ISSN : 2393-8188 (print)
2393-8196 (online)
www.milliyasrcollege.org.journal.php

URDU DOCUMENT IMAGE LAYOUT ANALYSIS AND CHALLENGES IN SEGMENTATION

Imran Khan Pathan[1], *Mohammed Zeeshan Shaikh*[2]

[1]Dept. of Computer Science, *Milliya Arts, Science & Management Science College, Beed,*
[2]*Head, Department of Computer Technology, YCIP, Beed,*
[1]*imk.pathan@gmail.com* ,[2]*Zeeshan.shaikh@gmail.com*

ABSTRACT:

Document Image Processing and Character Recognition are subdomains of Pattern Recognition which deals with digitization of documents and converting it into editable text. Digitization of documents like Various Forms, Magazines, Books, Broachers, Newspaper, Bank Slips, Check etc. is becoming essential with rise of automation in almost all official sectors. Urdu Document Image Processing is lagging behind due to dilemma of segmentation of Urdu document into paragraph, lines, words and characters. In present research paper an attempt is made to study various document segmentation techniques and test their accuracy on Urdu document segmentation.

KEY WORDS: Optical Character Recognition, Document Image Layout Analysis, Physical Layout Analysis, Logical Layout Analysis, Segmentation, Projection Profile

1. INTRODUCTION:

In an automatic Document Image Processing System a Physical document is digitized by means of digitizing devices and these scanned documents are called document images. Every document consists of different layout and structure according to their type. The process of identifying the basic components in a document and then reconstructing their hierarchical organization is called *Document Layout Analysis*. Document Layout Analysis is one of the significant step in converting document images into digital and editable form. It plays key role to find the different parts or element of a document like Titles, section, header, footer etc. and make it possible to read the document page in correct reading order. For example if a page consists of three columns then reading order should be correct for an OCR system. Organization of these elements is known as Document Layout and decides the type of document. Document entities like paragraph, columns, text lines, words, tables, figures, page background and watermark etc. are known as physical entities. Logical entities are the sentences, titles, captions, author names, footer, foot note etc. A document like admission form or a bank cheque consist different layout with various fields in it.

Human can easily recognize the type of document and identify it as magazines, books, broachers, newspaper, enrolment forms, research papers, bank slips or a simple page. But for Automation of the same using computer system is challenging due to large range of document types. The complexity of document layout varies for each type of document. The complexity of document increases if document contain complex background, headings, two or more columned text, Watermark, artistic text formatting etc.

2. Document Layout Analysis

In Document Structure and Layout Analysis the document image is decompose into various component regions according to their functional roles and relationships. Every Document image consists of different layout and structure accordingly. It is required to be split during Document Image Analysis. The process of identifying the basic components in a document and then reconstructing their hierarchical organization is called Layout Analysis [1].

A typical Document Layout Analysis System include noise removal, skew correction, segmentation i.e. structure and layout analysis and the output regions are forwarded for further character recognition system Figure 1-1 illustrate it.

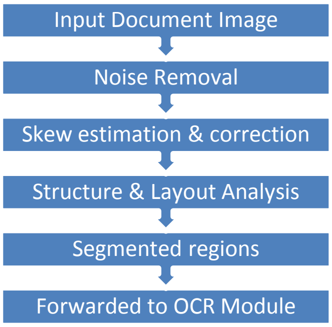

Fig.1-1 Document Image Layout Analysis System

In Document Structure and Layout Analysis, the document image is decomposed into various regions according to their functional roles and relationships [2]. In which document image is segmented into homogeneous regions and assigned logical meaning to them. Figure 1-2 shows layout of an Urdu image that needs to be segmented and to identify the physical and logical entities.

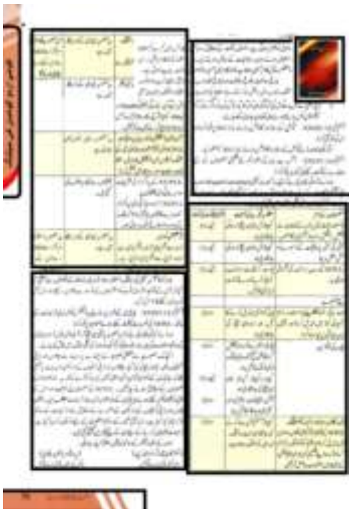

Figure 1-2 Document layout of an Urdu document image

Each document has specific geometric page structure. Basically a document Layout Analysis can be categorized into Physical Layout Analysis and Logical Layout Analysis. Physical layout analysis deals with the identification of the geometric structure and its segmentation into uniform regions. Logical layout analysis assigns logical meaning to these regions like title, authors, footer etc.

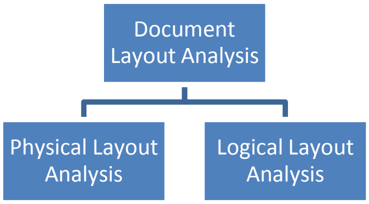

Figure 1-3 Document Layout Analysis

2.1. Physical Layout Analysis:

In the physical layout analysis, the page decomposition is performed. An image is segmented into homogeneous blocks of maximum size which are classified into a set of predefined data type which is called block classification. In page segmentation, only the geometric layout of page is considered and block classification step provides the knowledge of data type e.g. text, picture, or chart etc. In [3], an appropriate review on various page decomposition techniques has been given.

On the basis of processing order, the Document layout analysis algorithms are primarily divided into top-down approaches and bottom-up approaches. Top-down method splits the complete document image into smaller homogeneous meaningful regions. And bottom-up approach begins with pixel level. It typically uses connected component analysis and merging techniques. If neighbouring connected pixel has similar features, it groups them into larger regions like words, line, non-text etc. [1, 2]. On the basis of the objectives, the page decomposition algorithms are broadly classified into four categories.

- Text Segmentation Approach
- Page Segmentation Approach
- Segmentation / Classification Mixed Approach
- Block Classification Approach.

Text Segmentation Approach:

The algorithms of this group directly extract and segment the text by analyzing the document image. The extracted text is arranged in the hierarchical structure as classified columns, paragraphs, lines, word etc. Such type of algorithms are preferred in case of document contains only text or some non-text elements. The techniques which are useful for text segmentation are Connected Component Analysis, Projection Profile Methods and Texture Based or Local Analysis etc.

Page Segmentation Approach:

In this approach, the document image is partitioned into homogeneous regions. Few famous techniques for the page segmentation are such as Smearing technique, Projection profile analysis, Texture based or local analysis, Analysis of background structure.

Segmentation/Classification mixed Approach:

In this approach, the document image is segmented and classified simultaneously, because it is not possible to split the segmentation step from classification step. For this connected component analysis, Smearing and texture or local analysis techniques are used.

Block Classification Approach:

Here, the algorithms give label to the previously segmented regions. The major part of labeling is based on feature extraction and linear discriminate classifiers. The techniques for block classification approach can be grouped as Feature extraction, Linear Discriminant classifiers, Binary classification and Neural Networks.

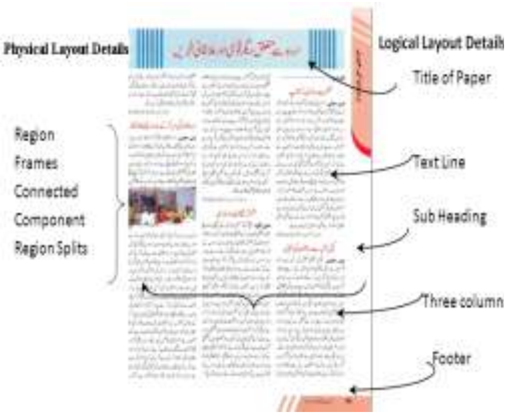

Figure 1-4 Physical and logical structure of Urdu Document

2.2. Logical Layout Analysis:

Once physical layout analysis is done, the document image is split into various components. And then in logical layout analysis gives specific meanings to every component like title, author, header, footer etc. Usually features like font type; font size, formatting, and size of blocks are considered for logical layout analysis. Indeed, the role played by a layout component represents meta-information that could be exploited to label the document in order to help its filing, handling and retrieval in a collection or library. The logical components can be organized in a hierarchical structure, which is called the logical structure [4].

The Figure 1-4 illustrates the Physical and Logical Layout Analysis components of Urdu document image. Here, Physical Layout Analysis gives details like regions, frames, connected components etc. And logical layout analysis provides details regarding title, text lines, sub heading, number of columns and footer.

3. Urdu Document Layout Analysis:

Urdu is a very complex script and Urdu Document Image Processing is considered difficult as compared with other scripts. Research on Urdu Document Layout Analysis is lagging behind due to complexities of Urdu script and challenges in Urdu text segmentation. Character Recognition of

Urdu Naskh Script based document is more difficult than Roman script. Character recognition of Nastaliq Script is even more complicated due to the inherent characteristics of its writing style. In any typical Document Layout Analysis character recognition system, there are four levels of segmentation namely Paragraph, lines, words and characters. First the text region is segmented into lines of text, latter each of the lines of text is split into words and then each word into constituent characters. In case of Urdu script, the segmentation of words into character is more challenging than any other script and is an open field of research.

Even segmentation of lines into words, which is quite straight forward in most of the scripts, is not easy because of vertical overlapping in words and ligatures. Another unique feature of Urdu is that the Urdu words are usually written without short vowels or diacritic symbols [5]. In case of machine translation, Urdu language suffers from segmentation dilemma. In a fruitful research on automatic Urdu to Hindi translation system [6], the segmentation problems in Urdu language are discussed with appropriate examples.

3.1. Segmentation Problem in Urdu documents due to Overlapping:

Segmentation is considered as one of the difficult tasks in Urdu character recognition system due to its writing style and overlapping words and characters.

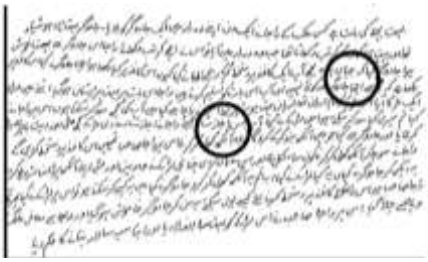

Figure 1-5 Lines and words overlapping in handwritten Urdu document image

Figure 1-5 shows the lines and words of handwritten Urdu which are overlapped (highlighted by circle) and hence, it is difficult to segment by using projection profile and other techniques. Segmentation of line from the paragraph seems to be very difficult task.

Word Segmentation problem:

Word segmentation is a non-trivial task, as there is not much difference between inter-word and intra-word vertical gap. It is very difficult to judge if the two adjacent ligatures belong to same or different words. For example, in Figure 1-6, for a person not familiar in Urdu script, it is very hard to tell how many words are there in text line by looking at inter-word gap.

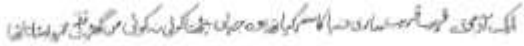

Figure 1-6 Handwritten urdu line with overlapped words

Character segmentation:

It is another difficult task in Urdu Document Segmentation and the majority of the researchers working on Urdu Document Image Processing stay away from this phase. It is very difficult to break a ligature into isolated characters. As shown in Figure 1-7 it is very complex to segment the characters from the word due to overlapping nature of Urdu word.

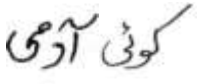

Figure 1-7 Overlapping of Urdu Sub-words

As character segmentation is a very complex task, so researchers mainly prefer to segmentation free approach in Urdu Optical Character Recognition. In segmentation free approach, the ligature as a whole is used instead of segmenting it into smaller units.

4. Results based on Project Profile

Projection Profile has been tested on some parts of Document Images to test its accuracy of this technique for segmentation of lines in Urdu document images. First this method is tested on printed English text lines to illustrate how it can be used to segment the words from line and characters from words respectively. As shown in Figure 1-8 all the characters in printed English text can be easily segment using projection profile method. But Urdu is a cursive script in which projection analysis method is not satisfactory, the reasons are horizontal overlapping and touching characters. Figure 1-9 shows the failure of projection profile method in case of cursive Urdu text. The accuracy of projection profile based segmentation is poorer in case of handwritten Urdu text as shown in Figure 1-10.

Simple vertical projection analysis the histogram based method doesn't work effectively for tilted text like Urdu. Another complication in Urdu character segmentation is the presence of diacritic and secondary components, which may generate false segmentation points. Vertical projection analysis method is not much suitable in case of handwritten cursive text because it consists of slants, overlapped, broken and touching characters. In our experimental work projection profile method is only partially successful for line segmentation of Urdu Document Image. Segmentation due to overlapping, secondary strokes, diacritic and absence of accurate baseline are the primary reasons behind slow progress in automatic Urdu Document Image Processing and Character Recognition System. Result in Figure 1-11Shows the Line Segmentation Problem due to Overlapping in Urdu Script which is one of the major dilemmas in Urdu Document Segmentation. The combination of projection profile method with Connected Component Analysis or Morphological Image Processing may be useful for such segmentation.

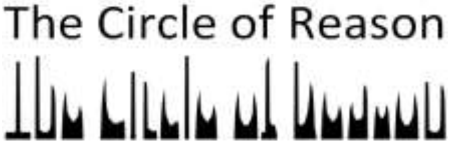

Figure 1-8 Projection profile of printed English text

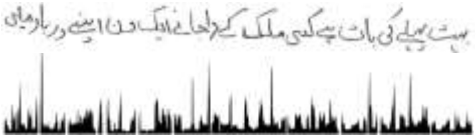

Figure 1-9 Projection profile of printed Urdu text

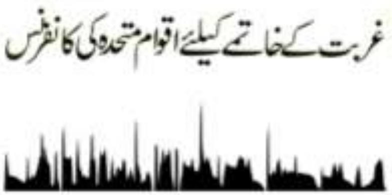

Figure 1-10 Projection profile of handwritten Urdu text

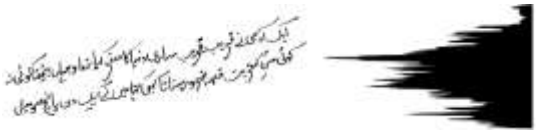

Figure 1-11 Line Segmentation Problem due to Overlapping in Urdu Script

5. CONCLUSION:

In present research Projection Profile based segmentation method has been tested to study its results on the Urdu Document Image. But projection profile based method doesn't found suitable to split the Urdu Document Image into Paragraph, Lines, Words and Characters respectively. The reasons are word and character overlapping, presence of the secondary strokes, diacritic and absence of accurate baseline in Urdu Script.

6. REFERENCES:

[1] Simone Marinai: Introduction to Document Analysis and Recognition, Studies in Computational Intelligence (SCI) 90, 1–20 (2008)

[2] Anoop M Namboodiri and Anil Jain Document Structure and Layout Analysis, in Digital Document Processing: Major Directions and Recent Advances B. B. Chaudhuri (ed.), Springer-Verlag, London, (ISBN:978-1-84628-501-1), Jan. 2007, pp. 29--48

[3]R. Cattoni and T. Coianiz and S. Messelodi and C. M. Modena, ‘Geometric layout analysis techniques for document image understanding: a review, 1998

[4]S. Ferilli, Automatic Digital Document Processing and Management, Advances in Pattern Recognition, DOI 10.1007/978-0-85729-198-1_5, © Springer-Verlag London Limited 2011.

[5] Project proposal, (2011) "Software Requirement, Design, & Testing documents, Development of Robust Document Image Understanding System for Documents in Indian Scripts Phase II" ,Sponsored By, Ministry of Communication & Information Technology, Govt. of India. Phase II TDIL.

[6] Gurpreet Singh Lehal, (2010) " A Word Segmentation System for Handling Space Omission Problem in Urdu Script", Proc. of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010, pages 43–50,