Text Extraction from PDF document

D. Sasirekha
Research Scholar
Karpagam University
Coimbatore

E. Chandra, Ph.D
Director, computer Science
Dr SNS Rajalakshmi College of arts and Science
Coimbatore

ABSTRACT

Documents in PDF format are nowadays called the Universal document format. PDF to speech converter systems involves many steps to achieve. Text extraction is the primary step From PDF to do further processing. In this paper we start with the brief discussion about the steps involved in extracting the text from PDF documents. The aim of this paper is to give the introduction with some basic concepts on PDF, and with text extraction concepts, which will be useful for the readers who are less familiar in this area of research.

Keywords

Text extraction, PDF, Text extraction technique

1. INTRODUCTION

Nowadays increasing number of documents are available in PDF document [Fig.1] as considered as a favorite file format. The reason behind this are listed below [1][2][3][4]

It is an open standard for document exchange, created by Adobe Systems in 1993 is used for representing documents in a manner independent of software, hardware, and operating systems.

Each PDF file encapsulates a complete layout description of the document, including the text, fonts, graphics, and other information needed to display it. PDF files allow to download complex information efficiently by compression technique.

PDF files care for the source file information and appear like original documents even compiled from multiple formats with text, drawings, multimedia, video, 3D, maps, full-color graphics, photos, and easily share files with others using different platform such as Mac, Windows, Linux, etc. Simple file sharing makes that will look the same (layout, fonts) on multiple computer systems.

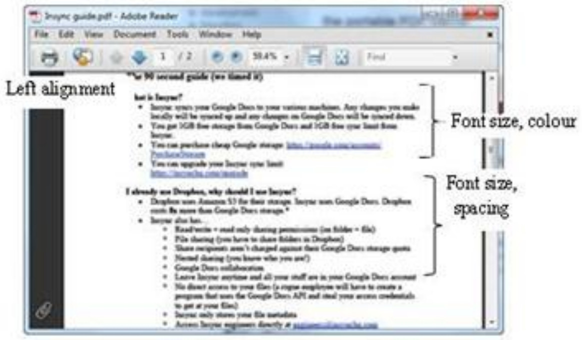

Insync guide.pdf - Adobe Reader
File Edit View Document Tools Window Help
1 / 2
50.4%
Find
Left alignment
Font size, colour
Font size, spacing

Fig 1:Example of Portable Document Format

File sharing is protected from unauthorized viewing, printing, copying, or editing. Creating files with hyperlinks and bookmarks that can be easily shared via email and on the internet.

The text extraction from PDF is a challenging problem to make document analysis, classification and process. Significant amount research is developing in recent times. The major issue in extracting the text depends on the font size, font style, alignment Layout and table.

In the next chapters we explained about the steps involved in extracting text from PDF.

.

2. TEXT EXTRACTION- OVERVIEW

In PDF document, the text is stored as a set of strings objects; each object gathers a combination of characters, coordinating the font and other information to make Glyphs (visual representation of characters [6].

The advantage of PDF document analysis is that the character and layout information obtained from the PDF parser is much richer and more accurate than acquired from OCR[5].

Even then, there exist some challenges in extracting text from PDF documents. This is due to the reason that the PDF documents are generated by different tools and composed of different type of objects that makes vary in font, font size, alignment and the gap between the text lines.

3. RELATED ARCHITECTURE

The Document Analysis presented in PDF helps to examine the appearance and geometric position of text and image in the complete document [9]. The geometric position the point having geometric elements {eg. Group of text lines or a photograph that form a "block" followed by paragraph break}. Detecting the edges first and generating the line feature vector map would help in detecting text regions [10].

4. STEPS FOR TEXT EXTRACTION

The main steps involved [Fig 2] in the text extraction are

1. Layout Analysis
2. Segmentation
3. Character / Symbol recognition
4. Structure Recognition

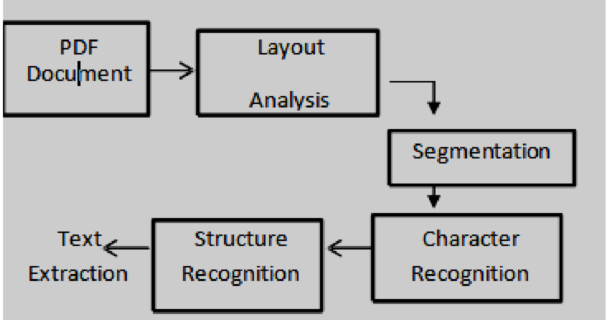
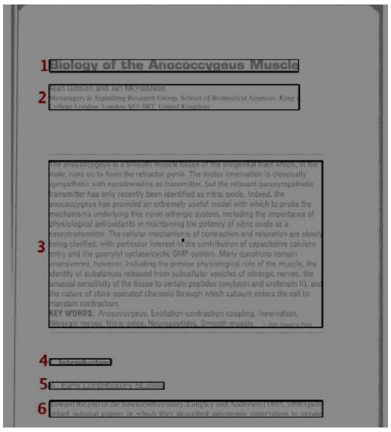


Fig 2: Steps for Text Extraction

4.1 Layout Analysis

Layout analysis [Fig 3] is a key step in document capture conversions into electronic formats, optical character recognition (OCR), information retrieval from scanned documents, There are number of novel geometric algorithms and statistical methods available. Layout analysis systems built from these algorithms are applicable to a wide variety of languages and layouts, and have proven to be robust to the presence of noise and spurious features in a page image.



1 Biology of the Anococcygeus Muscle

2 Alan Gibson and Ian McFadzean
Messengers & Signalling Research Group, School of Biomedical Sciences, King's College London, London SE1 9RT, United Kingdom

3 The anococcygeus is a smooth muscle tissue of the urogenital tract which, in the male, runs on to form the retractor penis. The motor innervation is classically sympathetic with noradrenaline as transmitter, but the relaxant parasympathetic transmitter has only recently been identified as nitric oxide. Indeed, the anococcygeus has provided an extremely useful model with which to probe the mechanisms underlying this novel nitrergic system, including the importance of physiological antioxidants in maintaining the potency of nitric oxide as a neurotransmitter. The cellular mechanisms of contraction and relaxation are slowly being clarified, with particular interest in the contribution of capacitative calcium entry and the guanylyl cyclase/cyclic GMP system. Many questions remain unanswered, however, including the precise physiological role of the muscle, the identity of substances released from subcellular vesicles of nitrergic nerves, the unusual sensitivity of the tissue to certain peptides (oxytocin and urotensin II), and the nature of store-operated channels through which calcium enters the cell to maintain contraction.
KEY WORDS: Anococcygeus, Excitation-contraction coupling, Innervation, Nitrergic nerves, Nitric oxide, Neuropeptides, Smooth muscle. © 2001 Academic Press

4 I. Introduction

5 A. Early Contributory Studies

6 Toward the end of the nineteenth century, Langley and Anderson (1895, 1896) published seminal papers in which they described autonomic innervation to organs

Fig 3: Example of a PDF document with its Geometrical Layout Information

The layout of a document refers to the physical location and boundaries of various regions in the document .The progression of Layout Analysis aims to decompose a document into a hierarchy of homogenous regions, such as photos, background, text blocks, text lines, words, characters, etc.

The layout analysis algorithm is classified mainly into two groups depending on their approach. Bottom-up algorithms start with the smallest components of a document (pixels or connected components) that group them to form larger region. The top-down algorithms start with the complete document and divide it repeatedly to form smaller and smaller regions. Each approaches its own advantage and they work good in specific situations. In addition hybrid approach is available that uses a combination of top-down and bottom-up strategies [13]

The set of logical or functional entities in a document, along with their inter-relationships is referred to as the Logical Structure of the document. This is normally performed on the results of the layout analysis stage. But in many complex documents, layout analysis would require some of the logical information about the regions to perform correct segmentation [11].

4.2 Segmentation

Segmentation algorithms are designed to handle complex document layouts and backgrounds, Document[Fig 4] .Segmentation is a basic task in storage and retrieval systems, and is either attacked by bottom-up or top down methods.

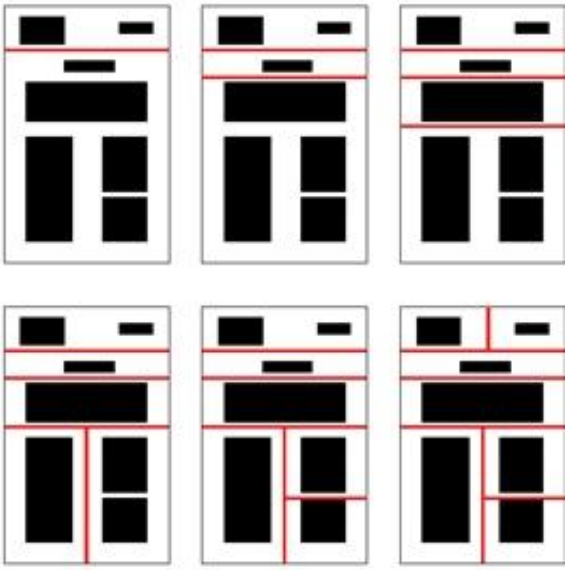


Fig 4: An example for sequential cuts that segment a page into different regions

Page segmentation algorithms can be broadly classified into three categories: bottom-up ,top-down and hybrid algorithms. The traditional document segmentation algorithms give good results on most documents with complex layouts but assume the script in the document to be simple as in English. These algorithms fail to give good results on the documents with complex scripts[12]. Segmentation is a labeling process which consists in assigning the same label to spatially aligned units (such as pixels, connected components or characteristic points). The choice of a segmentation technique depends on the complexity of the structure of the PDF document. Therefore the aim of the segmentation algorithm should be that partitions a document with complex scripts.

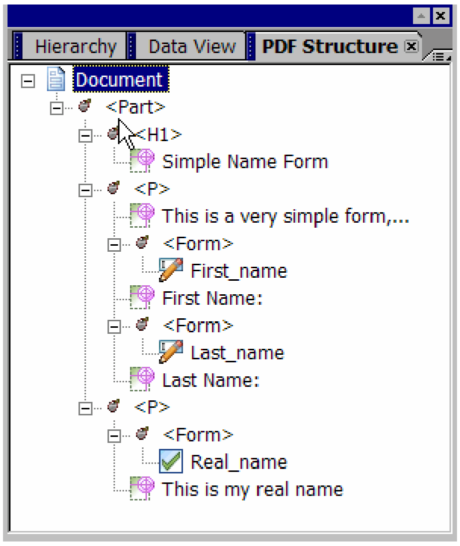
4.3 Character Recognition

OCR or Optical character recognition is a system through which a computer looks at the pattern in an image. This translation of image containing to text to actual machine encoded text. This is the first important step in starting of character recognition.

PDF character recognition is the process by which characters are recognized from PDF files and placed into text searchable ones. Optical character recognition (OCR) technology is an important part of PDF character recognition system and it is responsible for the extraction of text from PDF files. Without PDF character recognition, scanned PDF files have a number of drawbacks which limit their usage [8][9].

One major drawback associated with scanned PDF files is that they cannot be searched through with a text string. They have to be read through page by page to get to the relevant information. Once such information is located it cannot be extracted from scanned PDF files. PDF character recognition system helps to overcome all these problems. Therefore, the manual extraction process can be side stepped and information can be brought in faster.

4.4 Structure Recognition



```
Document
  <Part>
    <H1>
      Simple Name Form
    <P>
      This is a very simple form,...
      <Form>
        First_name
      First Name:
      <Form>
        Last_name
      Last Name:
    <P>
      <Form>
        Real_name
      This is my real name
```

Fig 5: PDF structure palette

Structure understanding [Fig 5] deals with logical descriptions of regions rather than their physical attributes. The logical structure of a document is a mapping from the physical regions in the documents to their logical labels. Document structure analysis is the process of assigning the logical labels to physical regions identified during layout analysis.

Document physical structure analysis procedures also have performance uncertainties and so may provide uncertain input to the logical structure analysis process. Stochastic models, represented by stochastic grammars and related parsing techniques could be used to address these problems. The input to the parser could be regarded as probabilistic to reflect uncertainty due to erroneous physical layout analysis results. Physical layout and logical structure analysis algorithms based on stochastic language models are in trend [13].

5 CONCLUSION

This paper made a clear and a simple overview of working of text extraction from PDF document in step by step process. There are many ready-made systems available in the market and also much improvisation is currently going on in research area to extract the text regions from the document containing images, tables, borders etc. We expect the researchers to do advancements in each step of text extraction that make extracting the text regions smooth in complex document format.

6. REFERENCES

[1] http://desktoppub.about.com/od/electronicpublishing/g/pdf.htm

[2] http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml

[3] http://www.techterms.com/definition/pdf

[4] http://www.webopedia.com/TERM/P/PDF.html

[5] Lin, X., Gao, L., Tang, Z., Lin, X., & Hu, X. 2011. Mathematical formula identification in PDF documents. In Document Analysis and Recognition (ICDAR), 2011 International Conference on (pp. 1419-1423)

[6] AJEDIG, M. A., Li, F., & ur Rehman, A. 2011. A PDF Text Extractor Based on PDF-Renderer. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1)

[7] Gupta, G., Niranjan, S., Shrivastava, A., & Sinha, R. 2006. Document Layout Analysis and Classification and Its Application in OCR. In Enterprise Distributed Object Computing Conference Workshops, 2006. EDOCW'06. 10th IEEE International (pp. 58-58)

[8] Williams S.Lovegrove and David F.Brailsford 1995 Document analysis of PDF files: methods, results and implications”, Electronic publishing ,vol.8 (2&3),20-220.

[9] S.Audithan, R M. Chandrasekaran 2009 Document text extraction from document images using Haar Discrete Wavelet Transform” , EJSR.

[10] Claudie Faure, Nicole Vincent 2009 Simultaneous detection of vertical and horizontal text lines based on perceptual organization Proc. SPIE 7247, Document Recognition and Retrieval XVI, 72470M doi:10.1117/12.805504,2009

[11] K.S. Sesh Kumar, Anoop M. Namboodiri, and C.V. Jawahar 2006 Learning segmentation of documents with complex scripts ICVGIP’06 Proceedings of the 5th Indian Conference on Computer Vision, Graphics and Image Processing, pp. 749-760.

[12] Song Mao, Azriel Rosenfeld, and Tapas Kanungo 2003 Document structure analysis algorithms: A literature survey Vol. 5010 of SPIE Proceedings, SPIE, pp. 197-207.

[13] Tamir Hassan“ Object-Level Document Analysis of PDF Files”, DocEng’09, September 16-18, 2009, Munich, Germany.