Document Layout Analysis & Classification and Its Application in OCR

Gaurav Gupta, Shobhit Niranjan, Ankit Shrivastava
Dept. of Electrical Engineering
Indian Institute of Technology Kanpur
gauravg@iitk.ac.in, nshobhit@iitk.ac.in, ankits@iitk.ac.in

Dr RMK Sinha
Dept. of Electrical Engineering and Computer Science
Indian Institute of Technology Kanpur
rmk@cse.iitk.ac.in

Abstract—Digitization of paper-bound documents is one of the foremost commercial interests worldwide. First step in all such applications is transforming a paper bound document into an electronic document by scanning, subsequently applying to the image OCR to generate textual information from the document image. In this paper we describe our work that acts as a pre-processing stage for OCR application. Automatic document layout extraction and segmentation is done using spatial configuration of various text/image segments represented as bounded boxes; this segmented layout is than analyzed with certain heuristic tests and each segment is assigned labels (title, authors, abstract, body, header, footer etc). This information is than passed on to OCR module as an XML interface, accelerating it’s performance by allowing it to label recognized text segments and identifying only those parts of the document which have text resulting saving in computation. Although, the work has been motivated for application to an automated machine translation system preserving the overall document layout, it has a number of other applications such as in information retrieval, search etc. This information is also being used to classify technical documents into three categories which can be extended to any number of classes based on spatial configuration heuristics.

I. INTRODUCTION

With the ever-increasing use of computers and the Internet, electronic storage of information is a growing need. Much of this information is initially in the form of a physical document, and needs to be converted to digital information [1].

The segmentation and layout analysis algorithms form part of Document Image Processing that lead to a number of applications in ‘OCR’, information retrieval, machine translation and in development of automated document “understanding” system [2].

Document layout analysis, which is the most important step in the document analysis and recognition processes, is the function of separating text, graphics, and images, and then extracting isolated text blocks (layout objects) such as titles, paragraphs, headers, footers and captions [3]. It is difficult to extract textual information from differently formatted documents because the documents do not have the exactly identical formats [4]. There have been various attempts to translate document images into logical text forms such as SGML. Some researchers have used the search templates or translation criteria including structural data and translation rules [5] [6] [7]. These templates or rules are generally format specific and the systems are not designed to allow an end user to generate and input new templates or rules easily and often require a great deal of manual labour.

In order to efficiently extract a given piece of textual information from sufficiently similar documents, the documents must each have a corresponding text image area where the desired information is to be found. The structural data are generally organized to maintain logical as well as positional relationships among the textual elements and textual image areas in a specific type of a document. In order to solve this and other problems, especially when a large number of documents must be analyzed, we design a robust layout extraction algorithm which works with almost every type of document image. Subsequently we identify the important features of the input document and design classification based on this. The feature identification algorithm is based on the relations graph representation of the document. In this paper we focus on different types of technical documents in standard formats.

II. PREVIOUS WORK DONE

Mostly bottom-up approach has been used to generate layout information from a document. Ishitani [3] generates the layout based on emergent computation model. The approach adopted by Mitchell P.E. et. al. [1] is to extract all the patterns from an image, and store them in a linked list. Patterns are defined by merging adjacent rectangles, that are thought of as a rectangular region of loosely connected black pixel of a character. Then layout is generated using the process of grouping patterns into blocks based on spatial proximity and classification. After that ordering of merged blocks is done in a natural manner that helped in classification. In the past, such systems have been built that generate near perfect document layout extraction and document labeling along with their classification. However, they are based on a hybrid approach of using geometrical layout properties as well as textual features as extracted by a commercial ’OCR’ system [8], [9]. On the other hand, our system itself is the pre-processing stage for OCR and so do not need OCR to extract document layout extraction, labeling and classification. For classification of document in Li et. al. [2] first introduce a new segmentation methodology, then describe the document layout structure representation in terms of a Directed Weight Graph (DWG). However, Kochi et. al. [4] uses an algorithm
based on a distance measure based similarity criterion that is highly dependent on the relative position of fields of user defined template and fields of extracted layout. In this case the user is required to define a new template for every different layout structure.

III. OUR METHODOLOGY

A. Layout extraction

1) Pattern Extraction layer: First we generate the binary image of the give image by thresholding it at 80% of its intensity and inverting the resulting image. In the binary image white pixels represent the characters, graphics and lines etc. The first step in pattern extraction is to locate rectangular regions called ‘rect’. A rect is a rectangular region of loosely connected white pixels [1], that encloses a certain logical part of the document. We considered simple 8-neighborhood connectivity and performed connected component (contour) analysis of the binary image leading to the segmentation of the textual components. For next part of algorithm we use the minimum bounding rectangle of contours. These rectangles were then sorted top-to-bottom and left-to-right order, using 2D point information of leftmost-topmost corner. Smaller connected patterns were discarded based on the assumption that they may have originated due to noise dependent on image acquisition system and does not in any way contribute to the final layout. Also punctuation marks were neglected using smaller size criterion e.g. comma, full-stop etc. At this level we also segregate the fonts based on the height of the bounding rect using $avgh$ (average height) as threshold. Two thresholds are used to classify fonts into three categories - small, normal and large. Sample output of font analysis is shown in figure 1.

$$\frac{2 \times avgh}{3} \leq normal \ fonts_h \leq \frac{3 \times avgh}{2}$$

2) Region growing layer: Having obtained letter level bounding boxes, we merge them horizontally to obtain sentence level bounding boxes, and subsequently vertical merging to obtain rects enclosing paragraphs. If two local regions or patterns overlap horizontally within predefined scope (on both sides: left and right), they are merged with each other i.e. the horizontal distance between them is less than the minimum value of their horizontal merge scopes. These distances are defined as the minimum distance between a character in one region and a character in the other region. This allows detection of the distance between two local regions even if their boundaries overlap [3]. The output after this step for a input image is shown in figure 2(a) and (b).

3) Region analysis layer: The aim of region analysis layer is to decide whether a region is textual, graphical or tabular. Regions are classified individually based on size, aspect ratio, area and run length statistics. Rules are defined as follows:

- After region growing layer graphical parts are combined such that their aspect ratio lies within an upper and a lower bound. Generally they have rectangular structure and have larger area (compared to other bounding boxes after contour analysis).
- After horizontal merging, regions that have textual component will have large aspect ratio and their height close to average-height.
- Lines will have very large aspect ratio as compared to textual and graphical components.
- Tabular region is detected using X-Y cut algorithm based on projection profiles of bounding boxes as suggested by Liang et al [9].

The final extracted image part for a input image is shown in figure 3(a) and 3(b).

4) Pattern Blocking: After region growing by merging the local patterns horizontally, and region analysis step, we need to club these horizontal blocks to generate global layout structure for feature detection and classification stage. For this we simply use morphological operation Dilation [11] as region growing algorithm on the binary image of above stage. The structuring element for dilation is chosen as rectangle and limit of dilation is determined experimentally.

B. Layout analysis

In this part we perform layout analysis for document classification and identification.

1) Generic modeling of paper layout: Here we define a standard model for technical paper; similar structure can be modeled for other type of documents also e.g. office letters, news-papers etc. We characterize a generalized technical paper using standard features:

- Such a document image may have logo of the editor which is either graphical or textual. We make use of the property that it’s bounding box has aspect ratio close to 1 and logos appear either at left/right most corner or both.
- It must have (in case of presence of logo after it) title of the paper at the top. Bounding box related to this field has larger aspect ratio and only textual component appears in it.
- After ‘Title’ field it must have authors’ name, their affiliations, address etc. They appear before abstract block and bounding boxes are of large aspect ratio (more than 2).
• It must have an ‘abstract’ block after the title which may spread in to two columns or single column. It comes strictly after the author field and have very large area and normal aspect ratio (0.5 to 2.5)
• Above fields are followed by rest of the parts of a technical paper e.g. Introduction, section, subsection which may have textual and/or graphical components. This may be in single column structure or double column structure.

2) Feature extraction and identification:
Based on above generic modeling we analyze the extracted layout and create vertical and horizontal histograms of the output image. Using the histograms scanning top to bottom and simultaneously using certain heuristics, we segregate the paper into structure of above mentioned components. We also identify the fields and rectangles bounding them. Based on this information we generate XML file compatible to OCR system having above mentioned generic structure.

3) Classification:
Having found the exact structure of the given layout we can classify the document images into different categories by simple comparison of document structures with standard templates based on above model. For example, IEEE papers have logo at left corner, title, authors, abstract (single column), and paper structure double column while LNCS format can be summarized as logo at left corner, title, author name and addresses, and abstract and paper structure both as 2 column type. We categorize papers into three categories based on following two criteria:

1) The abstract of the document could be spanning the
TABLE I

<table>
<thead>
<tr>
<th>Paper Type 1</th>
<th>Paper Type 2</th>
<th>Paper Type 3</th>
<th>none</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper Type 1</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Paper Type 2</td>
<td>0</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Paper Type 3</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

We demonstrated that our layout extraction algorithm is robust and yields a high performance. We also identify logical fields of the documents and perform classification based on these. We generate XML output denoting nature of various fields such as title, authors, abstract etc. This XML file is used for OCR of different text zones with additional information on expected font size and the nature of text. The file, thus generated besides providing electronic storage of the information contained in the document, has a variety of applications. The document can be reproduced in its original form. As the title, headings, sub-headings, figure captions and other zones in the document are identified and stored, both the tasks of information retrieval and text summarization get additional handles. Further, in case of an automated machine translation of the textual content of the document, after the OCR output is fed to the language translator, the XML file is used for reformatting of the output so that the original document layout is preserved to a large extent. Further work is continuing in this direction.

REFERENCES