An Agent Based Algorithm for Document Analysis (ABADA)

K. Jambi, M. Saleh, H. Barhamtooshi, F. Essa and A. Ezz
Computer Science Department
King Abdul Aziz University
Jeddah, Saudi Arabia

ABSTRACT The aim of document layout analysis is to extract the geometric structure for the document image. The introduced system is deigned to work with a variety of documents, without prior knowledge about the nature of the document. This algorithm mainly depends on dividing the document into strips or runs, these runs ease the document handling and enable the ability of handling a part of the document and then handling the other part. This is too important when handling memory, and when downloading documents from networks. It can be calssified as a hybrid tecnique of top-down, and bottom-up, to overcome the disadvantages in each staregy. Agent tecnology is used to ease the operation of the system through the network. Expermental results reveal the proposed approach is effective.

1. Introduction

Documents can be viewed as paper-based documents, or digitized documents. Paper-based documents are processed for information extraction by human which make these documents a labor force consuming. Moreover, These documents need much space for storing. On the other hand, digitized documents can be viewed as electronic images which are machine searchable. This ability makes electronic documents preferable for its fast search capabilities as well as less storage media.

Layout analysis is the extraction of geometric structure from the document image [1]. The given document image is generally in binary format. It is segmented into several objects due to location, contents, homogenity, and other attributes. Document layout analysis plays an important role in automated documet processing environment as it classifies different parts of the document before deciding which later subsystems to use[2]. Structural layout analysis can be achieved using top-down or bottom-up techniques [3].

In this work, a document image processing system is developed. This system is based on an agent technology. The document image analysis system is used to analyze and classify the document contents in details with a high accuracy. Agents are used to go through networks in parrallel to reduce the time and to collaborate the documents to get specific predefined documents. This paper deals with two different concepts. The first concept covers the intelligent agent technology, and the second one deals with the process of document analysis. In the following two sections, we will discuss these two concepts.

1.1 Document Analysis

The concept of Document Image Processing is used for performing analysis and understanding several types of documents. A document is a set of related pages to express a subject. A page is a set of objects (range from text, graphics, or images) that take part in describing the mentioned subject. The page is segmented into equi height parts called runs.