MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

An Extraction Method of Search Indexes for Graph Image Retrieval

Masashi Koga,Tatsuya Murakami,
and Yoshihiro Shima
Central Research Laboratory
Hitachi, Ltd.

1-280,Higashi-Koigakubo,Kokubunji
Tokyo 185 JAPAN

Hiromichi Fujisawa

Software Development Center
Hitachi, Ltd.

549-6,Shinano-machi,Totsuka-ku
Yokohama-shi, Kanagawa 244, Japan

ABSTRACT

The increased capacity of document image retrieval systems has necessitated an efficient method for indexing parts of document images. This is particularly important for graphs, because graphs provide the relationships between two items a user is searching for. This paper proposes a method that recognizes and indexes different kinds of graphs. Coordinates extracted from a graph are used to assign the graph to a category and to extract the labels attached to coordinates. The category is registered as an index of the graph. The extracted label images are recognized, converted into ASCII data, and also stored as indexes. We propose an automaton type recognition method that is based on a knowledge of graph formats. It, by detecting the coordinates and labels from among candidates, recognizes the graph type. The transitions of the automaton are driven by local features of the images. We also propose a system that retrieves parts of document images, and the automatic indexing method is also applied to this system.

INTRODUCTION

Recent advances in computers and storage devices have enormously expanded the capacity of electronic document filing systems [1,2]. These systems are expected both to provide an efficient environment for retrieval and to reduce the amount of space needed for storage. Efficient retrieval will require that adequate indexes are given to each document, and that it is possible to retrieve parts of documents (charts, tables, photographs, etc.). We must therefore develop a method for dividing document images into parts and giving an index to each part.

We can, of course, input the indexes and the coordinates of the parts by using a mouse or keyboard, but for a large system this would be impractically slow. An automated dividing and indexing method must therefore be developed. There is already a method for analyzing full-color document images and segmenting them into parts, such as photograph, charts, and text[3]. And there are methods that analyze the logical structure of document and extract strings adequate for search indexes [4-7]. These methods extract the indexes of diagrams from the document space outside the diagrams and they ignore the strings inside diagrams. Sometimes, however, the most effective indexing strings are within the diagram. Even though the diagrams (such as graphs) are often the most important parts of documents, an effective method for extracting indexing strings from diagrams in document images has not yet been found.

This paper proposes a method that recognizes different kinds of graphs by extracting two coordinates from a graph (as well as the labels attached to the coordinates) by using the format rules for graphs. Earlier, we analyzed various kinds of graphs and found that they could be classified into categories. The extracted classifications are stored as one index of the graph. The extracted label images are recognized, converted into ASCII data, and also stored as an index.

SUBJECTS OF GRAPH IMAGE INDEXING

Many types of graphs are used in technical documents (for example, bar graph,

line graph, etc.). Their type of coordinates also can be classified into many categories (Fig. 1), but only a few categories are frequently appeared. We searched 6 technical journals (a total of 600 pages) for graphs and examined. 95% of the graphs found in the journals are line graphs and the rest are bar graphs. Of the line graphs, 51% are classified as box type, 19% are L type, 7% are cross type, and most of the others are variations of the three types. So, a method which can index graph of these three types of their variation is enough to cover the most of graphs used in technical journals. Some types of graphs are difficult to recognize because of more pictorial elements included. Still, it is true that certain kinds of graphs can be analyzed and indexed automatically.

Fig. 1 : Type of Coordinates of Graph

PROCESS OF ANALYSIS OF GRAPH IMAGE

The concept overview of the advanced document file system which can automatically index the graph image is shown in Fig. 2. The document image may be color image to store the document with photograph or colored diagram, but in the case of graph images, only binary image is used. Page images are stored in page image database as the conventional filing system. At the same time, graph area in the documents are automatically extracted and stored in partial image database. The graph image are analyzed by the method described in this paper, then coordinates type are identified and labels of the coordinates are extracted. The types and the labels are stored in index database and used to retrieve the graph image. To operate the retrieval engine, visual user interface is provided which enables the operator to input the query and browse the retrieved image in visual environment.

The flow of index extraction of the newly developed method are shown in Fig. 3. First, candidate elements of a graph image are

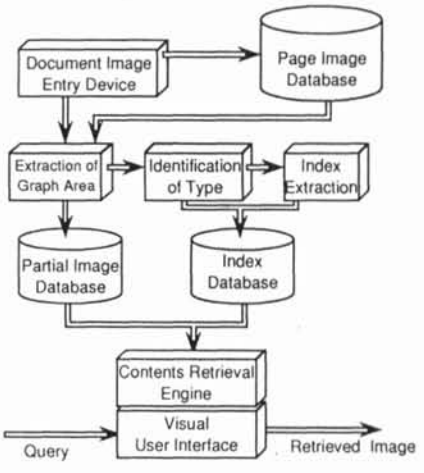


Fig. 2 : Advanced Document Image Retrieval System

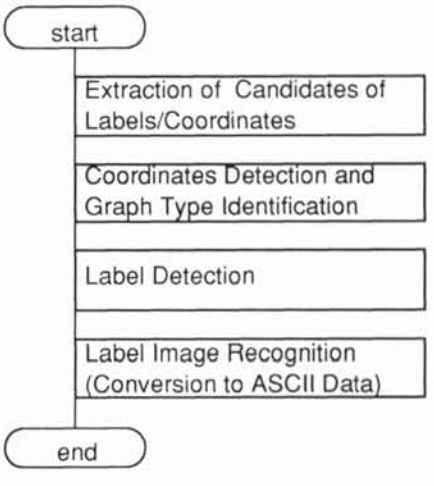


Fig. 3 : Flow of Indexes Extraction

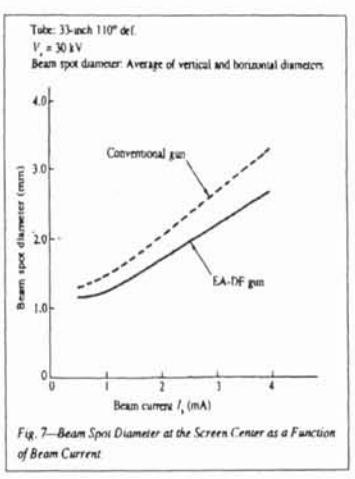
extracted by analyzing the projection. Candidate elements are parts of graph image which are dense with the black dots and where coordinates of the graph or labels of the coordinates may be placed. Secondly, the candidates are analyzed referring the knowledge of form of graphs, then the coordinates are detected from the candidates and type of the coordinates is identified. Next to this step, label images of the coordinates are detected using the information about the position and the type of the coordinates. Finally, the label images are recognized using the technique of OCR. The ASCII data obtained by the recognition are registered as indexes of the graph as well as the coordinates type.

Fig. 4 shows an example of the results of candidates elements extraction. Original image of graph is shown in (a). The curved line in the right half of (b) shows the projection in horizontal direction. The horizontal line drawn in the left half of (b) shows candidates elements of horizontal coordinates and labels of the coordinates. The candidates elements are extracted by finding the peak of the projection. Then the candidates elements are classified into two categories. The width of each candidates elements are measured by examining the projection. If the width is smaller than a threshold, the candidates element is classified as a candidate of the coordinates. If not, the candidate is classified as a candidate of the labels. The projection in vertical direction near each candidate elements of coordinates are calculated and the ends of lines placed around the candidate elements are detected. The candidate elements of vertical coordinates are generated just in the same way.

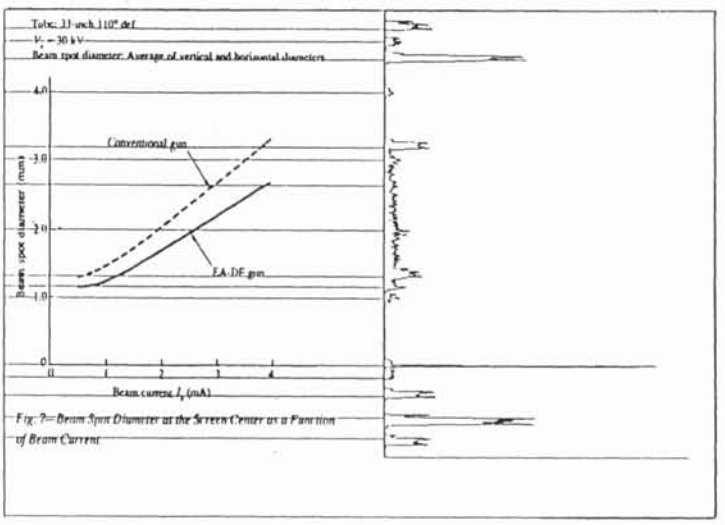
An algorithm which analyzes the candidates elements based on the knowledge of form of graph was developed. An automaton model, shown in Fig. 5, is applied to the algorithm. By the algorithm information of the candidate elements of coordinates are examined one by one and transitions of the automaton take place according to the information. The horizontal candidates are examined first, then vertical candidates are examined referring the horizontal candidates already examined. For example, if a candidate elements of coordinates horizontal is examined when the state is S0, the transition of T1 takes place. If a vertical candidate is examined next, transition t11, t12, t13 or t14 take place according to the relativity of the candidates already examined and newly input. Inadequate candidates are rejected, and a series of the transitions finishes when all candidates are examined. The detected coordinates are the candidate elements which are not rejected, and the identified graph type is the state where the transition finished. Label images are extracted afterwards, using the position and the type of coordinates and the knowledge of layout of labels of graph. Fig. 6 and Fig. 7 show examples of experimental results. Coordinates and labels are extracted from other imagery elements.

PERFORMANCE ANALYSIS

We implemented this method on an workstation. The program is written in C language. 28 samples, collected from a technical journal (Hitachi Review Vol. 37 - No. 5), are tested. From 26 of them, the graph type and coordinates positions are



(a) Original image



(b) Projection and Candidates

Fig. 4: An Example of the Results of Candidates Elements Extraction

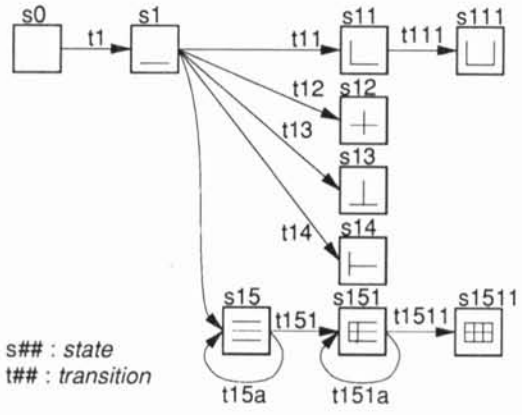
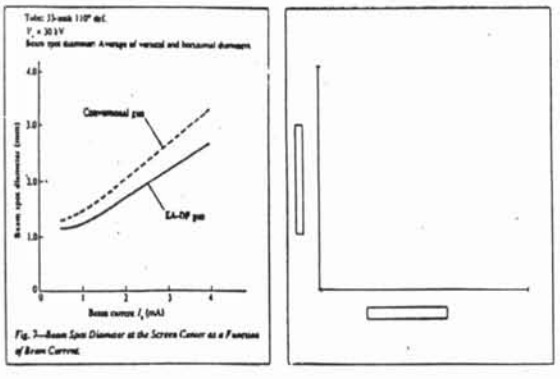
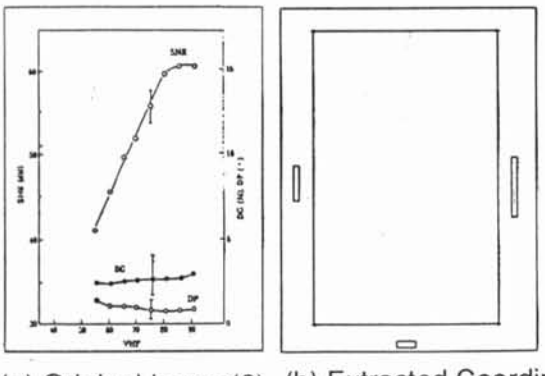


Fig. 5: Automaton model used to recognize the coordinates of a graph



(a) Original Image(1) (b) Extracted Coordinates and Labels(1)

Fig. 6 Example of Result of Experiments(1)



(a) Original Image(2) (b) Extracted Coordinates and Labels(2)

Fig. 7 Example of Result of Experiments(2)

extracted successfully. Coordinates are not extracted correctly from 2 samples because of break of coordinates line. Label image positions are extracted successfully from 25 samples, but not be extracted correctly from 3 samples in which one of the labels touches to coordinates or other strings.

CONCLUSION AND FUTURE WORKS

We developed a method to extract the coordinates type and labels of the coordinates of graph. We applied an automaton model to analyze the coordinates structure. Various types of graph images are correctly analyzed by the method.

The following are the future works.

(1) To develop a method to index more complex diagrams such as flow chart etc.

(2) To develop a user interface environment to facilitate the retrieval of partial image

ACKNOWLEDGEMENTS

This work was performed under the management of INTAP as a part of the Large Scale Project of AIST/MITI "Interoperable Database Systems", by NEDO (New Energy and Industrial Technology Development Organization).

REFERENCES

[1] S. Itoh and N. Takahashi, "HITFILE 650 Optical Disk Filing System", Hitachi Review, Vol. 36, No. 4, 1987, pp. 213 - 220

[2] H. Fujisawa, "Artificial Intelligence as applied to Optical Image Filing", Japanese J. Applied Physics, Vol. 26, Supplement 26-4, 1987, pp. 205-210

[3] Y. Shima, T. Murakami, et al., "A Segmentation Method of Color Document Images for Multimedia Content Retrieval System", RIAO '88,User-Oriented Content-Based Text and Image Handling Conference (AFIPS), Cambridge, Mar. 1988

[4] J. Higashino, H. Fujisawa, et al.,"Document Image Understanding using a Form Definition Language", Proc. Ann. Meeting of IECE, S10-2, Mar. 1985 (in Japanese).

[5] J. Higashino, H. Fujisawa, Y. Nakano and M. Ejiri, "A Knowledge-based Segmentation Method for Document Understanding", 8th Int. Conf. Pattern Recognition, Paris, Oct., 1986, pp. 745-748

[6] H. Fujisawa, et. al. "Document Analysis and Decomposition Method for Multimedia Contents Retrieval" Proc. the Second Int. Symposium on Inter operable Information Systems, Nov. 1988

[7] H. Yashiro, H. Fujisawa, et al., "A New Method of Document Structure Extraction using Generic Layout Knowledge.", MIV-89, Tokyo, Apl. 1989, pp. 282-287

[8] K. Wong, R, Casey, and F. Wahl, "Document Analysis System", IBM J. Research and Development, Vol. 26, No. 6, 1982, pp. 647-656

[9] K. Inagaki, T. Kato, T. Hiroshima, and T. Sakai, "MACSYM : A Hierarchical Image Processing System for Event-Driven Pattern Understanding of Documents", Pattern Recognition, Vol. 17, No. 1, 1984, pp.85-108

[10] K. Kubota, O. Iwaki, et al., "Document Understanding System", Proc. 7th Int. Conf. Pattern Recognition, 1984, pp.612-614