Spotting Where to Read on Pages – Retrieval of Relevant Parts from Page Images

Koichi Kise, Masaaki Tsujino, and Keinosuke Matsumoto

Department of Computer and Systems Sciences,
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuencho, Sakai, Osaka 599-8531 Japan
kise@cs.osakafu-u.ac.jp

Abstract. This paper presents a new method of document image retrieval that is capable of spotting parts of page images relevant to a user's query. This enables us to improve the usability of retrieval, since a user can find where to read on retrieved pages. The effectiveness of retrieval can also be improved because the method is little influenced by irrelevant parts on pages. The method is based on the assumption that parts of page images which densely contain keywords in a query are relevant to it. The characteristics of the proposed method are as follows: (1) Two-dimensional density distributions of keywords are calculated for ranking parts of page images, (2) The method relies only on the distribution of characters so as not to be affected by the errors of layout analysis. Based on the experimental results of retrieving Japanese newspaper articles, we have shown that the proposed method is superior to a method without the function of dealing with parts, and sometimes equivalent to a method of electronic document retrieval that works on error-free text.

1 Introduction

Document image retrieval is a task to retrieve document images relevant to user's information needs. This technology allows us to replace a huge amount of paper documents with document images, and thus enables us to solve the problem of space occupied by paper documents.

Although a number of researches have been made on document image retrieval [1], there is still room for improvement. We focus here on the problems of *indexing*, *retrieval* and *presentation*.

Indexing. Users' information needs are typically represented as keywords, while the database stores document images. This gap between symbols (character codes) and signals (pixels) poses the problem of indexing. Since manual indexing is prohibitive, automatic indexing should be employed. Researchers have proposed a number of methods ranging from the indexing based on text produced by OCR [2] to the indexing based on the image features [3,4,5]. The problem here is how to obtain the robustness against low-quality images. For instance, in case that the indexing based on OCR results is employed,

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 388–399, 2002.

it is necessary to cope with OCR errors. The OCR errors are not limited to the misrecognition of individual characters, but include the errors in layout analysis and identification of reading order.

Retrieval. The effectiveness of document (or page) ranking is an important problem of retrieval, though most of the existing methods deal mainly with keyword spotting on page images. In order to obtain the ranking, we should define and utilize a measure of similarity between a user's query (a set of keywords) and a page image, according to the spotted keywords.

Presentation. It seems that the problem of presentation is often overlooked. This problem is caused by the disparity between the size of page images and the size of images that can be displayed. For example, newspaper pages scanned with the resolution of, say, 200 dpi, are too large for ordinary displays. Images of A4 pages could cause the same problem if we use PDA's. Thus it is important to locate where to read on pages in addition to select pages which contain information relevant to a query.

In the field of electronic document retrieval, researchers have faced the similar problems of document ranking and presentation, and proposed the scheme called "passage retrieval" [6,7] to solve the problems. The task of passage retrieval is to retrieve not whole documents but their parts, or *passages* relevant to a query. This enables us to solve the problem of presentation in a natural way. It has also been shown that, as compared to the similarity based on whole documents, the similarity based on passages improves the quality of document ranking [9]. This suggests that the problems of document image retrieval could also be solved by applying passage retrieval.

In this paper, we experimentally validate the above suggestion. We propose a method of passage retrieval for document images by extending our method [8,9] for electronic documents. In the context of document image retrieval, passages correspond to parts of page images. The characteristic features of the proposed method are as follows:

- For indexing, we do not employ the results of neither layout analysis nor identification of reading order so as to obtain the robustness against errors. Document images are indexed based only on character positions and codes obtained by segmentation and recognition of characters.
- We assume that parts of page images are relevant to a query if these parts *densely* contain keywords in a query. Two-dimensional distributions called "density distributions of keywords" are calculated for finding such parts.

The organization of this paper is as follows. In Sect. 2, we give an overview of the related work in the fields of both electronic documents and document images. Section 3 describes the proposed method of retrieving parts of document images. In Sect. 4, we experimentally evaluate the proposed method in comparison with a standard method of electronic document retrieval.

2 Related Work

2.1 Retrieval of Electronic Documents

Since the mid 90's, researchers in the field of information retrieval have proposed a retrieval scheme called "passage retrieval" [6,7]. Passage retrieval is advantageous to conventional document retrieval in the following points:

1. It provides us the direct access to passages, which relieves us from the burden of finding relevant parts in the retrieved documents.
2. It improves the ranking of documents as follows. In general, long documents contain multiple topics. Even if a topic in such documents is relevant to a query, the rest may be irrelevant. This results in disturbing document ranking by conventional document retrieval, since there is no way to distinguish relevant topics from irrelevant topics. On the other hand, passage retrieval enables us to avoid the influence of irrelevant topics in documents.

The authors have also proposed a method of passage retrieval called "density distribution". The density distribution was first introduced to locate the explanation of a word in long documents [10] and applied to passage retrieval by the authors [8,9]. In this paper, we extend this method to be applicable to the retrieval of page images.

2.2 Retrieval of Document Images

In the field of document image retrieval, a central issue has been how to locate words on page images. This would be because, after locating words, the task of document image retrieval is considered to be equivalent to that of electronic document retrieval. For the retrieval of page images written in ideograms such as Kanji characters, we can also consider the possibility of locating characters. Words and characters can be located in several ways including the application of OCR [4], utilization of a special set of symbols such as "character shape coding" [3], and the matching with real-valued feature vectors [4,5].

However, if we build systems of document image retrieval by replacing the function of locating words in electronic document retrieval, we will suffer the similar difficulties about the selection of relevant parts as well as the influence of irrelevant parts.

We consider that the layout of pages provides us fruitful hints for solving these problems in the document image domain. This is because words and characters in the same topic are laid out closer with one another on a page. The method proposed in this paper embodies this idea without suffering from the errors of layout analysis.

3 Retrieval of Image Parts

3.1 Overview

The proposed method employs three types of processing: indexing, retrieval and presentation. The process of indexing is applied to document images in advance

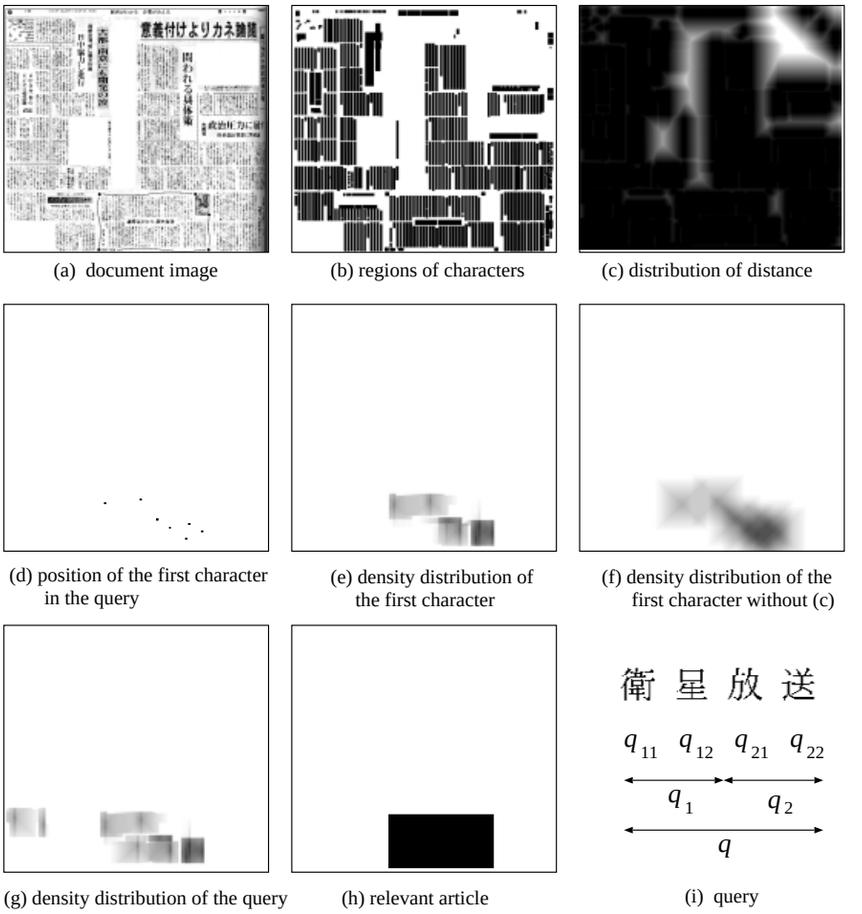


Fig. 1. Examples of processing results

of receiving a query. Every time a query is received, the process of retrieval starts to find relevant parts of images, which are then displayed by the process of presentation. The current implementation is just for Japanese printed documents, though we consider that central ideas of the method are independent of the type of languages.

In what follows, each step of processing is explained using examples shown in Fig. 1. In Fig. 1(a), a page of a newspaper that contains an article about "satellite broadcasting" is shown. Figure 1(i) illustrates four Kanji characters (q_{11}, q_{12}, q_{21}, q_{22}) which mean "satellite broadcasting". Taken as input the query in Fig. 1(i), the method produces the density distribution of the query as shown in Fig. 1(g) based on the intermediate results of Figs. 1 (b),(c),(d) and (e). Figure 1(h) illustrates the correct region of the article about "satellite broadcasting"; the output (Fig. 1(g)) agrees well with the correct region.

3.2 Indexing

For documents written in western languages, it is natural to take words as units of indexing, because it is relatively easy to extract them from page images. On the other hand, for documents written in agglutinative languages such as Japanese, it is difficult to extract words because there is no space between them. Morphological analysis is required to identify words from a sequence of characters. Instead of applying morphological analysis to recognized characters[1], we simply utilize recognized characters as units of indexing.

Segmentation and Recognition of Characters. As the first step of indexing, we apply segmentation and recognition of characters to page images. Figure 1(b) illustrates the regions of characters segmented from the page image in Fig. 1(a). In Fig. 1(b), regions of characters are shown as black rectangles.

Extraction of a Layout Feature. In general, we can browse pages efficiently with the help of layout. This would be partly because physical components such as text-lines and characters laid out closer with one another are of the same topic, and wide space indicates the boundaries of topics.

In order to employ such fruitful information in a way robust to noise on images, a low level feature about the layout is extracted from each page image p. The feature we utilize here is the minimum distance $K^{(p)}(x, y)$ from a pixel (x, y) to a character closest to the pixel. This can be easily obtained by applying the distance transform to a page image p. An example of $K^{(p)}(x, y)$ is shown in Fig. 1(c), where lighter pixels indicate longer distance to the closest characters.

3.3 Retrieval

Query Processing. The first step of retrieval is the processing of a query. Since a user may represent a query as sentences or phrases, we first apply morphological analysis to extract as keywords nouns from a query. Then each keyword is further decomposed into characters.

For example, a query q (*satellite broadcasting*) shown in Fig. 1(i) is decomposed into two keywords q_1 (*satellite*) and q_2 (*broadcasting*). The first keyword q_1 consists of two characters q_{11} and q_{12} in Fig. 1(i), and the second keyword q_2 includes two characters q_{21} and q_{22}.

In the following, keywords extracted from a query q are represented as $\{q_1, ..., q_n\}$, and characters contained in a keyword q_i are represented as $\{q_{i1}, ..., q_{im}\}$.

[1] This is not a trivial task because morphological analyzers are not designed to deal with errors of character recognition.

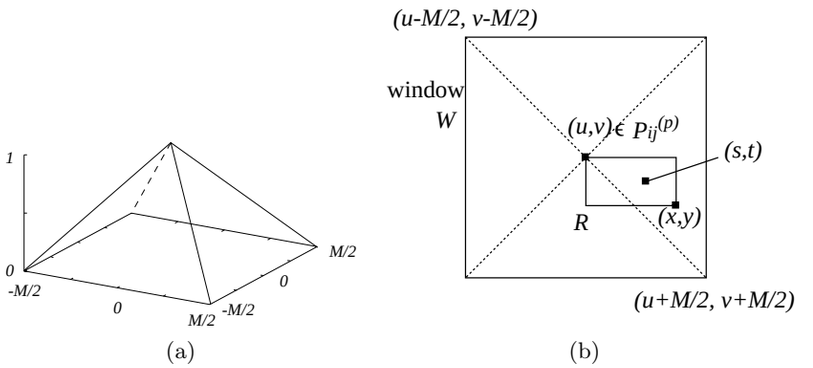


Fig. 2. Window function and points (u, v), (x, y) and (s, t)

Calculation of Density Distributions. The role of this step is to calculate the density distribution of a query for each page in order to obtain information for spotting relevant parts. The processing consists of the following four steps.

(1) Locating Characters on Pages

As shown in Fig. 1(b), character regions are represented as rectangles. We indicate positions of characters using the centers of rectangles. In this step, we first find all positions of each character q_{ij} on pages. An example is shown in Fig. 1(d). In the following, the positions of a character q_{ij} in a page image p are represented by a set of points $P_{ij}^{(p)} = \{(u, v)\}$ where (u, v) indicates the center of a rectangle.

(2) Calculation of the Density Distribution of a Character

The density distribution $D_{ij}^{(p)}(x, y)$ of a character q_{ij} is obtained by smoothing the distribution of points $P_{ij}^{(p)}$ with the distance $K^{(p)}(x, y)$. Figure 1(e) illustrates the density distribution obtained from Fig. 1(d) with Fig. 1(c). In Fig. 1(e), darker regions contain the character more densely. The distribution is given by

$$D_{ij}^{(p)}(x, y) = \sum_{(u,v)\in P_{ij}^{(p)}} W(x - u, y - v)\alpha^{(p)}(x, y, u, v) \ , \tag{1}$$

where W is a window function for smoothing, and $\alpha^{(p)}$ is a weight obtained from $K^{(p)}(x, y)$. As a window function, we utilize a square pyramid shown in Fig. 2(a) where M is the window width.

The weight $\alpha^{(p)}$ is to control the influence of a point (u, v) on the distribution using the distance $K^{(p)}(s, t)$:

$$\alpha^{(p)}(x, y, u, v) = \begin{cases} T - \max\limits_{(s,t)\in R} K^{(p)}(s, t) & \text{if } \max\limits_{(s,t)\in R} K^{(p)}(s, t) < T \ , \\ 0 & \text{otherwise} \ , \end{cases} \tag{2}$$

where T is a threshold and (s, t) is the point within the rectangle R defined by two points (u, v) and (x, y) as shown in Fig. 2(b). If the rectangle R includes

a point whose distance $K^{(p)}(s,t)$ is larger, the value of the weight α becomes smaller. When the distance exceeds the threshold T, there is no influence from the point (u,v) to the point (x,y).

If we do not employ the weight $\alpha^{(p)}$, the distribution shown in Fig. 1(f) is obtained from Fig. 1(d). The weight $\alpha^{(p)}$ changes this distribution into the distribution shown in Fig. 1(e).

(3) Calculation of the Density Distribution of a Keyword

Next, the density distribution $D_i^{(p)}(x,y)$ of a keyword q_i is calculated based on the distribution of $D_{ij}^{(p)}(x,y)$ of each character q_{ij} in q_i:

$$D_i^{(p)}(x,y) = \min_j \beta_{ij} D_{ij}^{(p)}(x,y) \ . \tag{3}$$

The operation "min" is applied to the distributions of characters, because we are interested in the *co-occurrence* of characters to identify a keyword.

In (3), β_{ij} indicates the weight of $D_{ij}^{(p)}$ which represents the importance of a character q_{ij} for the identification of a keyword q_i. If a character q_{ij} is ubiquitous on pages, its distribution conveys little information for the identification, and thus a small weight is used[2]. The above notion can be expressed as:

$$\beta_{ij} = \sum_{p,x,y} (\max_{p,x,y} D_{ij}^{(p)}(x,y) - D_{ij}^{(p)}(x,y)) \ . \tag{4}$$

(4) Calculation of the Density Distribution of a Query

As the last step, the density distribution of a whole query q is calculated based on the distributions of keywords q_i as follows:

$$D^{(p)}(x,y) = \max_i D_i^{(p)}(x,y) \ . \tag{5}$$

This time, the operation "max" is applied since the parts which contain one of the keywords would be relevant to a query.

3.4 Presentation

Each page image p is now associated with its density distribution $D^{(p)}(x,y)$ of a query. In our method, the maximum of the density distribution in a page image:

$$\max_{x,y} D^{(p)}(x,y) \tag{6}$$

is employed as a score of a page image. Page images in the database are ranked according to their score (in descending order of the score). Then the top-ranked page image is presented to a user by displaying its part whose density is the maximum.

[2] The aim of the weight β_{ij} is similar to the *inverse document frequency (IDF)* that is often employed for term weighting in IR systems.

Table 1. Queries used in the experiments

query id	keywords(translation)	no. of relevant articles
1	任天堂 (Nintendo), セガ (Sega)	4
2	農薬 (agricultural chemical)	3
3	液晶 (liquid crystal)	3
4	減税 (tax reduction)	4
5	衛星 (satellite), 放送 (broadcasting)	3
6	賃貸 (rent), 住宅 (house)	4
7	核兵器 (nuclear weapon)	4

Table 2. Statistics on page images and queries for experiments

no. of pages	25
no. of articles	249
size of page images	$8,000 \times 6,000$ pixels
scanning resolution	800 dpi
size of characters in body text regions	50×50 pixels
no. of queries	7
ave. no. of relevant articles for a query	3.57 articles
ave. no. of keywords in a query	1.48 keywords
ave. no. of characters in a keyword	2.2 characters

4 Experimental Results

4.1 Data for Experiments

We prepared the data of page images and queries based on a test collection available for evaluation of Japanese information retrieval (IR) systems. The collection we utilized is called BMIR-J2 [12], which consists of 50 queries for 5,080 articles of Mainichi Shinbun newspaper issued in 1994 and the groundtruth (relevance judgements) for the queries.

Based on BMIR-J2, we prepared the document image database with the groundtruth. The queries in BMIR-J2 are classified into several types. We utilized the queries which belong to the basic type as listed in Table 1. Next, we randomly selected, for each query, three to four articles whose subjects are relevant to it. Then, pages that contain those articles were obtained from the microcopy of the newspaper. In the scanned pages, the parts which do not correspond to 5,080 articles of BMIR-J2 were manually erased. The statistics on page images and queries are shown in Table 2.

The results of character segmentation and recognition were obtained using a commercial OCR. The recognition rate for the characters in the queries was 74.1% and the false alarm rate was 2.3%. We consider that such a low recognition rate was due to the quality of the images which were obtained not from the original pages but from the microcopy pages.

In the processing of queries, we applied a Japanese morphological analyzer JUMAN [13] to obtain keywords.

4.2 Methods for Comparison

Since the electronic version of articles are also available, we can apply a standard IR method for the same set of queries and articles. In the experiments, we employed as a method of comparison the vector space model(VSM) with the tf–idf term weighting [11]. The VSM is applied to the recognized text as well as to the clean (error-free) text. In the following, the VSM applied to the clean text is referred to as "VSM" and the VSM to the recognized text is referred to as "VSM (OCR)".

In addition to them, we employed a modified version of the proposed method in order to evaluate the effectiveness of taking image parts into account. The modification is to exclude the function for dealing with image parts as follows. For each character q_{ij} in a keyword q_i, the weight tf_{ij}, which corresponds to D_{ij} in (1), is computed as

$$\mathrm{tf}_{ij} = (\text{the number of occurrence of } q_{ij} \text{ in a page}) \quad . \tag{7}$$

The weight w_i for a keyword q_i is then given by

$$w_i = \min_j (\mathrm{idf}_{ij} \cdot \mathrm{tf}_{ij}) \quad , \tag{8}$$

which corresponds to (3). In (8), idf_{ij} is the inverse document frequency given by

$$\mathrm{idf}_{ij} = \log \frac{N}{N_{ij}} \quad , \tag{9}$$

where N and N_{ij} represent the number of all pages in the database and the number of pages that contain the character q_{ij}, respectively. Finally, the score of a page is obtained by

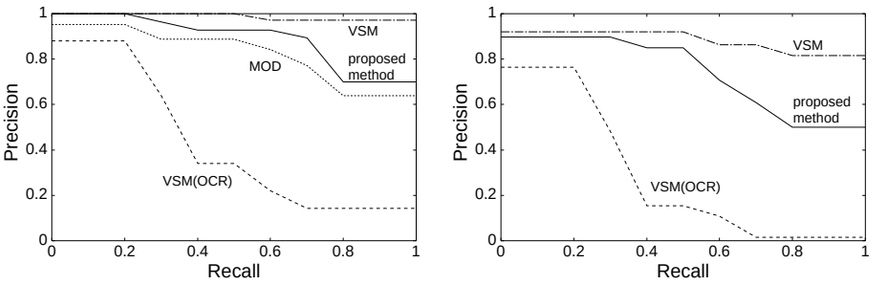
$$\max_i w_i \quad , \tag{10}$$

which corresponds to (5). In the following, this method is called "MOD".

4.3 Criteria for Evaluation

A common way to evaluate the performance of retrieval methods is to compute recall and precision [11]. Let X, Y be a set of retrieved documents and a set of relevant documents, respectively. The recall R and the precision P is defined as $R = |X \cap Y|/|Y|$, $P = |X \cap Y|/|X|$. When we have multiple queries for evaluation, the precision for each query is calculated at some fixed points of recall (recall levels) with the help of interpolation, and then averaged over all queries at each recall level. Since this produces some points of recall and precision for each method, they are typically shown in recall-precision graphs.

In addition to the recall-precision graphs, it is sometimes convenient to have a single value that summarizes the performance. For this purpose, we utilized *the average precision (non-interpolated) over all relevant documents* [11].



(a) retrieval of pages (b) retrieval of articles

Fig. 3. Experimental results

4.4 Tasks

In this paper, we have two retrieval tasks, i.e., retrieval of pages and that of articles. For the task of retrieval of pages, methods were required to rank *pages* for evaluation. In the proposed method, for example, the ranking was obtained according to the score defined by (6). For the task of retrieval of articles, on the other hand, *articles* were ranked according to the score. Although there is no obstacle to compute the score of articles by the VSM, it is not possible for the proposed method, because there is no way to obtain the exact regions of articles by itself. In order to compute as well the score of articles for the proposed method, the regions of articles were prepared beforehand and the score of each article was determined as the maximum of the density distribution $D^{(p)}(x, y)$ within its region.

4.5 Values of Parameters

The proposed method employs two parameters: the window width M in Fig. 2 and the threshold T in (2). The values of these parameters were experimentally determined as follows. Since the number of queries was small, we applied leave-one-out, which means that values of the parameters for i-th query were determined using the data for all queries except for the i-th query. We examined all the combination of the values for $M = 100 \sim 3,000$ with the step of 50 and $T = 10 \sim 100$ with the step of 10 to obtain the best values in terms of the average precision for retrieval of pages.

4.6 Results and Discussion

The results of experiments are shown in Fig. 3. Because MOD is only for page ranking, its result is shown only for the task of "retrieval of pages".

As shown in Fig. 3, the proposed method was superior to VSM(OCR) for both of the tasks (retrieval of pages and articles), though these two methods

equally took as input error-prone recognition results (74% recognition rate). The low performance of VSM(OCR) was reasonable because it employed no mechanism for compensation of recognition errors such as the approximate string matching. Although the proposed method was without any compensation as well, it outperformed VSM(OCR), because (1) in order to obtain appropriate ranking by VSM(OCR), it is required that all characters in a keyword are correctly recognized, but (2) the proposed method does not require it. In other words, some characters correctly recognized within the window helped ranking.

The proposed method was inferior to the VSM (applied to the clean text) for both of the tasks. However, we were surprised that the difference was not so big except for the performance at high recall levels (0.7 – 1.0). This difference at high recall levels was caused mainly by the errors of character recognition. Some relevant documents contain only a few characters of a keyword, so that the performance was influenced by the errors. Thus in order to improve the performance at high recall levels, it is required to employ a method to cope with recognition errors.

Finally, let us compare the proposed method with MOD on the task of page retrieval. The graph in Fig. 3(a) shows a small but clear difference at all recall levels. This indicates that the the proposed method which takes into account the locality of occurrence of characters and keywords enables us to improve precision of retrieval.

5 Conclusion

We have presented a method of document image retrieval which enables us to spot where to read on pages. The characteristics of the proposed method are as follows:

1. The method realizes the notion that relevance of parts of images can be measured by the density distributions of keywords,
2. The method relies only on the distribution of characters so as to obtain the robustness against errors of layout analysis.

From the experiments on the retrieval of both pages and articles, it has been shown that the proposed method outperforms the methods (VSM and MOD) that also work on recognized characters. It has also been shown that at low and middle recall levels the proposed method is almost equivalent to the VSM that works on error-free text.

Due to the difficulty of preparing the groundtruth for the articles on pages, the number of pages and articles were too small to draw a definite conclusion. In addition, the documents employed in the experiments were limited only to those written in Japanese, though we consider that the fundamental notions of the method are independent of languages. Thus the future work is to scale up the experiments as well as to apply the method to documents written in other languages.

Acknowledgment. This research was supported in part by Grant-in-Aid for Scientific Research (C) and (B) from Japan Society for the Promotion of Science (No.14580453, No.14380182). We used BMIR-J2 based on the Mainichi Shinbun CD-ROM'94 data collection, as well as a Japanese morphological analyzer JUMAN. We are grateful to those who permitted us to use their materials and programs.

References

1. Doermann, D.: The Indexing and Retrieval of Document Images: A Survey, *Computer Vision and Image Processing*, Vol. 70, No. 3, pp.287–298, 1998.
2. Ohta, M., Takasu, A., Adachi, J.: Retrieval Methods for English-Text with Missrecognized OCR Characters, *Proc. of the 4th ICDAR*, pp.957–961, 1997.
3. Smeaton, A. F., Spitz, A. L.: Using Character Shape Coding for Information Retrieval, *Proc. of 4th ICDAR*, pp.974–978, 1997.
4. Ohta, Y., Mori, R., Sakai, T.: Retrieval of Chinese Character Sequence Using Pictorial Features — The Case of Names on Visiting Cards —, *Trans. IECE, Japan*, Vol. J64-D, No. 11, pp.997–1004, 1981 (in Japanese).
5. Nakanishi, T, Omachi, S., Aso, H.: High Precision Keyword Search System Adapted to Low Quality Document Images, *Tech. Report of IEICE*, PRMU98-232, 1999 (in Japanese).
6. Salton, G., Singhal, A., Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, in *Proc. Hypertext '96*, pp.53-65, 1996.
7. Callan, J. P.: Passage-Level Evidence in Document Retrieval, in *Proc. SIGIR '94*, pp.302-310,1994.
8. Kise, K., Mizuno, H., Yamaguchi, M., Matsumoto, K.: On the Use of Density Distribution of Keywords for Automated Generation of Hypertext Links from Arbitrary Parts of Documents, in *Proc. of the 5th ICDAR*, pp.301–304, 1999.
9. Kise, K., Junker, M., Dengel, A., Matsumoto, K.: Experimental Evaluation of Passage-Based Document Retrieval, in *Proc. of the 6th ICDAR*, pp.592–596, 2001.
10. Kurohashi, S., Shiraki, N., Nagao, M.: A Method for Detecting Important Descriptions of a Word Based on Its Density Distribution in Text, *Trans. Information Processing Society of Japan*, Vol.38, No.4, pp.845–853, 1997 (In Japanese).
11. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison-Wesley Pub. Co., 1999.
12. Sakai, T., et al.: BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems, *SIGIR Forum*, Vol.33, No.1, pp.13–17, 1999.
13. ⟨ URL:http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html ⟩.