Volume 7, Issue 1, January 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

K-D Indexing in Printed Trilingual Documents

[1]Mahesha D M, [2]Gopalan N P

[1]Dept. of Computer Science, Centre for Research and Development, PRIST University, Thanjavur, Tamilnadu, India

[2] Department of Computer Applications, NIT, Trichirappalli, Tamilnadu, India

DOI: 10.23956/ijarcsse/V6I12/0222

Abstract— In this work, we proposed indexing model for script identification. Rectangular White Space analysis algorithm is used to analyze and identify heterogeneous layouts of document images. To speed up the script identification, we focus on designing an indexing mechanism for tri-lingual scripts for optimizing the subsequent robust identification system. For representation, we extract features from Gabor responses and also using scale invariant feature transform. We considered a set of global features and index by Kd-tree. For experimentation, we have used our own database. Experimental results reveal that indexing prior to identification is faster than conventional identification method in terms of time for scripts.

Keywords— Segmentation; Section finding; Section Merge; Feature Extraction; Indexing

I.INTRODUCTION

Anything which conveys information is known as a document. Generally, a document is a knowledge container. Most of the times we acquire knowledge from documents such as Newspapers, Textbooks, Scientific journals, Magazines, Technical reports, Office files, Postal letters, Bank cheques, Application forms etc. (Tang et al., [1]). To understand the huge information, an extensive amount of manual processing is required and such a manual processing is very much time consuming. To overcome this difficulty, it is essential to automate the manual process which needs efficient algorithms. This automation process is considered as document image processing (DIP). In general, the document image processing is divided into text processing and graphics processing. Text processing is further divided into character recognition and page layout analysis. Graphics processing is further divided into line processing and region processing as shown in Figure1.

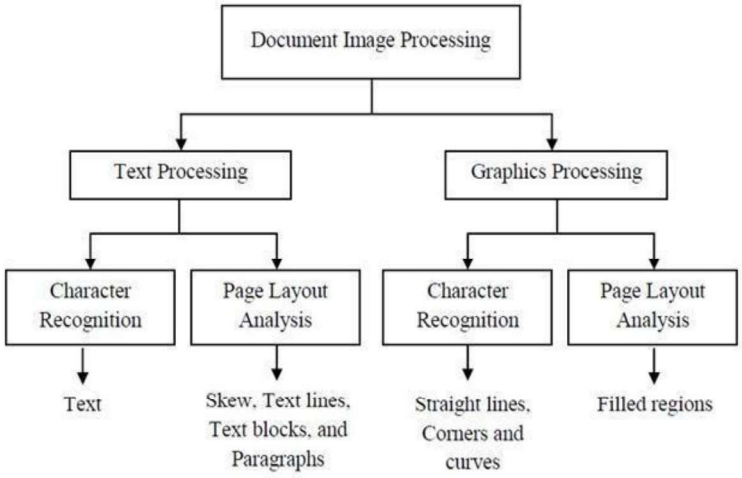

Figure 1. Hierarchy of document image processing with subcategories

A. Stages in Document Image Processing

The document image processing involves three basic steps at conceptual levels, which are document image analysis, document image recognition and document image understanding. Within these three levels, there are several other interacting modules such as image acquisition, binarization, block segmentation, block classification, logical block grouping, character and word recognition, picture processing and analysis, graphic analysis, picture understanding, text understanding and graphics understanding. The interactions between these processes and data flow between levels are shown in Figure 2.

Volume 7, Issue 2, February 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Segmentation of Tri-Lingual Documents

[1]Mahesha D M, [2]Bhavya D N, [3]Nandini H M
[1]Department of Studies in Computer Science, Karnataka State Open University, Mysore,India
[2]Department of Studies in Computer Science, Karnataka State Open University, Mysore,India
[3]Department of Studies in information technology, Karnataka State Open University, Mysore, India

Abstract— Physical layout analysis 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. In Page Segmentation, Top-down methods are simple and efficient but fail in non Manhattan layouts. In contrast, Bottom-up approaches adapt non Manhattan layouts easily than the top down approaches, but heavily depend on the threshold, parameters and extensive computations for layout identification. On the other hand, Hybrid methods (Bruel [31], Bruel [32]) suits well for layout identification by eliminating the dependency on threshold and parameters. But this analyzes the white background of the image with small white rectangles and merges them to locate the content blocks. Merging of small white rectangles makes the identification process tedious since large number of small white rectangles gets involved in the image. In addition, this approach heavily relies on heuristics for merging operations, which affects the segmentation rate considerably. In all the above reported methods (Bottom up and Hybrid approaches), connected component analysis (requires more number of pixel visits) is required to identify black and white components from the image. Therefore, the above shortcomings motivated this research towards designing a White Space analysis technique which eliminates the usage of the connected component analysis (to identify white spaces), heuristics, threshold and prior knowledge. As a result, in this thesis, Rectangular White Space Analysis (RWSA) 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 and this hybrid approach has been explained in the subsequent section.

Keywords— Segmentation; Section finding; Section Merge; Feature Extraction; Indexing

I.INTRODUCTION

Anything which conveys information is known as a document. Generally, a document is a knowledge container. Most of the times we acquire knowledge from documents such as Newspapers, Textbooks, Scientific journals, Magazines, Technical reports, Office files, Postal letters, Bank cheques, Application forms etc. (Tang et al., [1]). To understand the huge information, an extensive amount of manual processing is required and such a manual processing is very much time consuming. To overcome this difficulty, it is essential to automate the manual process which needs efficient algorithms. This automation process is considered as document image processing (DIP). In general, the document image processing is divided into text processing and graphics processing. Text processing is further divided into character recognition and page layout analysis. Graphics processing is further divided into line processing and region processing as shown in Figure 1.

A. Stages in Document Image Processing

The document image processing involves three basic steps at conceptual levels, which are document image analysis, document image recognition and document image understanding. Within these three levels, there are several other interacting modules such as image acquisition, binarization, block segmentation, block classification, logical block grouping, character and word recognition, picture processing and analysis, graphic analysis, picture understanding, text understanding and graphics understanding. The interactions between these processes and data flow between levels are shown in Figure 2.

1) Document Image Analysis

Document image analysis is a process of recovering syntactic and semantic information from images of documents, prominently from scanned versions of paper documents. There are two distinct tasks in document image analysis. The first has a syntactical goal consisting of the identification of basic components of the document, the so-called document objects. The second has a semantic goal consisting of the identification of the role and meaning of the document objects in order to have an interpretation of the whole original document. The structural analysis, on the other hand involves usage of layout clues to identify headlines, locate different lines, etc. In general, image analysis involves

International Journal of Computer Engineering and Applications,
Volume XII, Issue I, Jan. 18, www.ijcea.com ISSN 2321-3469

TEXTURAL FEATURES IN SCRIPT IDENTIFICATION FOR PRINTED BILINGUAL DOCUMENTS

Mahesha D M [1], Bhavya D N [2], Nandini H M [3]

[1]*Department of Studies in Computer Science, Karnataka State Open University, Mysore, India*
[2]*Department of Studies in Computer Science, Karnataka State Open University, Mysore, India*
[3]*Department of Studies in Information Technology, Karnataka State Open University, Mysore, India*

ABSTRACT:

In this work, we investigate the effect of texture features for script classification. Rectangular White Space analysis algorithm is used to analyze and identify heterogeneous layouts of document images. The texture features, namely the color texture moments, Local binary pattern (LBP) and responses of Gabor, LM-filter, S-filter, R-filter are extracted, and combinations of these are considered in the classification. In this work, a probabilistic neural network and Nearest Neighbor are used for classification. To corroborate the efficacy of the proposed method, an experiment was conducted on our own data set. The experiment was conducted for various sizes of the datasets, to study the effect of classification accuracy, and the results show that the combination of multiple features vastly improves the performance.

Keywords: Segmentation, Section finding, Section Merge, Feature Extraction, Classification.

[1] INTRODUCTION

Anything which conveys information is known as a document. Generally, a document is a knowledge container. Most of the times we acquire knowledge from documents such as Newspapers, Textbooks, Scientific journals, Magazines, Technical reports, Office files, Postal letters, Bank cheques, Application forms etc. (Tang et al., [1]). To understand the huge information, an extensive amount of manual processing is required and such a manual processing is very much time consuming. To overcome this difficulty, it is essential to automate the manual process which needs efficient algorithms. This automation process is considered as document