A Lattice-Based Method for Keyword Spotting in Online Chinese Handwriting

Heng Zhang, Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun East Road, Beijing 100190, P.R. China
{hzhang07, liucl}@nlpr.ia.ac.cn

Abstract——**This paper proposes a lattice-based method for keyword spotting in online Chinese handwriting to improve the trade-off between accuracy and speed, and to overcome the out-of-vocabulary (OOV) problem of lexicon-driven approach. Using a character string recognition algorithm, the lattice-based method generates a candidate lattice of N-best list. We observe that search multiple candidate strings reduces the precision rate while improving the recall rate compared to the top-rank string. We propose a post-processing method using word confusion network (WCN) for candidate pruning in the lattice in order to alleviate the precision loss of searching multiple candidate strings. Our experimental results on a large database CASIA-OLHWDB2.0 demonstrate the effectiveness of the proposed method.**

Keywords-Lattice-based keyword spotting; N-best list; post-processing

I. INTRODUCTION

Online handwritten documents are constantly produced with the wide use of digitizing tablets, tablet PCs and digital pens (such the Anoto Pen). This entails efficient retrieval techniques with higher recall and less time to exploit the semantic information in the documents. In addition to the difficulties of layout segmentation and character segmentation as well as the character shape variation, Chinese documents suffer from the large alphabet (over 5,000 characters are daily used) and the difficulty of word segmentation (there is no space between words). This requires more efficient indexing and search techniques to search the document in every character and match with a very large lexicon.

For large database retrieval, it is necessary to build an index containing the candidate recognition results. According to the indexing technique, handwritten document retrieval methods can be categorized into two groups: indexing by character recognition (OCR) [1-4] and lexicon-driven indexing [5-8]. OCR-based methods generate the index from the character candidates (stored in a lattice) of high scores output by character recognizers. The lattice is independent of a lexicon (vocabulary), but to guarantee high recall of retrieval, the lattice needs to store a large number of candidate characters. This can consume a large volume of disk space. In lexicon-driven methods, a word recognizer is used to generate multiple word candidates for the words contained in a pre-compiled lexicon, or the document is matched with all the word models corresponding to the words in the lexicon to generate matched candidates. From the candidates, a word in the lexicon can be searched quickly. However, when the query word is out-of-vocabulary (OOV), it cannot be retrieved.

Both the OCR-based method and the lexicon-driven method suffer from the compromise between the storage size of candidates and the recall rate of retrieval. The lexicon-driven method also suffers from the OOV problem though it is more storage efficient. In lexicon-driven methods, how to build word models and improve the word matching accuracy is a focus of research. In OCR-based methods, improving the accuracy of character recognition (helping reduce the number of candidate characters) is always beneficial to retrieval.

In this paper, we propose a lattice-based method for keyword spotting in online Chinese handwritten documents. The lattice-based method generates candidate string recognition results (N-best list), which are stored in the index database and then the query word is searched in the N-best strings. This method has been proven effective in spoken document retrieval [9]. Its advantage is that the linguistic context is exploited by (lexicon-free) string recognition to generate compact candidates of words in the N-best list. To improve the recall rate of searching in multiple candidate strings while alleviate the loss of precision, we propose a post-processing method for candidate pruning in the lattice, using the linguistic model and character recognition scores to remove some implausible words.

Our string recognition method is based on the energy function defined in [10] integrating character recognition, geometric context and statistical language model. Our experimental results on a large database CASIA-OLHWDB2.0 [11] demonstrate that the proposed lattice-based keyword spotting method achieve both high recall and precision rates.

II. SYSTEM OVERVIEW

Figure 1 illustrates the proposed document retrieval system, containing two main parts: data storage and keyword spotting. The online documents are first segmented into text lines [12] using temporal and spatial information, after which each text line is over-segmented into primitive segments. Between each pair of consecutive segments, there is a candidate segmentation point. Consecutive segments are combined to generate candidate character patterns, represented in a segmentation candidate lattice. With a character recognizer, each candidate character is assigned a number of candidate classes. Each path in the enlarged lattice denotes a possible segmentation-recognition result of the text line. The N-best paths generated by beam-search strategy [13] are post-processed by incorporating linguistic context and character recognition scores. We prune the lattice formed by the N-best list using the word confusion network (WCN) [14]. The pruned lattice are stored as index, and in retrieval, each query word will be compared with the characters in the index.

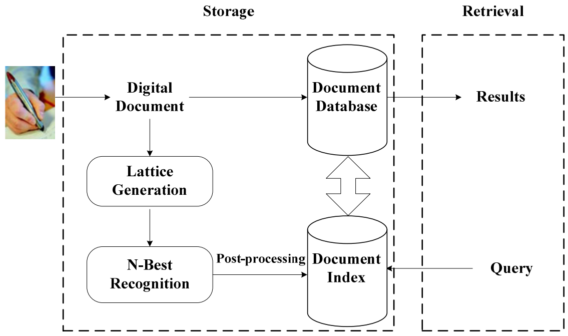
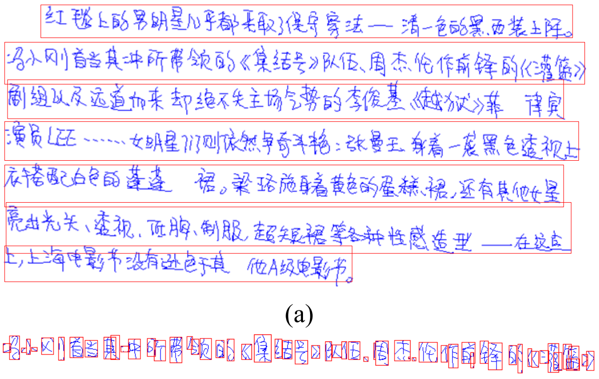


Figure 1. Overview of the keyword spotting system.

III. Lattice Generation and N-best Recognition

The input document is first segmented into text lines according to the time and space interval between consecutive strokes (the pre-segmentation step of [12]). Each line is over-segmented into primitive segments according to the off-stroke distance (pen-lift between two successive strokes) and the spatial overlap between strokes. Examples of text line segmentation and over-segmentation are shown in Figure 2.



(a)

(b)

冯小刚首当其冲所带领的《集结号》队伍、周杰伦作前锋的《灌篮》

(c)

Figure 2. Text line segmentation (a), over-segmentation of a text line (b) and the corresponding ground-truth (c).

After over-segmentation, consecutive segments are combined to generate candidate character patterns [8], which are represented in a segmentation candidate lattice (Figure 3). After assigning candidate classes to each candidate character pattern using a character recognizer, the candidate segmentation-recognition paths are evaluated by combining the classification scores, geometric context and linguistic context, and the N-best paths are obtained using the beam search strategy.

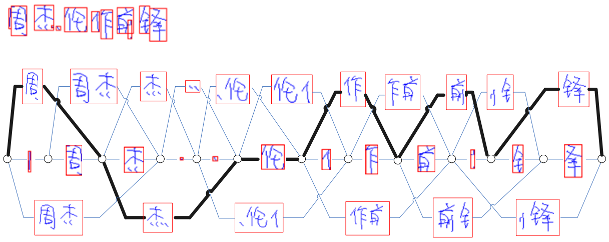
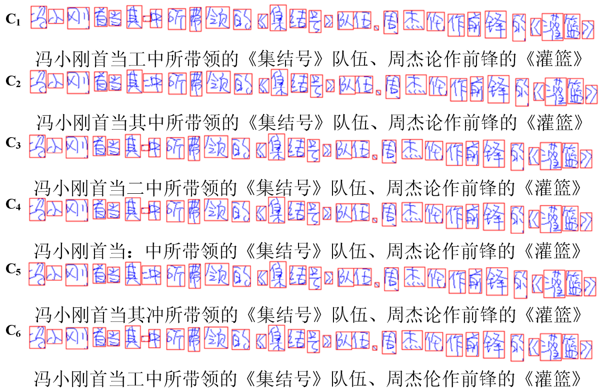


Figure 3. Example of segmentation candidate lattice. The thick lines denote the desired segmentation path.

In beam-search, the partial paths ending at each candidate segmentation point are sorted and at most N partial paths with minimum costs are retained for extension. The paths are evaluated by the energy function defined in [10] and we weight the likelihood of each character pattern with its number of constituent segments (similar to [15]), while the parameters are optimized with the minimum classification error (MCE) criterion [10]. The combined feature functions include the character recognition scores, the bi-gram language model and four geometric models [10] [16]. Figure 4 shows an example of 6-best recognition results. The word "首当其冲" and the character "伦" are only correctly recognized in path 5 and path 6, which necessitates retaining multiple recognition hypotheses. Besides, the results such as "其中" in path 2 and "二中" in path 3 are wrong and should be pruned.



冯小刚首当工中所带领的《集结号》队伍、周杰论作前锋的《灌篮》

冯小刚首当其中所带领的《集结号》队伍、周杰论作前锋的《灌篮》

冯小刚首当二中所带领的《集结号》队伍、周杰论作前锋的《灌篮》

冯小刚首当：中所带领的《集结号》队伍、周杰论作前锋的《灌篮》

冯小刚首当其冲所带领的《集结号》队伍、周杰论作前锋的《灌篮》

冯小刚首当工中所带领的《集结号》队伍、周杰伦作前锋的《灌篮》

Figure 4. Example of 6-best recognition results.

IV. Post-Processing

To prune the lattice formed by the N-best paths using linguistic context and character recognition scores, we construct a WCN [14] to remove the implausible words. The WCN is a simplified lattice and it forces the competing words to be in the same group by aligning the words occurring at the same interval in the lattice. The posterior probabilities of the words in the WCN are computed as confidence scores. Figure 5 shows the typical structures of the traditional lattice and the WCN, where we can see that WCN is more efficient in size and structure.

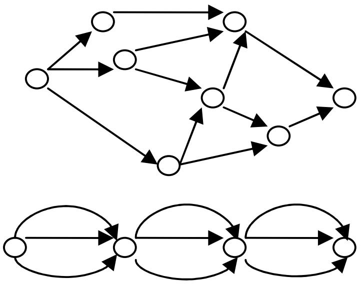


Figure 5. Traditional lattice (upper) and the derived WCN (lower).

Figure 6 illustrates the post-processing procedure, including two parts: WCN construction and candidates pruning. The N-best paths $C_1=(L_1,R_1,S_1)$, … ,$C_N=(L_N,R_N,S_N)$ are taken as input, where L_i, R_i and S_i denote the *i*th sequence (string) of labels, recognition scores and candidate segmentation points, respectively.

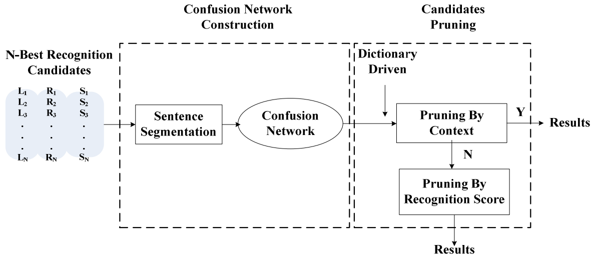
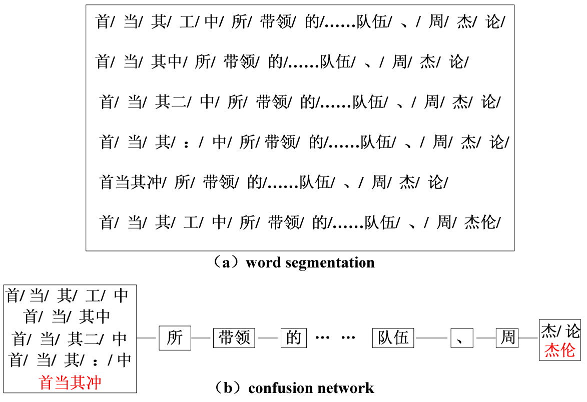


Figure 6. Flowchart of the post-processing procedure.

A. WCN Construction



首/ 当/ 其/ 工/ 中/ 所/ 带领/ 的/……队伍/ 、/ 周/ 杰/ 论/
首/ 当/ 其中/ 所/ 带领/ 的/……队伍/ 、/ 周/ 杰/ 论/
首/ 当/ 其二/ 中/ 所/ 带领/ 的/……队伍/ 、/ 周/ 杰/ 论/
首/ 当/ 其/ ：/ 中/ 所/ 带领/ 的/……队伍/ 、/ 周/ 杰/ 论/
首当其冲/ 所/ 带领/ 的/……队伍/ 、/ 周/ 杰/ 论/
首/ 当/ 其/ 工/ 中/ 所/ 带领/ 的/……队伍/ 、/ 周/ 杰伦/

（a）word segmentation

（b）confusion network

Figure 7. Examples of word segmentation and WCN.

Before constructing the WCN, the result label sequences (strings) are first segmented into word sequences using a Chinese lexical analysis system [17], and further aligned with the segmentation results. The WCN consists of equivalence cells each containing words or phrases with the same starting and ending segmentation points. The recognition scores of constituent characters and the path index are then assigned to each word.

Figure 7 shows the word segmentation results and the corresponding WCN. In Figure 7(b), each cell contains one or more units and each unit contains a word or a phrase.

B. Candidates Pruning

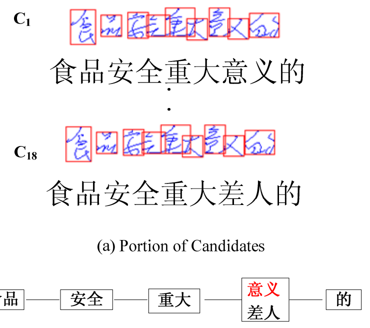
The candidate words in the WCN are pruned according to the linguistic context and the character recognition scores. Candidate words with more characters are considered more reliable than those with fewer characters, because of the stronger linguistic compatibility. Also, characters with higher recognition scores are more likely to be retained. Based on these observations, we prune the WCN using three heuristic rules.

(1) If a word can be found in a lexicon and it consists of three characters or more, the other words in the same cell will be pruned. In the first cell of Figure 7(b), only the 4-character word “首当其冲” are retained and others are removed. A 2-character word is not reliable enough to reject other words.

(2) When two consecutive multi-character words form a word bi-gram in a dictionary, they will be retained. In Figure 8, the two words “重大” and “意义” in two cells form a 4-character phrase, but the words “重大” and “差人” do not form a phrase, so the word “差人” is pruned. We used the lexicon and the dictionary downloaded from the resource of the Sougo labs [18], and organized them in tree structure (trie) as used in [19].

(3) After pruning with linguistic context, the remaining 2-character words are pruned using character recognition scores. If a word in the lower path has lower confidence score than one in a upper path with the same segmentation boundaries, it is pruned; If its score is higher, it is retained. In the last cell of Figure 7(b), the score of the word “杰伦” is greater than the that of “杰论”，so the word “杰伦” (the correct one) is retained although it belongs to a path of lower score.

After pruning by the above three steps, the characters in the remaining WCN are organized to the character lattice which is used as the index. While in spotting, words are searched only from characters in the same path (string).



(a) Portion of Candidates

(b) Word Confusion Network

Figure 8. Pruning based on word bi-gram. The red word is retained.

V. Experimental results

We evaluated the performance of the proposed method in experiments on an online Chinese handwriting database CASIA-OLHWDB2.0 [11], which contains 2,092 pages written by 420 persons. We used the 1,672 pages of 336 writers for training model parameters (for path evaluation) and the 420 pages of the remaining 84 writers for testing. The keyword retrieval performance is measured by recall, precision and F-measure, defined as follows:

$$recall = \frac{\#\text{correct detected words}}{\#\text{truth words}} \tag{1}$$

$$precision = \frac{\#\text{correct detected words}}{\#\text{detected words}} \tag{2}$$

$$F = \frac{2}{\dfrac{1}{recall} + \dfrac{1}{precision}} \tag{3}$$

We used a modified quadratic discriminant function (MQDF) classifier for character recognition, which was trained using the training character samples in datasets CASIA-OLHWDB1.0~1.2 (isolated characters) and CASIA-OLHWDB2.0~2.2 (characters segmented from text pages). The four geometric models [10, 16] were trained on the training set of CASIA-OLHWDB2.0. On the test character set, the accuracies of five classifiers (one (f_0) for character recognition and four for geometric models) are listed in Table I, where $f_1$and f_3 are QDF classifiers for unary and binary class-dependent geometric features, f_2 and f_4 are SVM classifiers for unary and binary class-independent geometric features.

TABLE I. Test accuracies (%) of five classifiers.

	f_0	f_1	f_2	f_3	f_4
accuracy	92.47	59.28	96.44	48.12	90.23

In over-segmentation, each candidate character pattern is assumed to have at most six segments. To guarantee high recall rate, the character recognizer assigns to each character pattern 20 candidate classes of highest scores.

We used the high-frequency words in the lexicon of the Sogou labs [18] as query words in testing. The top 60,000 frequently used words, including 39,049 two-character words, 9,972 two-character words and 9,438 four-character words, were tested in our experiments.

A. Retrieval Results

We first give the results of keyword spotting from the N-best list without post-processing. The retrieval results for words of different lengths are listed in Table II, Table III and Table IV, respectively. We can see that when search multiple candidate strings (N>1), the recall rate is improved compared to search the top rank string only though the precision rate is decreased. If we keep the precision rate above 90%, searching N-best paths can boost the recall rate by 1.29% for 2-character words, 3.42% for 3-character words and 2.74% for 4-character words, respectively.

TABLE II. Retrieval results of 2-character words without post-processing.

N	1	3	5	10	20	50
recall	92.85	93.83	94.14	94.54	94.80	95.12
precision	96.43	93.04	90.76	86.26	82.06	73.83
F	94.61	93.44	92.42	90.21	87.97	83.13

TABLE III. Retrieval results of 3-character words without post-processing.

N	1	3	5	10	20	50
recall	89.65	91.41	91.74	92.06	92.71	93.07
precision	98.98	98.28	98.05	97.12	96.22	94.12
F	94.08	94.72	94.79	94.52	94.44	93.59

TABLE IV. Retrieval results of 4-character words without post-processing.

N	1	3	5	10	20	50
recall	88.09	89.55	89.80	90.34	90.57	90.83
precision	99.47	99.19	99.05	98.76	98.60	98.24
F	93.43	94.12	94.20	94.36	94.42	94.39

TABLE V. Retrieval esults of 2-character words with post-processing.

N	5	10	20	30	40	50
recall	94.04	94.50	94.65	94.72	94.82	94.84
precision	90.94	90.06	89.40	88.94	88.59	88.21
F	92.46	92.23	91.95	91.74	91.59	91.41

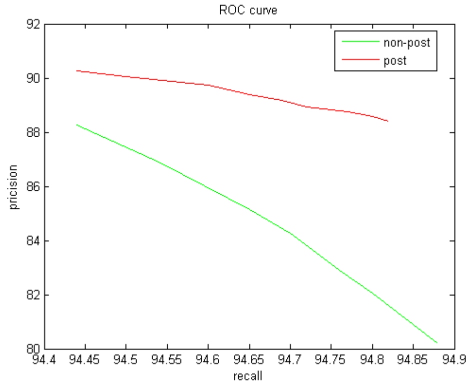


Figure 9. Recall and precision rates for the results of 2-character with post-processing and without post-processing.

By post-processing of N-best list, the decrease of precision rate can be largely alleviated while keeping high recall rates. The retrieval results with post-processing for 2-character words are shown in Table V. Comparing with the results in Table II, we can see that when increasing N, post-processing largely improves the precision rate while is the recall rate remains high. When keeping the precision rate above 90%, the recall rate is as high as 94.50%, higher than that of retrieval without post-processing (94.14%). Figure 9 shows the ROC curves (by varying N) of spotting 2-character words with and without post-processing. The benefit of post-processing is evident.

Our experiments were implemented on a PC with Intel(R) Core(TM)2 Duo CPU E8400 3.00 GHz processor and 2GB RAM, the average time of searching for a query word in all the test documents is 0.36s.

B. *Error Analysis*

Keyword spotting errors can be caused by string recognition errors (character segmentation or recognition errors in generating N-best list). In our experiments, segmentation errors mainly lie in the merging strokes of adjacent characters into segments in over-segmentation. Character recognition may exclude the correct class from the top ranks, and path evaluation and search may exclude the correct string. An example of string recognition error is shown in Figure 10, which shows both over-segmentation error and string recognition error. The off-stroke distance of the word "知名" is not long enough, so the last stroke of "知" and the first stroke of "名" are merged. Besides, the word "占有额" and the word "全方位" are not correctly recognized. By mis-segmentation or mis-recognition, the keyword cannot be retrieved as well.

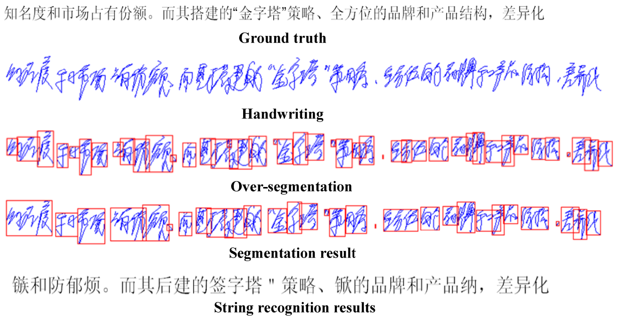

知名度和市场占有份额。而其搭建的"金字塔"策略、全方位的品牌和产品结构，差异化
Ground truth
Handwriting
Over-segmentation
Segmentation result
镞和防郁烦。而其后建的签字塔 " 策略、锨的品牌和产品纳，差异化
String recognition results

Figure 10. Example of sting recogniton error.

VI. CONCLUSIONS

In this paper, we presented a novel lattice-based keyword spotting method for online handwritten Chinese documents. By searching multiple string recognition paths (N-best list) generated by string recognition, the recall rate of word spotting is increased with considerable loss of precision. We proposed a post-processing method for pruning the word confusion network formed from the N-best strings. This was shown to be effective to improve the precision while keeping high recall rate. In retrieving 2-character words, we achieved 94.50% recall rate with precision above 90%.

ACKNOWLEDGMENT

The authors thank Xiang-Dong Zhou, Fei Yin, Da-Han Wang and Qiu-Feng Wang for helpful discussions.

REFERENCES

[1] H. Fujisawa, K. Marukawa, Full-text search and document recognition of Japanese text, *Proc. of 4th Symposium on DA & IR*, Las Vegas, Nevada, USA, 1995 , pp.55-80.

[2] T. Kameshiro, T. Hirano, Y. Okada, E. Yoda, A document image retrieval method tolerating recognition and segmentation errors of OCR using shape-feature and multiple candidates, *Proc. 5th ICDAR*, 1999, pp. 681-684.

[3] T. Kameshiro, T. Hirano, Y. Okada, A document retrieval method from handwritten characters based on OCR and character shape information, *Proc. 6th ICDAR*, 2001, pp. 597-601.

[4] N.R. Howe, S. Feng, R. Manmatha, Finding words in alphabet soup: inference on freeform character recognition for historical scripts, *Pattern Recognition*, 42(12): 3338-3347, 2009.

[5] D. Doermann, The indexing and retrieval of document images: a survey, *Computer Vision and Image Understanding*, 70(3): 287-298, 1998.

[6] H. Cao, A. Bhardwaj, V. Govindaraju, A probabilistic method for keyword retrieval in handwritten document images, *Pattern Recognition*, 42(12): 3374-3382, 2009.

[7] J.A. Rodriguez-Serrano, F. Perronnin, Handwritten word-spotting using hidden Markov models and universal vocabularies, *Pattern Recognition*, 42(9): 2106-2116, 2009.

[8] H. Zhang, D.-H. Wang, C.-L. Liu, Keyword spotting from online Chinese handwritten documents using one-vs-all trained character classifier, *Proc. 12th ICFHR*, 2010, pp.271-276.

[9] T. K. Chia, K. C. Sim, H. Li, H. T. Ng, Statistical lattice-based spoken document retrieval, *ACM Trans. on Information Systems*, 28(1): 1-30, 2010.

[10] X.-D. Zhou, C.-L. Liu, M. Nakagawa, Online handwritten Japanese character string recognition using conditional random fields, *Proc. 10th ICDAR*, 2009, pp.521-525.

[11] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, CASIA online and offline Chinese handwriting databases, submitted to *ICDAR2011*.

[12] X.-D. Zhou, D.-H. Wang, C.-L. Liu, A robust approach to text line grouping in online handwritten Japanese documents, *Pattern Recognition*, 42(9): 2077-2088, 2009.

[13] C.-L. Liu, H. Sako, H. Fujisawa, Effects of classifier structures and training regimes on integrated segmentation and recognition of handwritten numeral strings, *IEEE Trans. PAMI*, 26(11): 1395-1407, 2004.

[14] S. Quiniou, E. Anquetil, Use of a confusion network to detect and correct errors in an on-line handwritten sentence recognition, *Proc. 9th ICDAR*, 2007, pp. 382-386.

[15] Q.-F. Wang, F. Yin, C.-L. Liu, Integrating language model in handwriting Chinese text recognition, *Proc. 10th ICDAR*, 2009, pp.1036-1040.

[16] F. Yin, Q.-F. Wang, C.-L. Liu, Integrating geometric context for text alignment of handwritten Chinese documents, *Proc. 12th ICFHR*, 2010, pp.7-12.

[17] H.-P. Zhang, H.-K. Yu, D.-Y. Xiong, Q. Liu, HMM-based Chinese lexical analyzer ICTCLAS, *Proc. 2th SIGHAN*, 2003, pp.184-187.

[18] http://www.sogou.com/labs/resources.html.

[19] C.-L. Liu, M. Koga, H. Fujisawa, Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading, *IEEE Trans. PAMI*, 24(11): 1425-1437, 2002.