2012 10th IAPR International Workshop on Document Analysis Systems

Effect of Text/Non-text Classification for Ink Search employing String Recognition

Tomohisa Matsushita, Cheng Cheng, Yujiro Murata, Bilan Zhu and Masaki Nakagawa
Department of Computer and Information Sciences
Tokyo University of Agriculture and Technology
Tokyo, Japan
e-mail: 50011646201@st.tuat.ac.jp

***Abstract*—This paper presents the effect of text/non-text classification for ink search which employs string recognition. Pen or touch interfaces provides the benefit that users can write text and draw figures without changing the device or mode, but line drawings are troublesome for ink search. We propose the insertion of text/non-text classification before ink search and show its effect. For ink search, we employ our own engine to search keywords in the candidate lattice prepared by on-line handwritten Japanese text recognition, since this method produces higher search rate for Japanese text in digital ink than word spotting without ink recognition.**

Keywords-keyword search; on-line handwritten documents; digital ink; text/non-text classification; text line grouping; character recognition

I. INTRODUCTION

People can express almost anything by handwriting without any mental burden and any change of devices or modes. They can write text, figures, tables, formulas and so on by a pen or finger. Today, handwriting devices like tablets, tablet PCs, smart phones, pad PCs, digital pens and electronic white boards have been used by many people. Furthermore, electronic papers and paper-like PCs will be used in near future. In such an environment, so called digital ink (a time-sequence of pen-tip coordinates) will be stored and accumulated. Then, the search of keywords is necessary as it is essential for searching Web contents.

Searching keywords in printed or handwritten documents has been studied for many years; *Manmatha et al*. used Euclidean distance mapping and the SLH algorithm for searching [1]. *Lu et al*. presented a document retrieval technique that is capable of searching document images [2]. They do not depend on character recognition so that they are expected to work on degraded documents for which character recognition is not effective. On the other hand, *Marukawa et al*. employed character recognition and proposed a search method, which reduced search loss from incorrect recognition results by using two or more character recognition candidates and a confusion matrix [3]. *Ota et al*. further extended the above idea by producing search terms considering mis-segmentation as well as mis-recognition with the confidence from them and from bi-gram [4]. *Imagawa et al*. investigated reliability of recognition results using a neural network and they showed that both the recall rate and the precision rate were improved by their method [5].

Searching keywords in on-line digital ink has been gathering less attention. Due to the proliferation of pen-based or touch devices, however, demand for ink search is expanding. Between off-line paradigm and on-line paradigm, features, suitable segmentation methods and character recognition methods differ so that search methods will also differ.

Early work was made by *Lopresti et al*. [6]. They proposed ink search at several level of representations. For character level matching they showed performance prediction based on simulated text and fuzzy string matching. For stroke level, they presented stroke level matching algorithm and its performance. They continued the former research and formulated approximate string matching [7] and fuzzy logic [8], which was also valid for noisy text after OCR (off-line paradigm). *Senda et al*. presented a method to retrieve handwritten memos from handwritten queries [9]. They employed matching at the feature level. *Oda et al*. proposed a search method for on-line handwritten text [10]. This method employed writing-box-free handwritten text recognition [11]. *Cheng et al*. improved this method [12] They further improved it using noise reduction [13].

Regardless of whether off-line or on-line, previous studies have only focused on text, although handwriting is the most natural way to express figures, tables and so on with text.

Recently, free-format handwriting recognition has been also gathering attention, which has to consider recognition and segmentation at the same time. *Shilman et al.* proposed a segmentation method utilized adjacencies of strokes [14]. On the other hand, *Zhou et al*. proposed a classification method for text/non-text ink stroke in Japanese handwriting [15]. Also, *Zhou et al*. proposed a text line grouping method in on-line handwritten Japanese documents [16]. We have confirmed that the grouping method is able to improve text/non-text classification results.

In this paper, we employ the text/non-text classification method before ink search into text and non-text mixed digital ink, and we show its effect.

This paper is organized as follows. Section II describes our ink search method employed string recognition. Section III describes text/non-text classification. Section IV describes experiments and evaluations of our method. Section V draws conclusion.

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.33

II. INK SEARCH EMPLOYING STRING RECOGNITION

The search system employed in this paper is based on handwritten character recognition (HWX) [13]. For our database of digital ink, it is superior to word spotting [17] in our study. The system is given as input a keyword (i.e., a short string of Japanese character codes) and searches it in digital ink.

A. Basic Structre

The basic structure of this search system is shown in Figure 1. and Figure 2. It is composed of two phases.

The first phase is csr-lattice generation for search (Figure 1). Digital ink documents are recognized and multiple ways of segmentation and multiple candidates for each segmented character are represented in the form of a candidate segmentation and recognition lattice. Documents and their candidate segmentation and recognition lattices (csr-lattices) are stored.

The second phase is search (Figure 2). A user inputs a keyword and the search system searches the target keyword into csr-lattices and outputs the locations the keyword in digital ink documents.

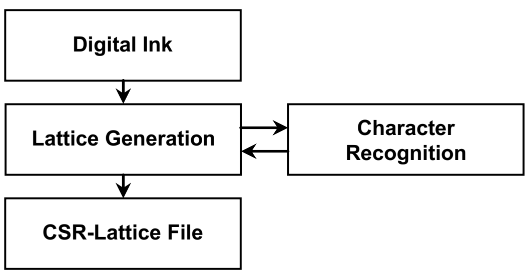


Figure 1. Csr-lattice generation phase.

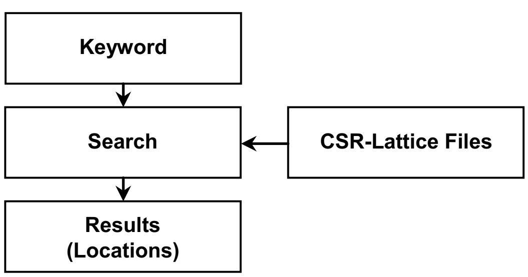


Figure 2. Search phase.

B. Before the search

Before the search is made, the system hypothetically segments the digital ink, divide an input sequence of pen strokes into segments each of which may form a single character pattern.

These segments are then considered for composite segments. That is, the system forms candidates of composite segments by combining (up to some determined number of) neighboring segments.

A recognition engine is applied to each segment or combined segment and a list of multiple candidate classes is associated with it where each class has a score representing how close it is to the segment. Multiple ways of segmentation and multiple candidates for each segment/combined segment are represented in the form of a candidate segmentation and recognition lattice (csr-lattice) as shown in Figure 3. Here, the segment for the digital ink “日月” is recognized and the following classes: “明”, “朋”, “門” are associated with their scores (for their scores see Figure 4).

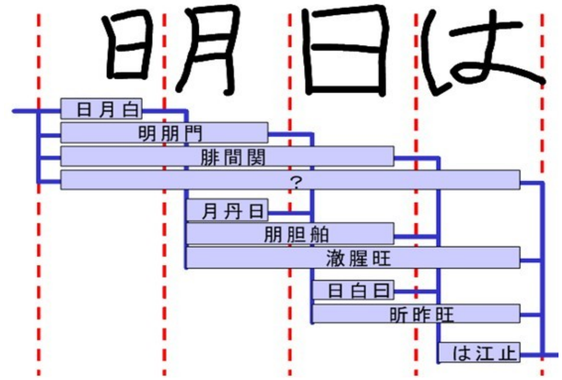
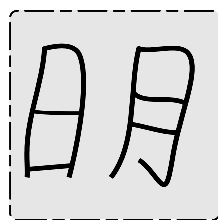


Figure 3. Csr-latticefor the digital ink “明日は”.



Rank	Code	Score
1st	明	891
2nd	朋	873
3rd	門	860
⋮	⋮	⋮

Figure 4. Recognition candidates for a combined segment “日月”.

C. Executing the search

Given a search keyword, the system searches for the keyword into the csr-lattices and find paths which match the input keyword. For each such path, the system outputs the positions of the start and end strokes of the path.

D. Improved search method

The method in [1], 10] for searching a keyword into csr-lattices is the Viterbi search. Recall and precision of the search depends on the correctness of the crs-lattices.

In order to increase the precision of its output as well as speed up the search, it is desirable to prune bad candidates from the search space. Two ways of pruning candidates are employed. The following candidates are pruned:

1) Candidates whose ranks in character recognition are below a threshold ***Tr****.*

2) Candidates whose recognition scores are below a threshold score ***Ts****.*

Nevertheless, the precision of the Viterbi search in [1, 10] is often affected by tiny discrepancies between strokes in digital ink. This happens when the correct set of strokes contain a subset which produces a high score as well. For example consider the two high scoring matches for the keyword “明日が” in Figure 5.

Figure 5. (a) The correct set of strokes for the keyword “明日が”. (b) A subset of the correct set of strokes which give a high score as well.

Although the strokes in (a) produces the highest score (since “が” matches “が” better than “か”), the strokes in (b) also produces a significantly high score for the input keyword. Since both (a) and (b) are output by the search, (b) reduces the search’s overall precision. This effect is especially noticeable in the case of the Japanese language since many Japanese hiragana/katakana characters have two slightly different neighbors: one for voiceless sound and one for voiced sound.

We use a straight-forward and efficient solution for this problem. We first sort the results from the Viterbi search by their scores. Then, starting from the result with the highest score, we remove all the results with segments of over a certain length that overlaps it.

E. Learning Parameters

For setting ***Ts*** and ***Tr***, we use the database “TUAT Nakagawa Lab. HANDS-kuchibue_d-97-06” (hereafter we refer to as Kuchibue).

Kuchibue is a set of on-line handwritten text patterns written by 120 participants with each composed of 11,962 character patterns written by a single participant (10,152 character patterns in meaningful context, 1,810 character patterns without context).

Oda et al. used 57 Kuchibue sets (i.e. 578,664 character patterns) to find the optimal ***Ts*** and ***Tr*** [1, 10]. We use these same values in all our experiments.

III. APPLYING TEXT/NON-TEXT CLASSIFICATION

Non-text strokes may disturb ink search. It is particularly troublesome for generating csr-lattices. In this paper, we insert text/non-text classification in the csr-lattice generation phase of this search system (Figure 1) and try to remove non-text strokes. The revised structure is shown in Figure 6.

A. Text/Non-text Classification

Zhou et al. proposed a classification method for text/non-text ink strokes in Japanese handwriting [5]. In this paper, we employ this method.

This method is based on Markov random fields (MRFs), which effectively utilize the spatial relationship between strokes. Support vector machine (SVM) classifiers are trained for individual stroke and stroke pair classification, and by converting the SVM outputs to probabilities, the likelihood clique potentials of MRF are derived.

This method is an effective approach for grouping text lines in on-line handwritten Japanese documents by combining temporal and spatial information. With decision functions optimized by supervised learning, the approach has few artificial parameters and utilizes little prior knowledge.

B. Text Line Grouping

We use text line grouping for fixing results of text/non-text classification. The improved structure of the text/non-text classification is shown in Figure 7.

Text strokes extracted by text/non-text classification are grouped into text lines. Bounding boxes of grouped text lines are located. If there are non-text strokes in the text lines' bounding boxes, these strokes are reclassified to text strokes (fixation).

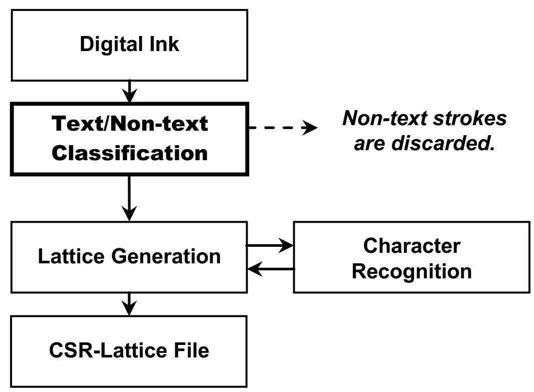


Figure 6. Revised csr-lattice generation phase.

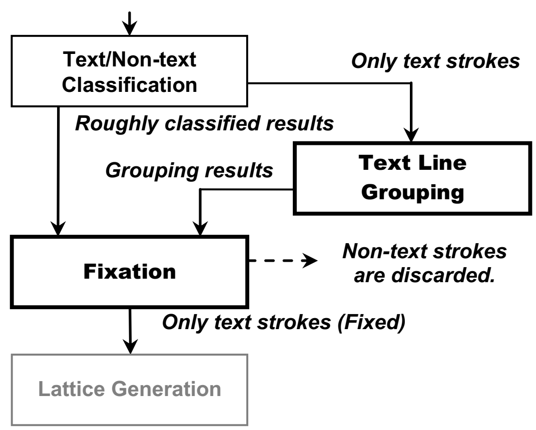


Figure 7. The improved structure of the text/non-text classification.

C. Example of CSR-Lattice Generation Phase

An example of classification is shown in Figure 8. This example is a weekly schedule table in nondescriptive Japanese style. We hope to classify its ruled lines to non-text and the other strokes to text and use only text strokes for the lattice generation.

First, text/non-text classification generates roughly classified results. At this stage, classified non-text includes some text strokes (shown in red) and classified text excludes these strokes.

Next, bounding boxes are output by text line grouping, and then, strokes detected in the boxes are reclassified to text strokes (fixation).

Finally, fixed text strokes are used for the csr-lattice generation.

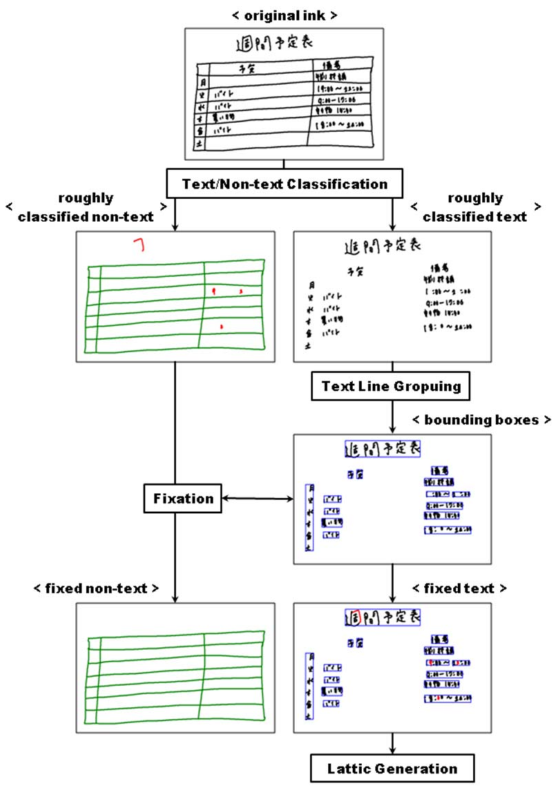


Figure 8. Example of csr-lattice generation phase

IV. Experiment and Evaluation

A. Criterions for Text/Non-text Classification

We evaluate the text/non-text classification by counting the *accuracy rate*. The *accuracy rate* is defined by (1).

$$accuracy\ rate = \frac{Number\ of\ corrent\ classified\ strokes}{Number\ of\ all\ strokes} \quad (1)$$

B. Criterions for Search

We evaluate the search system by counting the *f-measure*. The *f-measure* is defined by (2).

$$f = \frac{2}{\frac{1}{r} + \frac{1}{p}} \quad (2)$$

In (2), r is called recall, p is called precision and they are expressed by (3)-(4). The recall rate expresses the tolerance of the system to search losses, while the precision rate expresses the system's tolerance to search noises. The *f-measure* expresses the overall performance of the search method.

$$r = \frac{Number\ of\ correct\ search}{Number\ of\ search\ keywords\ in\ target\ data} \quad (3)$$

$$p = \frac{Number\ of\ correct\ search}{Number\ of\ searched\ items\ (include\ noise)} \quad (4)$$

C. Database used for Experiments

We employ the database "TUAT Nakagawa Lab. HANDS-Kondate_t_bf-2001-11" (hereafter we refer to as Kondate). Kondate is a set of on-line handwritten text patterns written by 100 participants. Its patterns consist of about 30-40 pages per a participant and 67 participants wrote about 10 text/non-text mixed pages per a participant. Furthermore, 36 participants' patterns of them had been used as the "learning pattern" for text/non-text classification. Therefore, we use the remaining 31 participants' patterns as the "test pattern" for these experiments. The 31 participants' patterns consist of total 310 pages.

D. Evaluation Results

1) Results of Text/Non-text Classification

We tested two classification methods with the test patterns of the Kondate database. One method is applying text/non-text classification and another one is applying fixation after the classification. The results are shown in TABLE I. Fixation increases the accuracy by 0.36 point.

TABLE I. Results of Text/Non-text Classification

Method	*Accuracy*
Original Classification	95.47%
Fixing Classification Results	95.83% (+0.36 point)

2) Results of Search

We tested following cases.

a) Do not apply text/non-text classification.

b) Apply text/non-text classification.

c) Apply also fixation after text/non-text classification.

The results are shown in Figure 9. If keyword length is short, precision is not enough (too many noises). As keyword length becomes longer, higher precision is derived. Text/non-text classification is effective as shown by the higher results of *b)* than those of *a)*. Additionally, *c)* is better than *b)*. It shows the effectiveness of the fixation by text bounding boxes.

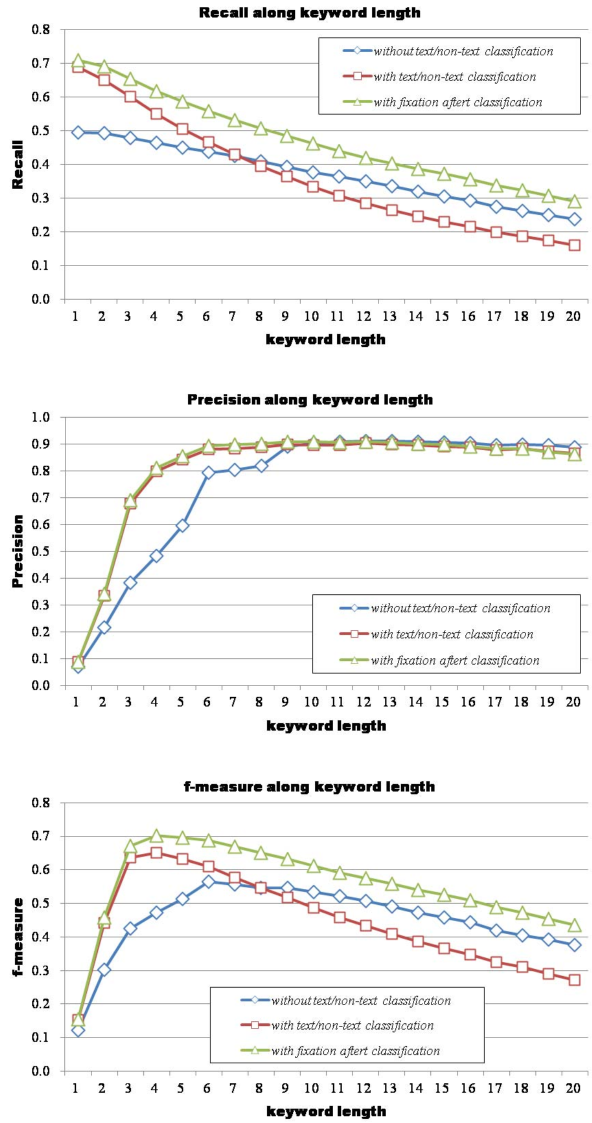


Figure 9. The results of search.

V. CONCLUSION AND FUTURE RESEARCH

In this paper, we proposed text/non-text classification for ink search employing string recognition. Non-text strokes are particularly troublesome for ink search.

It is confirmed that applying text/non-text classification to preparation phase (before generating lattices) of ink search is effective, fixing classification results by text line grouping contributes to the accuracy of text/non-text classification, and it also improves the performance for ink search.

REFERENCES

[1] R. Manmatha, H. Chengfeng and E.M. Riseman, "Word Spotting: A New Approach to Indexing Handwriting," In: Proc. of the IEEE Computer Vision and Pattern Recognition Conference, San Francisco, CA, USA, pp. 631-637, June 1996.

[2] L. Shijian, L. Linlin and C. L. Tan, "Document image retrieval through word shape coding," IEEE Trans. PAMI, vol. 30, No. 11, pp. 1913-1918, November 2008.

[3] K. Marukawa, H. Fujisawa and Y. Shima, "Evaluation of Information Retrieval Methods with Output of Character Recognition Based on Characteristic of Recognition Error (Japanese)," Trans. IEICE, vol. J79-D-II, No. 5, pp. 785-794, May 1996.

[4] M. Ohta, A. Takasu and J. Adachi, "Full-text search Methods for OCR-recognized Japanese Text with Misrecognized Characters (Japanese)," Trans. IPSJ, vol. 39, No. 3, pp. 625-635, March 1998.

[5] T. Imagawa, Y. Matsukawa, K. Kondo and T. Mekata, "A document image retrieval technique using each character recognition reliability (Japanese)," TECHNICAL REPORT OF IEICE PRMU99-72, pp. 63-68, 1999.

[6] D. Lopresti and A. Tomkins, "On the searchability of electronic ink," Proc. 4th International Workshop on Frontiers in Handwriting Recognition, pp. 156-165, December 1994.

[7] D. Lopresti and A. Tomkins, "Block Edit Models for Approximate String Matching," Theoretical Computer Science, vol. 181, No. 1, pp. 159-179, July 1997,

[8] D. Lopresti and J. Zhou, "Retrieval Strategies for Noisy Text," Proc. 5th Annual Symposium on Document Analysis and Information Retrieval, pp. 255-269, April 1996.

[9] S. Senda, M. Hamanaka and K. Yamada, "MemoPad: Software with functions of Box-free Japanese Character Recognition and Handwritten Query Search (Japanese)," TECHNICAL REPORT OF IEICE PRMU99-75, pp. 85-90, September 1999.

[10] H. Oda, A. Kitadai, M. Onuma and M.Nakagawa, "A search method for on-line handwritten text employing writing-box-free handwriting recognition," Proc. 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004), Tokyo, Japan, pp. 545-550, 2004.

[11] M. Nakagawa and M. Onuma, "On-line handwritten Japanese text recognition free from Constrains on Line Direction and Character Orientation," Proc. 7th International Conference on Document Analysis and Recognition (ICDAR-7 2003), pp. 519-523, Edinburgh, Scotland, August 2003.

[12] C. Cheng , B. Zhu andM. Nakagawa, "Ink Search employing Japanese String Recognition," CJK Joint Workshop on Pattern Recognition (CJKPR 2009), Nanjing, China, pp. 925-928, December 2009.

[13] C. Cheng, B. Zhu, X. Chen and M. Nakagawa, "Improvements in Keyword Search Japanese Characters within Handwritten Digital Ink," 10th International Conference on Document Analysis and Recognition (ICDAR-10 2009), Barcelona, Spain, pp. 863-866, October 2009.

[14] M. Shilman, P. Viola, and K. Chellapilla, "Recognition and Grouping of Handwritten Text in Diagrams and Equations," 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004) , Tokyo, Japan, pp. 569-574, 2004.

[15] X.-D. Zhou and C.-L. Liu, "Text/Non-text Ink Stroke Classification in Japanese Handwriting Based on Markov Random Fields, Document Analysis and Recognition," 9th International Conference on Document Analysis and Recognition (ICDAR-9 2007), Curitiba, Brazil, vol. 1, pp. 377-381, December 2007.

[16] X.-D. Zhou, D.-H. Wang and C.-L. Liu, "A robust approach to text line grouping in online handwritten Japanese documents," Pattern Recognition, vol. 42, No. 9, pp. 2077-2088, September 2009.

[17] P. Luangvilay, B.Zhu and M. Nakagawa, "An On-line Handwritten Text Search Method based on Directional Feature Matching," Proc. 11th International Conference on Document Analysis and Recognition (ICDAR-11 2011), pp. 683-686, Beijing, China, Sepember 2011.