Table Structure Identification from Document Images: A Survey

Sandeep Kaur
Research Student, M.Tech(CSE) Department of Computer Science, Punjabi University, Patiala,
Dharam Veer Sharma
Department of Computer Science, Punjabi University, Patiala

ABSTRACT
Table structure identification has received significant research attention in the past few years. The OCR (Optical Character Recognition) has faced many potential errors in the document images, so it is strongly required to make logical objects such as table explicit. So to get the deep understanding about the contents of the document, proper understanding of the document is required by means of many algorithms. In this research paper, I emphasized to describe various methods or algorithms which segment the scanned images into different blocks and detect any tabular structure in any form that may be present in the document.

INTRODUCTION
A large number of pages are to be scanned and analyzed to create document image libraries targeted to real world applications. Creating a document image library involves a chain of thorough and intense activities like scanning, pre-processing, segmentation, layout analysis, storage and retrieval, etc. despite being the most researched field in the domain of Document Image Analysis (DIA), the problems are yet to be solved up to the desired level of accuracy and efficiency. There are lots of methods proposed by different persons for the identification of tabular structure from the document images along with some benefits as well as some limitations. To find the tabular structure from document image classification on the presence of any tabular structure in a page lead to better segmentation at a lower computing cost. The term “tabular structure” resembles with a table. There are large number of methods and algorithms proposed by different persons for the detection/segmentation of tabular structure. In this research paper we present a brief review of the past work under the Table category.

Table: A table contains at least two rows and two columns, which may be fully or partially embedded in boxes formed by horizontal and vertical rule lines. Table detection and segmentation have been done in several ways at different times. The algorithms may be classified broadly into two types. These are as follows:
1. Based on the presence of rule lines in the table and
2. Based on the knowledge of table layout.

Identification of Table: Our main topic of interest is identification of table which can be done by the following steps:-
1. **Table Detection**: Locating the regions of a document with a tabular content.
2. **Table Structure Recognition**: Reconstructing the cellular structure of a table.
3. **Table Interpretation**: Rediscovering the meaning of the tabular structure. This includes:-
(a) Functional Analysis: Determining the function of cells and their abstract logical relationship. 
(b) Semantic interpretation: Understanding the semantic of the table in terms of entities represented in the table, their attributes, and the mutual relationships between such entities.

METHODOLOGIES
There are lots of works done on the identification and segmentation of the tabular structure from the document images. There are some popular and efficient approaches used for the identification and segmentation of document images along with advantages and disadvantages. Some of them are described as follows:

Table Structure Recognition Based On Robust Block Segmentation:  
(Thomas G.Kieninger)
This paper presents an efficient approach to identify tabular structures within either electronic or paper documents. The resulting T-Recs system takes word bounding box information as input, and outputs the corresponding logical text block units. Starting with an arbitrary word as block seed the algorithm recursively expands this block to all words that interleave with their vertical (north and south) neighbors. Since even smallest gaps of table columns prevent their words from mutual interleaving, this initial segmentation is able to identify and isolate such columns. In order to deal with some inherent segmentation errors caused by isolated lines (e.g. headers), overhanging words, or cells spawning more than one column, a series of post processing steps is added. These steps benefit from a very simple distinction between type 1 and type 2 blocks: types 1 block are those of at most one word per line, all others are of type 2. This distinction allows the selective application of heuristics to each group of blocks. The conjoint decomposition of column blocks into subsets of table cells leads to the final block segmentation of a homogeneous abstraction level. These segments serve the final layout analysis which identifies table environments and cells that are stretching over several rows and/or columns. The “Initial Clustering Algorithm” is used in this approach.

Strengths:
 a. It operates on either electronic (ASCII) or paper (OCR output) document.
 b. It neglects grid lines. This allows a faster image preprocessing.
 c. It disregards textual contents. Thus it can be applied to low quality documents such as facsimiles to perform its block segmentation.
 d. It detects table columns with very narrow gaps.
 e. It is able to handle non-Manhattan layout, since we do not perform any projections to find column candidates. Blocks do not even need to be left or right justified.
 f. It detects table-like structures within regular text without the presence of characteristic table headers.
 g. It detects table cells that span more than one document line.
 h. It is universally applicable to any document.

Weakness:
a. It would not be able to cluster single lines as in headers because they do not have vertical neighbor.
b. It also causes errors with regular paragraphs that have a space at the same x position on all lines. The paragraph would hence not be identified as only one homogeneous block.

**Conclusion:**
With T-Recs we have built an extremely robust and very fast system that already proved to be of interest for industrial partners. We are aiming towards three main goals:

a. He wants to further improve the results by involving even more document inherent features in the restructuring processes.

b. The improvement should affect the structure analysis of the segmented blocks.

c. He wants to develop techniques for an objective benchmarking of the results.


(N.Chen and D.Blostein)
This research paper is based on document image classification. Document image classification is an important step in Office Automation, Digital Libraries, and other document image analysis applications. There is great diversity in document image classifiers: they differ in the problems they solve, in the use of training data to construct class models, and in the choice of document features and classification algorithms. We survey this diverse literature using three components: the problem statement, the classifier architecture, and performance evaluation. This brings to light important issues in designing a document classifier, including the definition of document classes, the choice of document features and feature representation, and the choice of classification algorithms. We emphasize techniques that classify single-page typeset document images without using OCR results. Developing a general, adaptable, high-performance classifier is challenging due to the great variety of documents, the diverse criteria used to define document classes, and the ambiguity that arises due to ill-defined or fuzzy document classes. The classifier architecture includes four aspects:

a. Document features and Recognition stage

b. Feature representation

c. Class models and
d. Classification algorithms.

**Conclusion:**
In this research paper the document classification literature along three components: the problem statement, the classifier architecture, and performance evaluation. The problem statement is characterized in terms of the document space and the set of document classes. The classifier architecture includes four aspects: document features and recognition stage, features representation, class models and classification algorithms, and learning mechanisms. Advances in performance evaluation techniques for document classifiers are needed. Exiting standard document database have been used to test document classifiers.

**A survey of the Table Recognition: includes Models, Observations, Transformation and Inferences**

(Zanibbi et al.)
In this survey, the table recognition literature is presented as an interaction of table models, observations, transformations, and inferences. Table characteristics vary widely. Consequently, a great variety of computational approaches have been applied to table recognition. A table model defines the physical and logical structure of tables; the model is used to detect tables, and to analyze and decompose the detected tables. Observations perform feature measurements and data lookup, transformations alter or restructure data, and inferences generate and test hypothesis. This presentation clarifies the decisions made by a table recognizer, and the assumption and inferenceing techniques that underlie these decisions.

**Conclusion:**
This research paper presented the table recognition literature from the viewpoint that table recognizers may be understood as sequences of decisions supported by observation and transformations of available data. These decisions are made relative to a table model describing the location and composition of tables in a set of documents. Other avenues for future work include extending the proposed table models of Wang and others, further exploring content-based methods for recognition, improving experimental design and evaluation techniques at both the level of individual decisions and whole system defining corpii of ground-truth tables for use in the community, and exploring new observations, transformation, and inferences for use in a table recognition.

**Learning to detect Tables in a Scanned Document Images using Line Information**
(Kaser et al.)

This paper presents a method to detect table regions in document images by identifying the column and row line-separators and their properties. The method employs a run-length approach to identify the horizontal and vertical lines present in the input image. From each group of intersecting horizontal and vertical lines, a set of 26 low-level features are extracted and an SVM classifier is used to test if it belongs to a table or not. The performance of the method is evaluated on a heterogeneous corpus of French, English and Arabic documents that contain various types of table structures and compared with that of the Tesseract OCR system. The method yields promising results on a heterogeneous collection of documents. It uses a learned classifier to classify table regions without resorting to heuristic rules. Identification of the row and column line separators during the table detection stage implicitly gives the information about each cell of the table which is useful in subsequent table processing steps to understanding its logical structure.

**Conclusion:**
This research paper presented a new method for table detection in scanned document images using the properties of line separators. The method yields promising results on a heterogeneous collection of documents. It uses a learned classifier to classify table regions without resorting to a heuristic rules. Unlike other image-based techniques, this method relies only on the presence of lines and do not require any text analysis. This makes it insensitive to the layout, presence of multiple scripts and handwritten elements. While the method performs well for tables completely enclosed by lines, there are, in practice, other types of table layouts that may contain only parallel lines that separate the table headers or with no lines at all.
References: