IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 4 (Sep.-Oct. 2012), PP 31-35
www.iosrjournals.org

Layout Based Information Retrieval from Document Images

[1]D.Shobana, M.Sc(I.T).,[2]M.Phil(C.S).,(Ph.D)
(Department Of Computer Science, Research Scholar, Dravidian University)

Abstract : This research is intended to develop a layout based retrieval system for document image databases consisting of three phases: 1. At first, intelligent layout analysis algorithm has been designed to extract the layouts the document images physically with their edges and rectangles. 2. Every physically identified layout has been converted into a tree intermediary representation for indexing and storage in layout databases. Later when a query image has been supplied, it retrieves the similar layout images from the layout databases. 3. Finally, a logical layout analysis scheme has been proposed to identify the meaning of the layouts involved in the document images. In intelligent layout analysis system, White Space Analysis technique has been proposed to grab all the white spaces over the image in a single scan over the image with minimum pixel visits, and the white spaces are merged together without the assumptions of heuristics and threshold to segment the layouts. Moreover, two statistical properties have also been proposed in this thesis, to separate the text blocks and images from the identified layouts. In Layout based retrieval system, Tree Representation for physical layouts in a document image and indexing has been proposed in this thesis to enable the retrieval of document images based on layout similarities. This allows a user to find pages with a layout of similar page to the query image supplied by the user based on the tree structure. In logical layout analysis, this thesis attempts to analyze the layouts of the document images logically as title blocks, sub titles, text, images, content lines etc.. to produce more meaningful information from the document images.

Keywords: White Space Analysis (WSA) technique, Information Retrieval(IR), optical character recognition (OCR), Document Image Processing(DIP), Java Advanced Imaging (JAI).

I. Introduction:

The analysis of scanned documents is an important activity in the construction of digital paperless offices, digital libraries or other digital versions of originally printed documents. Both the digital or printed form of documents has its advantages. For instance, online digital libraries can provide improved distribution of information and more flexible access via search algorithms than traditional print libraries. On the other side printed document are still more comfortable for reading and modifying. The ultimate solution would be to deal with paper documents as we deal with other forms of computer media. However, adding existing print material into electronic libraries is a costly, slow process unless good automated procedures can be developed. The objective of document analysis systems is to recognize text, graphics and pictures in usually scanned images and to extract the intended information as a human would.

II. A Survey Based on the Research:

Accessing collections of document text is a problem that has been addressed by the information retrieval (IR) community for many years. For much of that time, however, it has been assumed that the systems would deal exclusively with clean and accurate data. Recently, techniques have been developing to deal with noisy data. The general consensus of the community had been that if sufficient computational resources are given, the text in document images could be recognized and converted so that standard retrieval techniques could be applied. Now, techniques are been emerging to retrieve information from document images without following an entire conversion.

Document Image Processing system can be applied for the purpose of document image retrieval. For example, a DIP system is used to automatically convert a document image to machine-readable text codes first and traditional information retrieval strategies are then applied. However, the performance of a DIP system relies heavily on the quality of the scanned images. It deteriorates severely if the images are of poor quality or complicated layout.

III. Layout Analysis Survey

Document layout analysis is an important technology prior to the optical character recognition (OCR). It includes page segmentation and zone classification. Its result is referred to homologous module for further analysis. The methods of document layout analysis can be classified into three types: the top-down , the bottom-up and the integration methods in the gross**.** The top-down method starts with the global document page, and

segment gradually up to document components. The bottom-up method is a process from part to whole. It incorporates pixels to connected components or characters, connected components to layout components.
Document page segmentation is a necessary step in a document processing system; its objective is to locate different types of contents such as text, graphics, and halftone images from the input document image. In Literature survey page segmentation task was done by two methods namely top-down and bottom-up methods. Basically, in bottom-up analysis connected components are extracted from the image and are subsequently aggregated in higher level structures, creating words, text lines, paragraphs, and so on(e.g. RLSA: Run Length Smoothing Algorithm).

A document image is composed of a variety of physical entities or regions such as text blocks, lines, words, figures, tables, and background. We could also assign functional or logical labels such as sentences, titles, captions, author names, and addresses to some of these regions. The process of document structure and layout analysis tries to decompose a given document image into its component regions and understand their functional roles and relationships. The processing is carried out in multiple steps, such as preprocessing, page decomposition, structure understanding, etc,
Document image physical layout analysis algorithms can be categorized into three classes:

- Top-down approaches,
- Bottom-up approaches,
- Hybrid approaches.

IV. Previous Work on Logical Layout Analysis

Logical layout analysis is required for newspaper document images. A newspaper document image is a visual representation of a printed page of a newspaper page. Typically a newspaper document image consists of blocks of text, i.e., letters, words, and sentences.., that are interspersed with half-tone pictures, line drawings, and symbolic icons. A newspaper document image is therefore a digital two-dimensional array representation of a newspaper document obtained by optically scanning and raster digitizing a hard copy document. Newspaper Document image analysis is the task of recognizing objects in a newspaper image by using techniques that extract homogeneous regions within the image.

Newspaper Document image understanding is the goal-oriented task of deriving a symbolic representation of the contents of a document image, which involves detecting and interpreting different blocks (like title, subtitle, author, contents, etc.), accounting for the interactions of the different components, and coordinating the interpretations to achieve an end result.

Documents generally conform to a certain geometric structure that dictates that the document be composed of a set of interconnecting rectangular printed regions, or blocks. The layout of the different rectangular blocks with respect to each other, i.e., the spatial relationships between these blocks, varies among different types of documents and is dependent on the layout conventions used for specific types of documents.
There is an underlying structure for all printed documents that is governed by certain basic constraints: First, the physical blocks into which all printed documents can be spatially divided represent meaningful physical divisions of the document. Second, each of the physical blocks of printed matter can be classified according to certain basic categories like text, photographs, etc. Third, these physical blocks can be logically grouped to make up units that represent meaningful logical entities in a document, e.g., a newspaper article, etc. Fourth, there exists a specific order in which the text blocks within each unit must be read in order for the information in the block to make sense syntactically and semantically. Thus, spatial knowledge is a very important factor in the task of classifying, grouping, and ordering object blocks in a document image, which is one of the primary objectives of document image understanding.

In content identification methodology, the image is segmented and scanned using Optical Character Recognition (OCR) and Artificial Intelligence techniques are applied to find headings, author names and content. This methodology is more complex. Optical Character Recognition technique does not efficiently identifies the characters like italic, Bold, etc. With these drawbacks in OCR, applying AI does not produce accurate results.

Another researcher produced background analysis system based on the description of white spaces inside regions; it classifies the regions in to text, graphics and images. In this, all blocks in newspaper are classified according to the distribution of white spaces. This algorithm is not more efficient and accurate. This often wrongly interprets small text blocks as titles since the description of white spaces in both cases are nearly same. As a result, this thesis attempts to analyze the layouts of the document images logically as title blocks, sub titles, text, images, content lines etc., to produce more meaningful information from the document images.

V. Objectives of this Research

This research is intended to develop a layout based retrieval system for document image databases consisting of three phases: 1. At first, intelligent layout analysis algorithm has been designed to extract the layouts the document images physically with their edges and rectangles. 2. Every physically identified layout

has been converted into a tree intermediary representation for indexing and storage in layout databases. Later when a query image has been supplied, it retrieves the similar layout images from the layout databases. 3. Finally, a logical layout analysis scheme has been proposed to identify the meaning of the layouts involved in the document images.

Therefore, this idea motivated us to develop a physical layout analysis system for document images which identifies the layouts properly by identifying their physical features and spatial locations. Later, this research has also been motivated to do layout retrieval of document images from document image databases. Finally, this research has also been motivated to develop a logical layout analysis system which analyzes the meanings of the layouts logically. As a result, this research consists of three objectives such as intelligent identification of physical layouts, layout based retrieval system and logical layout analysis from document images.

VI. Materials Used for this Research

Layout based information retrieval from document image databases has been developed successfully by using Java Advanced Imaging (JAI) Package and distance metrics algorithm as a material for this research.

VII. Physical Layout Analysis from Document Images

Physical layout analysis `qq2q1 intends to study the arrangement of layouts or locations of the regions present in a document image before understanding it. Before extracting the text or information from a document image, page segmentation (layout analysis) techniques need to be applied to identify the exact layout (area) where the text or image resides.

SYSTEM ARCHITECTURE

This thesis proposes an intelligent physical layout analysis system (Figure 1) to analyze the physical layouts from document images which consists of the following steps:

1. Preprocessing
2. Smearing
3. White Space Analysis
4. Layout Extraction
5. Text/Image Separation

Intelligent Layout Analysis System:

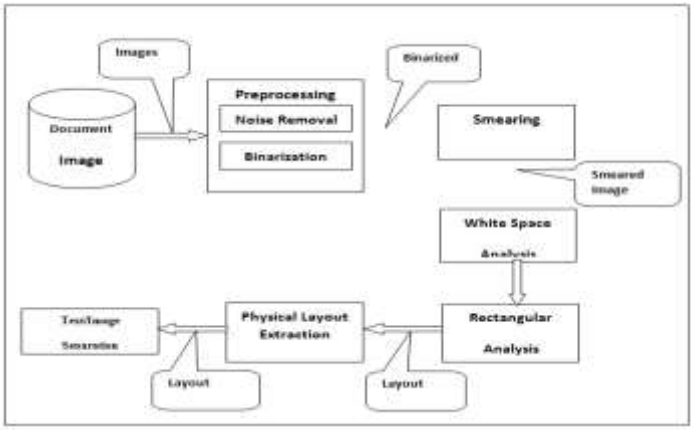

Figure 1: The Intelligent Layout Analysis System

VIII. A Process of Layout Based Information Retrieval:

In this thesis, a system has been developed to retrieve information from digital library documents which is large collections of scanned documents (newspaper, books and journals) based on their layout similarity. The main feature of this system is the ability of allowing a user to find pages with a layout similar to a query page. This allows users to retrieve meaningful pages with a low effort based on layout similarity. In this research, Tree Representation and indexing has been attempted for the retrieval of document images based on layout similarities. This allows a user to find pages with a layout similar page when a query image has been supplied, based on tree structure. Large collections of Newspaper document images are now available in Digital Libraries. Layout Based Retrieval from Newspaper document image system aims to identify the layout in complex document images using the white space analysis approach and retrieves the similar documents by reducing the time complexity and increasing the performance of image retrieval.

Document Image Retrieval (DIR) aims at finding relevant document images from a corpus of digitized pages. The basic idea of document image retrieval is to find documents relying on document image features only. Relevant sub-tasks include the retrieval of documents on the basis of their layout similarity. Most work has concentrated on the processing of converted text with image retrieval techniques. Only fewer methods approached the retrieval through layout similarity.

The proposed system identifies the layout in both simple and complex document images also termed as non-Manhattan images. A hybrid approach for document layout analysis is used here to identify the layouts. Identified zones in the layouts are converted into trees using tree structures and stored in database. Later, relevant images are retrieved from the database when a query image has been supplied, based on its tree structure.

Logical Layout Analysis from Document Images

Logical Layout Analysis to analyze the layouts of the document images logically as title blocks, sub titles, text, images, content lines etc., to produce more meaningful information from the document images.

System Design

The system architecture shown in Figure 2 describes the flow of processes involved in identification of logical layouts in a given input newspaper image. It consists of the following phases:

1. Preprocessing
2. Segmentation
3. Physical Layout Extraction
4. Logical Labeling
5. Logical Grouping of Layouts

Logical Layout Analysis Architecture:

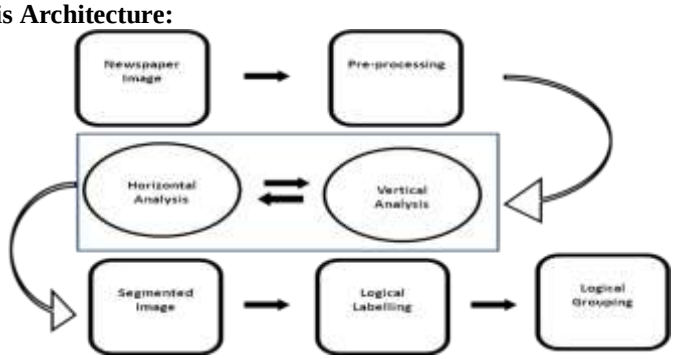

Figure2: The Logical Layout Analysis Architecture

IX. Conclusion:

Document Image Layout Analysis helps us to trace out physical structure of document image from document image corpus. In this research, a hybrid approach for analyzing complex document images or non Manhattan images has been proposed. Hybrid Layout Analysis works along with a text/image separation technique to separate the text and images present in the document images. This Hybrid layout analysis technique tries to record the various kinds of white spaces in the document image irrespective of the positions and identifies the layouts. This help in understanding the organization of layouts in images irrespective of their complexity. Two Statistical properties such as Black run length, Transition rate have been designed in this research to isolate the textual areas from images in the segmented blocks.

Also Logical grouping for those complex layouts are also efficiently done without any conflicts using Run length computation and Threshold calculation. Sampling is done over large number of samples is done and threshold is computed for small, medium and large texts which are maintained in database.

References

[1] Akiyama, T. and Hagita, N. 'Automated Entry System for Printed Documents', Pattern Recognition, Vol. 23, No. 11, pp. 1141-1154, 1990.

[2] Baird, H.S. 'Background Structure in Document Images', Document Image Analysis, H. Bunke, P. Wang, and H.S. Baird, eds., pp. 17-34, World Scientific, 1994.

[3] Baird, H.S., Jones, S.E. and Fortune, S.J. 'Image segmentation by shape-directed covers', In Proceedings of International Conference on Pattern Recognition, pp. 820–825, (Atlantic City, NJ), June 1990.

[4] Breuel, T.M. 'Two Geometric Algorithms for Layout Analysis', Document Analysis Systems, pp. 188-199, August 2002.

[5] Breuel, T.M. 'High Performance Document Layout Analysis', Proceedings of Symposium on Document Image Understanding Technology, April 2003.

[6] Breuel, T. M. 'An algorithm for finding maximal whitespace rectangles at arbitrary orientations for document layout analysis', In Seventh International Conference on Document Analysis and Recognition, pp. 66–70, Edinburgh, UK, August 2003.

[7] Faisal Shafait, Daniel Keysers, and Thomas M. Breuel 'Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms', IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 30, No. 6, June 2008.
[8] Ha, J., Haralick, R.M. and Phillips, I.T. 'Document Page Decomposition by the Bounding-Box Projection Technique', In Proceedings of the Third International Conference on Document Analysis and Recognition, pp. 1119-1122, Montreal, Canada, August 1995.
[9] Ishitani, Y. 'Logical structure analysis of document images based on emergent computation', In Proceedings of the International Conference on Document Analysis and Recognition, pp. 189–192, (Bangalore, India), September 1999.
[10] Kise, K. and Sato, A. 'Segmentation of Page Images Using the Area Voronoi Diagram', Computer Vision and Image Understanding, Vol. 70, No. 3, pp. 370-382, June 1998.
[11] Nagy, G., Seth, S. and Viswanathan, M. 'A Prototype Document Image Analysis System for Technical Journals', Computer, Vol. 25, No. 7, pp. 10-22, July 1992.
[12] Nagy, G. and Seth, S.C. 'Hierarchical Representation of Optically Scanned Documents', In Proceedings of the Seventh International Conference on Pattern Recognition, pp. 347-349, Montreal, Canada, 1984.
[13] Nagy, G., Seth, S.C. and Stoddard, S.D. 'Document analysis with an expert system', In Proceedings Pattern Recognition in Practice II, Amsterdam, The Nederland's, pp. 19-21, June 1985.
[14] O'Gorman, L. 'The Document Spectrum for Page Layout Analysis', IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 15, No. 11, pp. 1162-1173, November 1993.
[15] O'Gorman, L. and Kasturi, R. 'Document Image Analysis', IEEE Computer Society Press, Los Alamitos, CA, 1995.
[16] Pavlidis, T. and Zhou, J. 'Page Segmentation and Classification', Graphical Models and Image Processing, Vol. 54, No. 6, pp. 484-496, 1992.
[17] Shafait, F., Keysers, D. and Breuel, T.M. 'Pixel-accurate representation and evaluation of page segmentation in document images', In Eighteenth International Conference on Pattern Recognition, pp. 872–875, Hong Kong, China,Aug. 2006.
[18] Simone Marinai, Emanuele Marino and Giovanni Soda 'Layout Based Document Image Retrieval by Means of XY Tree Reduction', Proceedings of Eighth International Conference on Document Analysis and Recognition, pp. 432-436, August 2005.
[19] Simone Marinai, Emanuele Marino, Francesca Cesarini and Giovanni Soda 'A General System for the Retrieval of Document Images from Digital Libraries', Proceedings of the first International Workshop on Document Image Analysis for Libraries, Vol. 18, No. 14, pp. 274-299, 2004.
[20] Simone Marinai, Emanuele Marino, Francesca Cesarini and Giovanni Soda 'Tree Clustering for layout-based document image retrieval', Proceedings of the Second International conference on Document Image Analysis for Libraries, , pp. 337-359, 2006.
[21] Wong, K.J., Casey, R.G. and Wahl. F.M. 'Document Analysis System', IBM Journal of Reserch and Development, Vol. 26, No. 6, pp. 647-656, 1982.
[22] K. Kise, A. Sato, and M. Iwata, "Segmentation of Page Images Using the Area Voronoi Diagram," Computer Vision and Image Understanding, vol. 70, no. 3, pp. 370-382, June 1998.
[23] G. Nagy, S. Seth, and M. Viswanathan, "A Prototype Document Image Analysis System for Technical Journals," Computer, vol. 25, no. 7, pp. 10-22, July 1992.
[24] Simone Marinai, Emanuele Marino, Giovanni Soda.-"Tree clustering for layout- based document image retrieval" IEEE-2006
[25] Huaigu Cao, Rohit Prasad, Prem Natarajan, Ehry MacRostie-"Robust Page Segmentation Based on Smearing and Error Correction Unifying Top-down and Bottom-up Approaches".IEEE-2007.