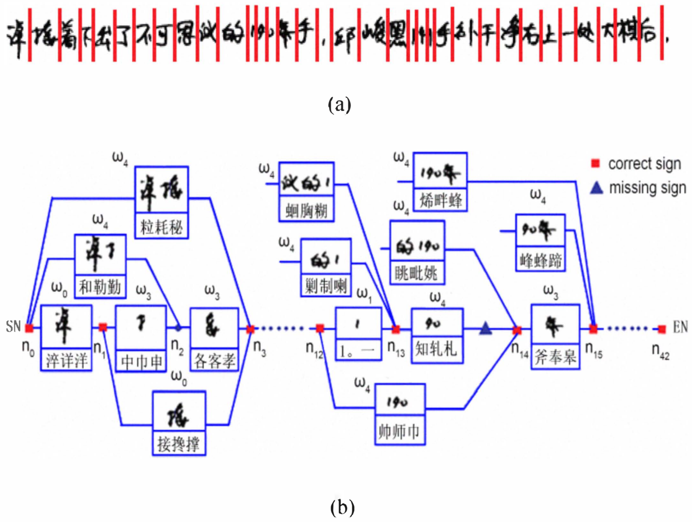


Figure 3. (a) The position of the ground truth segmentation paths in a text line, (b) the segmentation candidate lattice for the text line and the collected training samples.

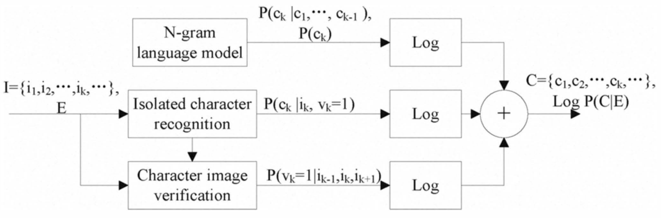

N-gram language model
$P(c_k\,|c_1,\cdots,c_{k-1})$,
$P(c_k)$
Log
$I=\{i_1,i_2,\cdots,i_k,\cdots\}$,
E
Isolated character recognition
$P(c_k\,|i_k, v_k=1)$
Log
+
$C=\{c_1,c_2,\cdots,c_k,\cdots\}$,
Log P(C|E)
Character image verification
$P(v_k=1|i_{k-1},i_k,i_{k+1})$
Log

Figure 4. Block diagram of evaluating a segmentation hypothesis.

sequences [7, 10], in order to reduce this kind of recognition error, normalization of $\log P(C\,|\,E)$ with respect to the number of characters in C is usually adopted. In this paper, we use beam search algorithm to find 10 best segmentation hypotheses of a text line according to the evaluation criterion in (4), and then select among them the one with the largest normalized value $\log P(C\,|\,E)/N$, where N is the number of characters in C. In this way of searching, examples of the best segmentation hypotheses of a text line with and without languge model respectively are shown in Fig. 5.

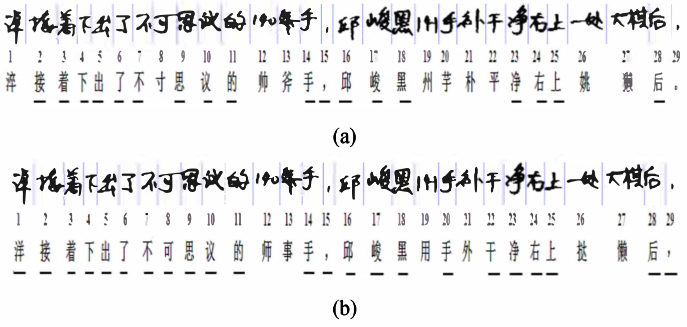

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
淬 接 着 下 出 了 不 寸 思 议 的 帅 斧 手 ， 邱 峻 黑 州 芋 朴 平 净 右 上 姚 獭 后 。
(a)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
洋 接 着 下 出 了 不 可 思 议 的 师 事 手 ， 邱 峻 黑 用 手 外 干 净 右 上 拢 懒 后 ，
(b)

Figure 5. Examples of the best segmentation hypotheses of a text line (a) without language model, and (b) with bi-gram language model, respectively. The final segmentation paths are drawn in blue lines, and the correctly recognized characters are underlined.

IV. Experiments

We test 383 text lines in the test set of HIT-MW database [8], which is a realistic handwritten offline Chinese text database written by multiple people. In this database, the average length of a text line is 21.51 characters. All the experiments are run on a PC with Dual-Core 1.6GHZ CPU and 1G memory.

For the first MQDF classifier in isolated character recognition, 160 sets of 3755 Chinese characters, 50 sets of 10 digits and 50 sets of 10 frequently used punctuations from our lab compose the training sample. The test samples come from the characters in the test set of HIT-MW database, and the recognition accuracy of 71.30% (one best recognition candidate) is achieved.

For the second MQDF classifier in character image verification, 100 text lines in the train set of HIT-MW database generate the training samples. And the text lines in the test set of HIT-MW database produce testing samples, whose correct classification rate reach 79.00%.

We use the same bi-gram language model as in [2] in our experiments, but preserve only 10 character recognition candidates in isolated character recognition. Since 10 character candidates may not enough for using bi-gram language model, we employ a two-stage searching scheme: in the first stage, without language model, 10 best segmentation hypotheses of a text line are found using beam search algorithm given the evaluation criterion in (4); then in the second stage, by Viterbi algorithm [11], bi-gram language model is used to assist the recognition and verification of the character images in each segmentation hypothesis found in the first stage.

The following two terms are used to evaluate the results of text line recognition: character-level correct recognition rate *CR*, and the *F* measure describing the segmentation accuracy of text lines, which are defined as

$$\begin{cases} CR = NC_c\,/\,NT_c \\ F = 2/(R^{-1}+P^{-1}) \\ R = NC_s\,/\,NT_s \\ P = NC_s\,/\,NA_s \end{cases}, \tag{9}$$

where NC_c is the number of correctly recognized characters, and NT_c is the number of total characters in text lines, NC_s is the number of correctly detected segmentation positions, NT_s is the number of ground truth segmentation positions in text lines, NA_s is the number of all detected segmentation positions.

The results of text line recognition using the evaluation criterion in (4) are listed in Table I. The first row lists the results when considering only isolated character recognition (Condition 1). The second row lists the results when considering both isolated character recognition and character image verification (Condition 2). The third and the fourth rows list the results when using isolated character recognition, character image verification and bi-gram language model, together in beam search algorithm (Condition 3) and separated in a two-stage searching scheme (Condition 4), respectively.

From Table I, we can see that the verification of character images using (8) largely reduces the wrong recognition of invalid character images, the recognition and segmentation accuracies are increased by 25.47% and 20.37%, respectively. In addition, the two-stage searching scheme is helpful to improve the recognition and segmentation accuracies given the limited number of character candidates when using bi-gram language model.

TABLE I. RESULTS OF THE PROPOSED METHOD

	***CR* (%)**	***F* (%)**	***R* (%)**	***P* (%)**	**Time (s/line)**
Condition 1	37.72	70.03	62.44	79.73	74.23
Condition 2	63.19	90.40	88.62	92.26	78.37
Condition 3	71.41	90.09	85.65	95.02	80.57
Condition 4	**73.97**	**91.55**	89.28	93.94	81.73

The comparison of the recognition accuracy *CR* between the proposed method and other methods is listed in Table II. All the methods are tested on the same data, and using the same bi-gram language model. The method in [4] uses empirical evaluation of segmentation scores for character image verification. Specifically, we perform this method using the same MQDF classifier as ours for isolated character recognition, and preserve the same number of segmentation hypotheses as ours in beam search algorithm. The method in [2] uses a more powerful MQDF classifier than ours for isolated character recognition, but just adopts basic measurements for character image verification. The method in [3] is a HMM-based recognition method that suffers from insufficient Chinese character training samples.

TABLE II. COMPARION OF *CR* (%) AMONG DIFFERENT METHODS

	The proposed method	**Method in [4]**	**Method in [3]**	**Method in [2]**
Without n-gram	**63.19**	41.55	44.22	57.60
Bi-gram	73.97	50.81	—	**77.18**

Table II shows that without using n-gram language model, the proposed method achieves the best recognition accuracy, which is increased by 21.64%, 18.97%, and 5.59% in comparison with [4], [3], and [2], respectively. On the other hand, when using the same bi-gram language model, the recognition accuracy of the proposed method is higher than that of [4] by 23.16%, but lower than that of [2] by 3.21%. This is because in our method, the number of character classes and the number of character candidates in isolated character recognition are much less than those in [2], which is disadvantageous for using bi-gram language model.

V. Conclusions

In this paper, we present a Bayesian-based probabilistic model for unconstrained handwritten offline Chinese text line recognition. In the probabilistic model, the recognition and verification of character sample images are incorporated with n-gram language model to give reliable evaluation of a segmentation hypothesis. Instead of complex computation, the verification of a character sample image is performed in a simple way. When testing on HIT-MW database, the proposed method shows effectiveness in text line recognition, outperforming the most recent results testing on the same data.

Acknowledgment

This work is supported in part by NSFC (Grant no. U0735004, 60772116) and GDSFC (Grant no. 07118074, 2007B010200048, 2008A050200004*)*). The authors would like to thank the National Laboratory of Pattern Recognition (NLPR) of Institute of Automation in Chinese Academy of Science for sharing the ground truth labeling of HIT-MW database and the n-gram language model.

References

[1] Y. Li, Chew L. Tan, X. Q. Ding, and C. S. Liu, "Contextual post-processing based on the confusion matrix in offline handwritten Chinese script recognition," Pattern Recognition, vol. 37, pp. 1901- 1912, 2004.

[2] Q. F. Wang, F. Yin, and C. L. Liu, "Integrating language model in handwritten Chinese text recognition," Proc. 10th Int. Conf. Document Analysis and Recognition, Catalonia, Spain, pp. 1036-1040, June 2009.

[3] T. H. Su, T. W. Zhang, D. J. Guan, and H. J. Huang, "Off-line recognition of realistic Chinese handwriting using segmentation-free strategy," Pattern Recognition, vol. 42, pp. 167-182, 2009.

[4] Q. Fu, X. Q. Ding, T. Liu, Y. Jiang, and Z. Ren, "A novel segmentation and recognition algorithm for Chinese handwritten address character strings," Proc. 18th Int. Conf. Pattern Recognition, Hong Kong, China, pp. 974-977, August 2006.

[5] J. Sadri, C. Y. Suen, and T. D. Bui, "A genetic framework using contextual knowledge for segmentation and recognition of handwritten numeral strings," Pattern Recogniton, vol. 40, pp. 898-919, 2007.

[6] Z. D. Feng, and Q. Huo, "Confidence guided progressive search and fast match techniques for high performance Chinese/English OCR," Proc. 16th Int. Conf. Pattern Recognition, Quebec City, Canada, pp. 89-92, August 2002.

[7] X. D. Zhou, J. L. Yu, C. L. Liu, T. Nagasaki, and K. Marukawa, "Online handwritten Japanese character string recognition incorporating geometric context," Proc. 9th Int. Conf. Document Analysis and Recognition, Curitiba, Parana, Brazil, pp. 48- 52, September 2007.

[8] T. H. Su., T. W. Zhang, and D. J. Guan, "Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text line," Int. J. Document Analysis and Recognition, vol. 10, pp. 27-38, 2007.

[9] T. Long, and L. W. Jin, "Building compact MQDF classifier for large character set recognition by subspace distribution sharing," Pattern Recognition, vol. 41, pp. 2916-2925, 2007.

[10] C. L. Liu, H. Sako, and H. Fujisawa, "Effects of classifier structures and training regimes on integrated segmentation and recognition of handwritten numeral strings," IEEE Transa. Pattern Analysis and Machine Intelligence, vol. 26, pp. 1395-1407, 2004.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," Proc. IEEE, vol. 77, pp. 257-285, 1989.