8—12

Image Based Word Retrieval Method for Unrestricted Textline Direction Documents

Takakazu Noge[1]
Graduate School of Information Science
Iwate Prefectural University

Takahiko Horiuchi[2]
Faculty of Software and Information Science
Iwate Prefectural University

Abstract

It is very important to perform a full-text retrieval search of document information accumulated in the past. Although the retrieval technologies for ascii text documents have been established, the highly precise character retrieval from the image based documents such as a bitmap image is not easy. In this paper, a word retrieval technique for a bitmap Japanese document image described with various layouts is proposed. The technique consists of character sequence extraction stage and word retrieval stage. As a result of the experiment using actual documents in of vertical writing and lateral writing mixture, it will be shown that the proposed technique is effective.

1 Introduction

Documents accumulated in the past has been processed electronically as a digital documents. In order to acquire required information from those document images, a full-text search is important technology. Although retrieval of the ascii text document is established as technology of a full-text search, retrieval of a bitmap document image has not resulted in practical use.

Conventional techniques about retrieval of a bitmap document image used OCR as pre-processing and stored the text by OCR with a bitmap [1],[2]. Although this method enables high-speed retrieval, since the accuracy of Japanese handwriting OCR is still not perfect and various devices are needed. In addition, there is also a problem of requiring a memory. The trial in which it will search directly from a bitmap document in recent years is also made [3]-[7]. Refs. [3] and [4] proposed retrieval techniques for English which consists of only lateral writing, and cannot be applied to Jpanese document where lateral writing and vertical writing are intermingled. Although Refs. [5]-[7] are carrying out for Chinese characters, it is difficult for Japanese to intermingle Chinese character and kana in which the character of form differs, and to apply those techniques for Chinese characters to kana characters.

In this paper, a word retrieval technique for the Japanese document with which vertical writing and lateral writing were intermingled is proposed. In character sequence extraction stage, a character sequence domain is extracted using a layout knowledge of a Japanese document, and word retrieval stage is performed in each character sequence domain. For processing time curtailment, matching of the head of only one character is performed for a query word. Moreover, only when it agrees, matching processing after the second character is performed in order. It experiments to 300 sheets of papers and a book, and performance is verified.

2 Word Retrieval Method

In this section, system devised by this paper explains in detail. This system is roughly divided into two stages. The first stage is character sequence extraction stage performed to the target bitmap document image. The second stage is word retrieval stage performed to the character sequence domain extracted by preprocessing.

About the first character sequence extraction stage, the 8-direction black pixel connecting method is first performed in the target document image. The connecting method is the method of connecting the black pixel, when the black pixel domain which adjoins in the 8-connection directions in the black pixel in an input document image is searched and a black pixel exists. And it carries out until a black pixel stops connecting this processing, and it asks for the minimum rectangle domain circumscribed to the domain of all this connected black pixel. Next, the rectangle which the domain overlapped in these rectangular domains is unified by those circumscription rectangles. Let the rectangle obtained as a result be a basic rectangle. Next, this basic rectangle is considered to be one character. Generally a connection margin is set as a basic rectangle from knowledge with the common Japanese document that it is narrower than spacing between characters, and

[1] Address: Sugo 152-52, Takizawa, Iwate, Iwate 020-0193 Japan. E-mail: **g231a026@edu.soft.iwate-pu.ac.jp**
[2] Address: Sugo 152-52, Takizawa, Iwate, Iwate 020-0193 Japan Japan. E-mail: **tah@soft.iwate-pu.ac.jp**

connection stage is performed. If it states in detail, it judges whether other basic rectangles exist in the margin of each basic rectangle. If it exists, the basic rectangle will be connected. Let the minimum rectangle domain circumscribed to these two connected basic rectangles be a new basic rectangle. And a margin is newly set also to this new basic rectangle.

Let W and H are be the width and the height of a basic rectangle, respectively. Let mW and mH be the width and the height of a margin, respectivery. Ten the initial margins can be expressed as follows:

$$mW = \begin{cases} W/3 & \text{(In the case of W>= 0.6xH and H>= 0.6xW)} \\ W/3 & \text{(In the case of W<0.6xH)} \\ W/5 & \text{(In the case of H<0.6xW)} \end{cases},$$

$$mH = \begin{cases} H/3 & \text{(In the case of W>= 0.6xH and H>= 0.6xW)} \\ H/5 & \text{(In the case of W<0.6xH)} \\ H/3 & \text{(In the case of H<0.6xW)} \end{cases}.$$

W and H are made into the width and the height of a new basic rectangle, and mW and mH are made into the width and the height of a margin, and w and h are made the vertical and lateral of the connection direction of a basic rectangle. About a margin setup of the second henceforth, it is expressed as follows:

$$mW = \begin{cases} 1 & \text{(In H>2xW, the connection direction is at w.)} \\ 1.5\times H & \text{(In W>2xH, the connection direction is at h.)} \\ W/5 & \text{(Regardless of the connection direction, in H<2xW and W<2xH, it is.)} \end{cases},$$

$$mH = \begin{cases} 1.5\times W & \text{(In H>2xW, the connection direction is at w.)} \\ 1 & \text{(In W>2xH, the connection direction is at h.)} \\ H/5 & \text{(Regardless of the connection direction, in H<2xW and W<2xH, it is.)} \end{cases}.$$

And in each basic rectangle, other basic rectangles do not exist the above processing in a margin, and it carries out until it stops performing connection processing. As for this technique, vertical writing and lateral writing can extract a character sequence domain with this algorithm as an advantage. A series of flows is shown in Fig. 1 about character sequence extraction stage.

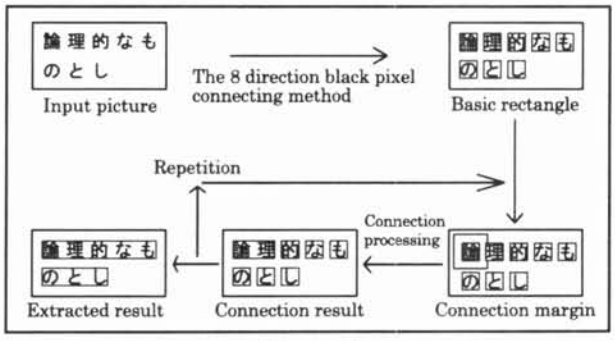

Figure 1.
Character sequence extraction algorithm

In the second word retrieval stage, matching of the head of one character of query words is performed to the inside of the extracted character line domain. And only when it agrees, matching processing after the second character is performed in order. In the matching technique, a thing called the simple similarity is used. The simple similarity can express them with $\cos\theta$ of the following formulas, when the character in input bitmap document image and retrieval character are expressed with the K-dimensional vectors a and b, respectively.

$$\cos\theta = (a,b) / (|a|\cdot|b|)$$

$$(0 \leqq \cos\theta \leqq 1)$$

Moreover, in case matching processing after the second character is performed, the circumference of the matched character is shifted and searched. This range to search is from immediately after the matched character to the half of the size of the character which the next searches. Within the limits of it is searched with the same matching processing.

In case the above word retrieval is carried out, a threshold is determined within the limits of $\cos\theta$, and it is considered that it agreed in more than a threshold. When the above processing is performed until all retrieval characters agreed, and all agree, the range is outputted as a retrieval result range.

By this technique, after a user inputs a query word, since the system made a query word into bitmap, the font of a query word can be freely chosen from the inside of a system. Font size is distinguished from the width and the height of a character sequence which were extracted on the occasion of character sequence extraction stage. A series of flows about word retrieval stage is shown in Fig. 2.

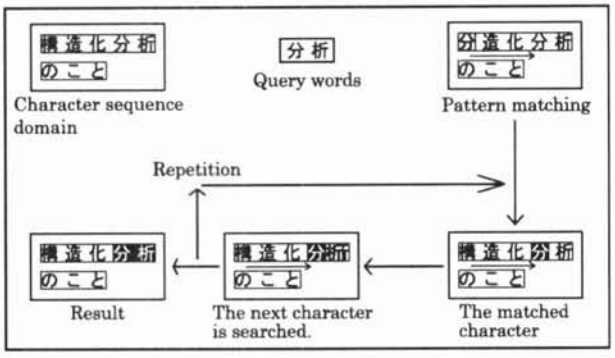

Figure 2. Word retrieval algorithm

3 Experimental Result

In order to verify the validity of this system, the comparison experiment was conducted in the character sequence extraction stage. Moreover, the evaluation experiment was conducted in the word retrieval stage.

The black run-length method which is one of the conventional techniques is used for having considered as comparison contrast in character sequence extraction stage. A formula called the rate of extraction shown below as the evaluation method is used.

・The rate of extraction = the number of extraction lines / the total number of lines

It extracted if it had not run out in the middle of the word even if extraction had run out in the middle of the sequence since it aimed at word retrieval this time.

Moreover, two formulas called the Recall rate and the precision rate which are shown below are used as the evaluation method in word retrieval stage.

・Recall rate = the number of correct words / the number total query words in the image

・Precision rate = the number of correct words / the number of searched words

The query words composed of 1-4 characters are selected from the inside of a document image. Moreover, it would be searched when the query words were contained in the retrieval result domain.

The document images used for the experiment are 100 paper images in the University of Tsukuba Electronic Library (http://www.tulips.tsukuba.ac. jp/), and document images which scanned 100 sheets from the books of 100 paper front pages in an electronic information communication society (http://www.ieice.org/jpn/index.html), and vertical writing. In character sequence extraction stage, it experimented to every 100 sheets [a total of 300] each of these. Moreover, in word retrieval stage, out of these 100 sheets each, it was random five sheets at a time, selected, and experimented by fluctuating a threshold about a total of 15 sheets.

The experiment result in character sequence extraction stage is shown in the following Table 1.

	The rate of extraction		
	(1)	(2)	(3)
University of Tsukuba	93.7%	97.6%	0.0%
Paper	95.0%	94.4%	0.0%
Books	94.0%	4.8%	93.2%

(1) The proposed technique

(2) Black run-length (vertical)

(3) Black run-length (horizontal)

Table 1. A result of extraction experiment

As the result of a character sequence extraction experiment, by using this technique regardless of the image of vertical writing and lateral writing, it was about 95% of the rate of extraction with this algorithm. About the black run length method considered as comparison contrast, when the direction of a line of the target document image differed from algorithm, a character line was not able to be extracted at all.

The experimental result in word retrieval stage is shown in the following Table 2.

Threshold (%)	One character	Two characters	Three characters	Four characters
26	91.98%	87.22%	85.35%	79.46%
28	85.54%	80.76%	79.69%	73.11%
30	83.09%	79.14%	79.69%	66.44%
32	79.80%	78.54%	79.69%	65.49%
34	65.89%	53.77%	40.89%	33.68%
36	60.22%	46.35%	35.27%	30.44%
38	53.54%	39.45%	30.02%	25.52%
40	44.46%	36.05%	28.35%	23.46%

(a)Recall rate

Table 2. A result of retrieval experiment

Since change of the recall rate was seen as a result of a word retrieval experiment when a threshold was 32 - 34%, 32% is considered to be the optimal. The recall rate at this time was about 75%, and the precision rate was about 72%.

4 Conclusion

In this paper, the word retrieval engine for a Japanese bitmap document was able to be built.

About the processing time of this system, they were about 1 - 4 seconds in a series of stages to character sequence extraction - word retrieval. It is a thing when manufacturing and experimenting in this system in the following manufacture environments about this measurement time.

- Use OS Windows2000
- Use Language Visual C++
- Machine Spec Pentium4 1.4GHz(CPU) 256MB(Memory)

This is considered to be the last waiting time which man would regard as unpleasant when searching a character.

Since a thing called a word consists of two or more characters, in case this system performs word retrieval from an experiment result to a document image, it is thought that it is to some extent effective. However, there is the necessity for an improvement from it being thought that the recognition accuracy of word retrieval is still low. As a reason nil why such recognition accuracy is low, although the font size of the target line is specified in the case of character sequence extraction stage, the state which is not desirable is considered to be the cause for the incorrect extraction and the input state of an object image by blur etc.

In character sequence extraction stage, since it is aimed at about 94% of rate of extraction, and the thing which does not contain things, such as a plate, in a image this time although it was high, it is a future subject to conquer such restrictions.

Threshold (%)	One character	Two characters	Three characters	Four characters
26	11.64%	33.36%	60.74%	89.99%
28	13.60%	47.23%	80.56%	91.32%
30	17.19%	65.05%	88.13%	93.75%
32	22.19%	74.18%	94.95%	95.83%
34	24.70%	79.21%	100.00%	100.00%
36	39.19%	94.44%	100.00%	100.00%
38	46.73%	95.00%	100.00%	100.00%
40	52.18%	100.00%	100.00%	100.00%

(b)Precision rate

References

[1] S.M.Harding, W.B.Croft and C.Weir: "Probabilistic Retrieval of OCR Degraded Text Using N-grams," Proc. European Conf. Digital Libraries, pp. 345-349, 1997.

[2] W.B.Croft, S.M.Harding, K.Taghva and J.Borsack: "An Evaluation of Information Retrieval Accuracy with Simulated OCR Output," Proc. Symp. Document Analysis and Information Retrieval, pp.115-126, 1994.

[3] F.R.Chen, L.D.Wilcox and D.S.Bloomberg, "Word Spotting in Scanned Image Using Hidden Markov Models," Proc. ICASSP-93, vol.5, pp.1-4, 1993.

[4] S.Kuo and O.F.Agazzi: "Keyword Spotting in Poorly Printed Document Using Pseudo 2-D Hidden Markov Models," IEEE Trans. PAMI, vol.16, no.8, pp.842-848, 1994.

[5] Y.He, Z.Jieng, B.Liu and H.Zhao, "Content-based Indexing and Retrieval Method of Chinese Document Images," Proc. 5th ICDAR, pp.685-688, 1999.

[6] C.L.Tan, W.Huang, Z.Yu amd Y.Xu: "Imaged Document Text Retrieval without OCR," IEEE Trans. PAMI, vol.24, no.6, pp.838-844, 2002.

[7] Y.Lu and C.L.Tan:"Word Spotting in Chinese Document Images without Layout Analysis", Proc. ICPR-2002, vol.2, pp.57-60, 2002.