SCUT-COUCH2009-TL: An Unconstrained Online Handwritten Chinese Text Lines Dataset

*Hanyu Yan, Lianwen Jin**
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, P.R.China
kingyhy@gmail.com, *lianwen.jin@gmail.com

Christian Viard-Gaudin, Harold Mouchère
IRCCyN, Université de Nantes, Rue Christian PAUC–BP 50609, 44306, Nantes Cedex 03 – France
Christian.Viard-Gaudin@univ-nantes.fr, Harold.Mouchere@univ-nantes.fr

Abstract

An unconstrained online handwritten Chinese text lines dataset, SCUT-COUCH2009-TL, a subset of SCUT-COUCH [1], is built to facilitate the research of unconstrained online Chinese text recognition. Texts for handcopying are sampled from China Daily corpus with a stratified random manner. The current vision of SCUT-COUCH2009-TL has 8,809 text lines (4,813 lines are collected by touch screen LCD and 3,996 by digital pen) and 159,866 characters in total that are written by more than 157 participants. To demonstrate that the dataset is practical, an over-segmentation, dynamic programming and semantic model based algorithm was presented for segmenting and recognizing the unconstrained online Chinese text lines. In preliminary experiments on the dataset, the proposed algorithm recognition achieves a baseline accuracy of 56.41%.

Keywords: SCUT-COUCH2009-TL, online Chinese handwritten dataset, online Chinese text line recognition

1. Introduction

Publicly available datasets are important for the handwriting recognition research. On the one hand, they provide a large number of training and testing data, resulting in high model fit and reliable confidence in statistic. On the other, they offer a means by which evaluation among different recognition algorithms can be performed. More and more handwriting researchers begin to pay much attention to the dataset standardization and evaluate their work using standard datasets.

During the last twenty years, numbers of handwriting datasets in different languages have been published in the literature. Most of the datasets are of offline data (images converted from paper documents). To name a few, there're the CEDAR English words and characters [2], English sentence database IAM [3], Japanese Kanji character databases ETL8B and ETL9B, Korean database PE92 [4], Indian database of ISI [5], Arabic databases [6], Chinese databases HCL2000 [7] and HIT-MW [8]. Databases of online handwritten data (trajectory data of strokes) are not so popular as offline ones because the collection of online data relies on special devices such as digitizing tablet, tablet PC and PDA. A few efforts in the area are the UNIPEN project [9], the Japanese online handwriting databases Kuchibue [10] [11], SCUT-COUCH2008 [12], and the very recent Chinese online handwriting database CASIA-OLHWDB1 [13], collected by Anoto pen. The French database IRONOFF contains both online and offline data, collected by attaching paper on digitizing tablet while writing [14].

To support research on recognition of unconstrained online Chinese handwriting, we have built an unconstrained online handwritten Chinese text dataset, SCUT-COUCH2009-TL[1]. It contains 159,866 characters that are written by more than 157 writers.Compared with HIT-MW and CASIA-OLHWDB1, our dataset possesses at least three advantages: (1). We divided the corpus into 7 subtopics: domestic news, international news, sports news, economic news, cultural news, academic trends, and education weekly; (2). Used two representative collection tools: touch screen LCD and digital pen; (3). We carry out our sample collection work in two schools (SCUT and Polytech Nantes), the latter being in France, while the former is from China.

To demonstrate that the database is practical, we conducted experiments using an over-segmentation scheme; a dynamic programming and semantic model based recognizer and obtained recognition accuracy of 56%.

The flowchart of building SCUT-COUCH2009-TL is shown in Figure 1. The rest of this paper is arranged as follows. The next section describes the sampling preparation. Then the dataset processing is introduced in section 3. Section 4 describes the segmenting and recognizing algorithms. The experiments result based on

[1] SCUT is the abbreviation of South China University of Technology, COUCH is the abbreviation of Comprehensive Online Unconstrained Chinese Handwriting, and TL is the abbreviation of Text Lines.

dataset are described at section 5. Finally, concluding remarks are given in section 6.

2. Sampling Preparation

To make our dataset more representative of the real Chinese handwriting, we have carefully taken care of all stages included in the data collection process. In this section, we describe the sampling of writer, selection of sampling devices, text sampling and layout designing respectively.

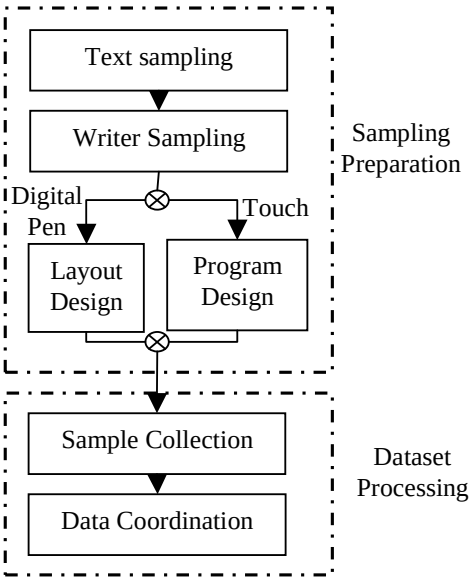


Figure 1. The flowchart of SCUT-COUCH2009-TL.

2.1 Writer Sampling

Generally, our strategy in selecting candidate writers is combining deliberation with randomness. To comprehend our dataset, we have implemented the sample collection in two schools: South China University of Technology and Polytech Nantes, which come from Guangzhou China and Nantes France respectively. Of course, in both cities selected writers were fluent to write Chinese texts. With the general set of candidate writers confirmed, we started a random selection of writers. The writer distribution on gender and city is shown in the following tables (Table 1~2).

Table 1. Gender distributions of writers

Items	Proportion
Male	79%
Female	21%

Table 2. City percentage of writers

City	Proportion
Nantes France	17%
Guangzhou China	83%

2.2 Collecting Devices

There exist many available devices and pens which can record the handwritten trajectory. It can be generally classified as touch screen devices or digital pens. The former are more portable and universally used as handwriting input devices in daily life, but you can only write on the screen. In the latter case, two kinds of technology are available. On one hand, you can use regular paper and write freely anywhere you want as long as you have clipped the receiver to the top of your paper; however the recorded trajectory would be not very accurate. On the other hand you can use digital pens with pre-printed dotted paper, which are much more accurate. Thus, we choose two typical tools: touch screen LCD and digital pen, which use regular paper and ultrasonic triangulation. As a result we have 66,913 character TS samples and 92,953 character digital pen samples.

2.3 Text Sampling

We choose China Daily corpus as the data source of our dataset. In the natural language processing field, China Daily is used as Chinese written language corpus, since it covers a comprehensive topics such as politics, economics, science and technology, culture, et al. [8] Using corpus as our data source instead of chaotic electronic texts demonstrates three advantages: (1). Linguistic context is automatically built in; (2). Dataset can be easily expanded with tremendous texts to sample from; (3). More frequently a character occurs, more training samples it possess.

We divide texts into 7 subsets according to topics: domestic news, international news, sports news, economic news, cultural news, academic trends, and education weekly. By doing this, we open the use of this dataset not only for purely recognition purposes but also for Information Retrieval tasks such as document categorization. All the topics are popular for nowadays people. Then we sampled texts with a stratified random manner. To reserve more data for future expansion, we only use texts of the China Daily 2009, date from January to June.

2.4 Collection Program and Layout Design

Since we have two collecting devices, we have designed collection program for LCD touch screens and layout for digital pens respectively.

2.4.1 Collection Program Design

Collection Program for touch screen LCD is running on the PC, so we should design a user interface program for writers. As illustrated in Figure 2, in our program, the input area is designed to be a box of static size, large enough to

accommodate as many characters as possibly required. The original texts to copy are laid on the left.

2.4.2 Layout Design

When we use paper to collect texts, it is the layout that serves as an interface to writers. Obviously, how to design the layout to make it friendly and informative is a nontrivial task. The design of the layout should fulfill two criteria. First, the layout should be simple and clear. Each form is divided into five distinct blocks: guideline block, text block to copy, writer's information block (name, age and gender), digital pen block, and writing block.

Second, as the digital pen receiver only permits writing with a good resolution in a small region (about the A5 paper size), we put the writer's information block, writing block at the right of form. Also we compress the writing guidelines to give more space reserved for handwriting.

After several recursions of feedback and modification, the final layout is illustrated in Figure 3. Each form is identified by an index, and each index means the topic of the forms and the paragraph number, e.g. **体育新闻** 1 (Sports News 1) means that the topic is sports news, and the paragraph index number is 1.

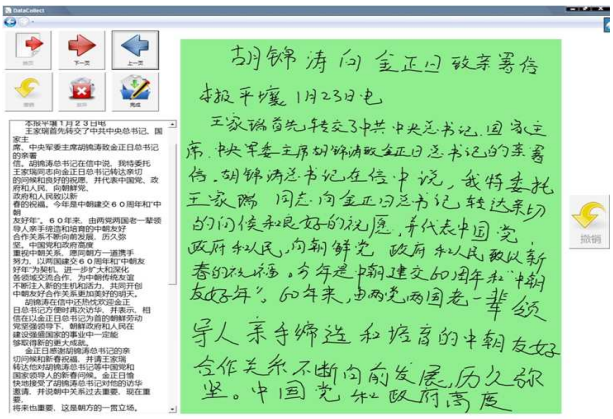


Figure 2. An illustration of layout for LCD touch screen.

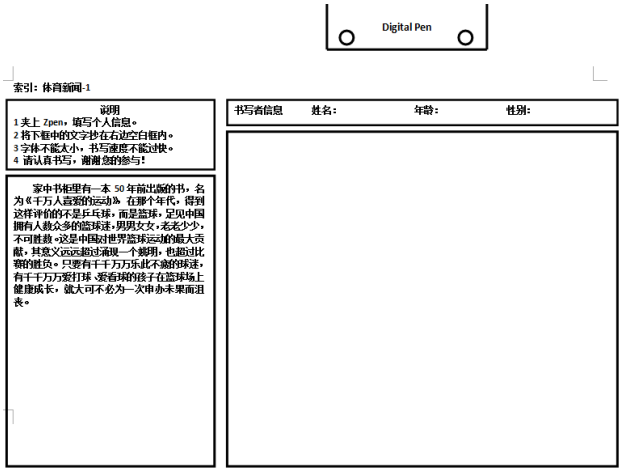


Figure 3. An illustration of Layout.

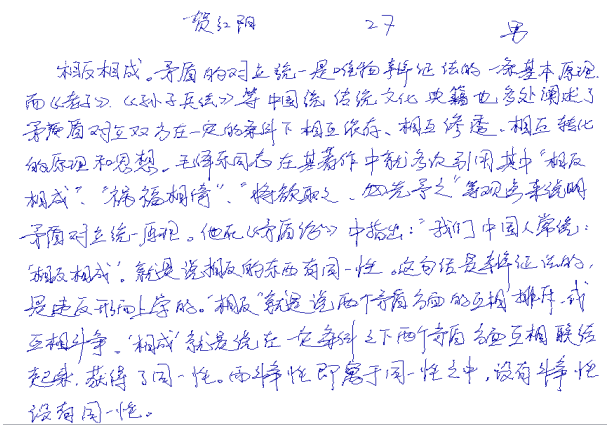


Figure 4. Sample of handwriting.

3. Dataset Processing and Analysis

Collected handwriting is transferred to computer, and saved as UNIPEN format [15]. The processing includes line segmentation, error correction, and dataset analysis. A sample of handwriting text is shown in figure 4.

3.1. Line Segmentation

To design a text recognition system, many stages are involved. One of them consists in segmenting the text into lines. In this dataset, the ground truth will be given at the line level so that evaluation can be done independently of the line segmentation process. As shown in figure 6, each sample has several lines. The illustration of line segmentation result is shown in figure 5.

3.2. Error Correction

A few samples, as shown in Figure 6, with wrongly characters and unacceptable errors have been detected while we were checking the collected data. We have corrected such examples by hand.

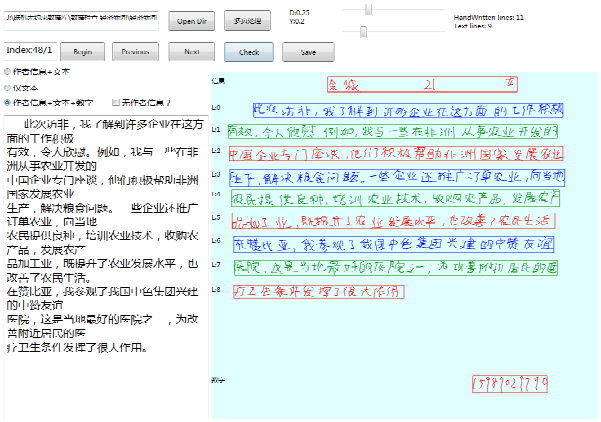
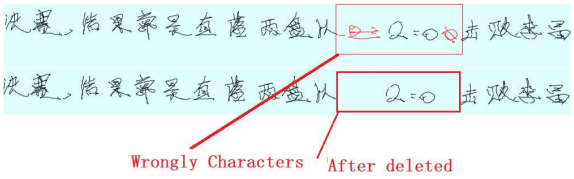
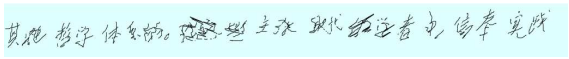


Figure 5. An illustration of Line segmentation program



(a) Wrongly characters and its correction

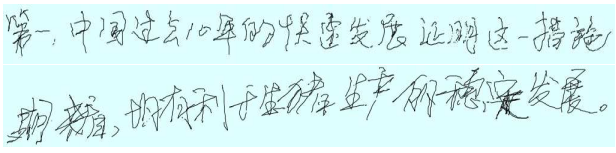


(b) The sample exhibits unacceptable errors

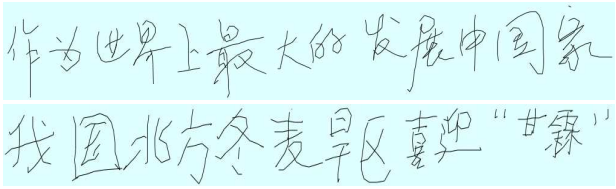
Figure 6. Error correction.

3.3. Dataset Analysis

SCUT-COUCH2009-TL dataset is a publicly-available unconstrained online handwritten Chinese text lines dataset, which is useful for online Chinese handwriting recognition. Figure 7 shows different kinds of samples we collected.



(a) Samples collected with digital pen



(b) Samples collected by touch screen LCD

Figure 7. Different kind of samples.

We have collected 421 legible digital pen sample texts, and 80 touch screen LCD sample texts. There represent 159,866 characters, including letters, punctuations besides Chinese characters, and these characters lead to 8,809 lines. By simple computation, we get following statistics: the total respective numbers of characters, paragraphs, article and line are shown in Table 3.

Moreover, the lexicon of the dataset has 2,632 entries (i.e. different classes of characters). In other words, each character in average occurs 60.73 times. The coverage over Sogou Text Classification Corpus [16] with 22,320,731 characters is 86%. This coverage shows our sampling schemes have a good representative capability.

Table3. The total respective numbers of character, paragraph, article and line.

Items	Character amount	Paragraph or article amount	Line amount
Touch	66,913	80(Article)	4,813
D. pen	92,953	421(Para.)	3,996
Total	159,866	-	8,809

4. Online Text Line Recognition

As a baseline for giving a first reference for this dataset, a new method for segmenting and recognizing these unconstrained online Chinese text lines is presented for evaluating the practical of the dataset.

4.1. Algorithm outline

The algorithm we proposed, first, over-segments the given text line into radical segments. Next, a dynamic programming scheme is used to evaluate each segmentation path. Finally, the optimal merging path can be obtained according to the recognition and semantic information, which are given by the isolated character classifier and trained from the dataset respectively. The algorithm's outline is shown in figure 8.

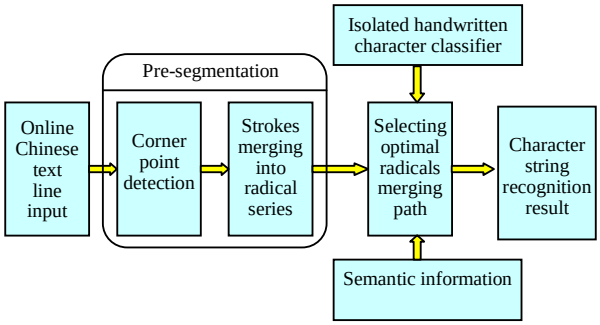


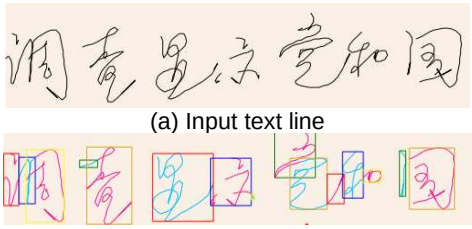
Figure 8. Algorithm flowchart.

4.2. Presegmentation

A sequence of sample points $\{x_i, y_i \mid i = 1, 2, .., N\}$ is obtained after some preprocessings, including normalization and resampling, where N is the total number of points. Next, we should find out all the potential segmentation points. A point p satisfying formula (1) would be considered as a corner point, i.e. a point with a high curvature, where t is the last corner point, and θ_i is the orientation of the point i.

$$\left| \theta_p - \frac{\sum_{i=t}^{p-1} \theta_i}{p - t} \right| > \frac{\pi}{2} \quad (1)$$

Using the corner points, we segment the point sequence into strokes. After all the potential segmentation points have been detected, we merge the strokes into radical segments using stroke order information and geometric information. An over-segmentation example is shown in figure 9.



(a) Input text line

(b) Presegmentation result

Figure 9. Over-segmentation.

4.3. Selection of candidate merging path set

An input text line is over-segmented into radical segments using the pre-segmentation processing. To reduce the computation, a rough evaluation only uses radicals' geometric feature to evaluate every possible character point sequence series. Each character point sequence has a score. Each character point sequences' rough evaluation score is the score summation of every character hypothesis in the series. The relative smaller set of candidate merging paths is obtained according to the rough evaluation score using dynamic programming method, [17].

4.4. Evaluation of optimal merging paths

The precise evaluation based on linguistic information will be used to select optimal merging path and optimal recognition results of the corresponding character point sequences. Suppose M character sequences are obtained according to the merging path S. $CharSQ_i$ denotes to the ith character sequence, $1 \le i \le M$. $Char_i$ denotes the corresponding recognition result of $CharSQ_i$. According to MAP criterion, the optimal path $\hat{S}$ and string recognition result $\hat{Char}_1, \hat{Char}_2, \cdots, \hat{Char}_M$ satisfy the expression (2). Among it, $p(Char_1, \cdots, Char_M \mid CharSQ_1, \cdots, CharSQ_N, S)$ expresses recognition likelihood of $Char_1, \cdots, Char_M$ under merging path S.

$$\hat{Char}_1, \cdots; \hat{Char}_M, \hat{S} = \underset{Char_1, \cdots; Char_M, S}{\arg\max}\ p(Char_1, \cdots; Char_M \mid CharSQ, \cdots; CharSQ_M, S) \quad (2)$$

In addition, we make the assumption that:

$$p(Char_1, \cdots, Char_M \mid CharSQ_1, \cdots, CharSQ_M, S) = p(Char_1) \times p(Char_1 \mid CharSQ_1) \times \prod_{j=2}^{M} p(Char_j) \times p(Char_j \mid CharSQ_j) \times p(Char_j \mid Char_{j-1}) \quad (3)$$

This assumption means that one character sequence's recognition result is only dependent on its own sequence. For any character, we calculate the prior probability $p(Char)$ and the transition probability $p(Char_i \mid Char_j)$ from the dataset of original texts.

The isolated character classifier [18] used in the paper is trained for classifying all Chinese character (GB1 and GB2) Arab digits and English characters. For any character point sequence $CharSQ$, the isolate character classifier could give 10 candidate recognition result characters, each with a recognition score. Formula (4) (5) is used to get the optimal result $\hat{Char}$ and its likelihood. $Char_i^{cnd}$ represents the ith candidate recognition result of $CharSQ$; $Score_j^{cnd}$ denotes the recognition score of $Char_j^{cnd}$, 1<i<10.

$$p(Char_j^{cnd} \mid CharSQ) = \frac{\exp(Score_j^{cnd})}{\sum_{i=1}^{10} \exp(Score_i^{cnd})} \quad (1 \le i, j \le 10) \quad (4)$$

$$p(Char \mid CharSQ) = \arg\max_{i \in [1,10]} \{ p(Char_i^{cnd} \mid CharSQ) \} \quad (5)$$

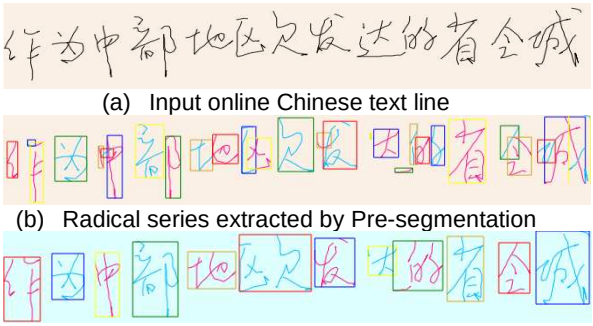
Using formula (3) to (5) in expression (2) we can get the optimal recognition result $\hat{Char}_1, \hat{Char}_2, \cdots, \hat{Char}_M$.

Table 4. The experiment results on dataset.

Topic/Device	Total character number	Recognition accuracy
Domestic news	66,320	56.50%
International news	22,848	59.51%
Sports news	21,937	55.30%
Economic news	20,858	53.45%
Cultural news	8,474	55.07%
Academic trends	6,407	51.89%
Education Weekly	12,434	60.39%
Touch screen LCD	66,320	56.50%
Digital pen	92,958	56.35%
Total	159,278	**56.41%**

5. Experiments and results

To evaluate the performance of the algorithm, experiments are carried out using the SCUT-COUCH2009-TL dataset. There are a total of 159,278 characters, and the presented algorithm recognition accuracy is 56%. The recognition accuracies based on topics and collecting devices are shown in Table 4. An example of segmentation and recognition result is shown in figure 10.



(a) Input online Chinese text line

(b) Radical series extracted by Pre-segmentation

(c) Segmentation result.

作为中部地欧发达的省会域

(d) Recognition result.

Figure 10. Segmentation and recognition result using proposed algorithm

6. Conclusion

We describe a publicly available database, SCUT-COUCH2009-TL, for research of unconstrained online Chinese text line recognition. The database contains 159,866 character samples written by 157 persons, in 7 topics. Preliminary experiments on the dataset using an over-segmentation based segmenting and recognizing algorithm shows that the dataset fulfills the goals it was practically challenging. The SCUT-COUCH2009-TL dataset and its latest detailed information are available at http://www.hcii-lab.net/data/scutcouch/.

Acknowledgments

We would like to thank Dapeng Tao, Huaide Zhan, etc. for helping to supervise data collection. We would also like to thank all those warmly cooperative volunteers. This work is supported in part by the research funding of NSFC (no. U0735004, 60772216) and GDNSF (no. 07118074) and from Atlanstic/University of Nantes.

References

[1] http://www.hcii-lab.net/data/scutcouch/

[2] J. Hull, A database for handwritten text recognition research, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994, 16(5): 550-554.

[3] U.-V. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, *International Journal Document Analysis and Recognition*, 2002, 5(1): 39-46.

[4] D-H. Kim, Y.-S. Hwang, S.-T. Park, E.-J. Kim, P. S.- H, S.-Y. Bang, Handwritten Korean character image database PE92, *IEICE Transactions Information and Systems*, 1996, E79-D(7): 943-950.

[5] V. Margner, H. El Abed, Databases and competitions: strategies to improve Arabic recognition, In: *Arabic and Chinese Handwriting Recognition*, S. Jaeger and D. Doermann (Eds.), LNCS Vol.4768, Springer, 2008, pp.82-103.

[6] U. Bhattacharya, B-B. Chaudhuri, Databases for research on recognition of handwritten characters of Indian scripts, *Proceeding of the 8th International Conference on Document Analysis and Recognition*, 2005, pp. 789-793.

[7] J. Guo, Z. Lin, H. Zhang, A new database model of offline handwritten Chinese characters and its applications, Acta *Electraonic Sinica*, 2000, 28(5): 115-116.

[8] T-H. Su, T-W. Zhang, D-J. Guan, Corpus-based HIT-MW database for offline recognition of generalpurpose Chinese handwritten text. *International Journal Document Analysis and Recognition*, 2007, 10(1): 27-38.

[9] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet, UNIPEN project of on-line data exchange and recognizer benchmarks, *Proceeding of the 12th International Conference on Pattern Recognition*, 1994, pp.29-33.

[10] M. Nakagawa, T. Higashiyama, Y. Yamanaka, S. Sawada, L. Higashigawa, K. Akiyama, On-line handwritten character pattern database sampled in a sequence of sentences without any writing instructions, *Proceeding of the 4th International Conference on Document Analysis and Recognition*, 1997, pp.376-381.

[11] K. Matsumoto, T. Fukushima, M. Nakagawa, Collection and analysis of on-line handwritten Japanese character patterns, *Proceeding of the 6th International Conference on Document Analysis and Recognition*, 2001, pp.496-500.

[12] Y. Li, L. Jin, X. Zhu, T. Long, SCUT-COUCH2008: A comprehensive online unconstrained Chinese handwriting dataset, *Proceeding of 11th International Conference on Frontiers and Handwriting Recognition*, 2008, pp.165- 170.

[13] D-H. Wang, C-L. Liu, J-L. Yu, X-D. Zhou, CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters, *Proceeding of the 10th International Conference on Document Analysis and Recognition*, 2009, pp.1206-1210.

[14] C. Viard-Gaudin, P.M. Lallican, S. Knerr, P. Binter, The IRESTE On/Off (IRONOFF) dual handwriting database, *Proceeding of the 5th International Conference on Document Analysis and Recognition*, 1999. pp. 455-458.

[15] http://hwr.nici.kun.nl/unipen/unipen-history.html

[16] http://www.sogou.com/labs/dl/c.html

[17] D. Eppstein, Finding the k shortest paths, *35th Annual Symposium on Foundations of Computer Science (FOCS 1994)*, 1994, pp.154-165

[18] T. Long, L-W. Jin, Building compact MQDF classifier for large character set recognition by subspace distribution sharing, *Pattern Recognition*, 2008, vol. 41, no. 9, pp.2916-2925.

[19] F. Qiang, Q-X.Ding, C-S. Liu, J. Yan, R. Zheng, A hidden Markov model based segmentation and recognition algorithm for Chinese handwritten address character strings,