A Unified Algorithm for Identification of Various Tabular Structures from Document Images

Sekhar Mandal, Bengal Engineering and Science University, Shibpur, India
Amit K. Das, Bengal Engineering and Science University, Shibpur, India
Partha Bhowmick, Indian Institute of Technology Kharagpur, India
Bhabatosh Chanda, Indian Statistical Institute, Kolkata, India

ABSTRACT
This paper presents a unified algorithm for segmentation and identification of various tabular structures from document page images. Such tabular structures include conventional tables and displayed math-zones, as well as Table of Contents (TOC) and Index pages. After analyzing the page composition, the algorithm initially classifies the input set of document pages into tabular and non-tabular pages. A tabular page contains at least one of the tabular structures, whereas a non-tabular page does not contain any. The approach is unified in the sense that it is able to identify all tabular structures from a tabular page, which leads to a considerable simplification of document image segmentation in a novel manner. Such unification also results in speeding up the segmentation process, because the existing methodologies produce time-consuming solutions for treating different tabular structures as separate physical entities. Distinguishing features of different kinds of tabular structures have been used in stages in order to ensure the simplicity and efficiency of the algorithm and demonstrated by exhaustive experimental results.

Keywords: Document Image Segmentation, Index Page Detection, Math-Zone Detection, Table Detection, Tabular Structures, TOC Detection

INTRODUCTION
Billions of pages are to be scanned and analyzed to create document image libraries targeted to real-world applications. The task is daunting; however, there is a pressing need for these libraries, as we witness a spurt of activities in recent times in industries as well as in academia. Creation of a document image library involves a chain of thorough and intense activities like scanning, per-processing, segmentation, layout analysis, storage and retrieval, etc. Hence, it is still constrained with the requirement

DOI: 10.4018/jdls.2011040103
of huge manual workloads—particularly for the segmentation and identification of page constituents as a part of the layout analysis. 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. Most of the currently used methods have been designed and tested with some typical applications in mind and have moderate performance apropos the huge variety of document pages in real life. An integrated approach to segmentation and physical analysis of document images is, therefore, imminent for broad classification of the pages utilizing some common criteria. A classification based on the presence of any tabular structure in a page may lead to better segmentation at a lower computing cost. It may be noted that, by the term “tabular structure”, we mean anything that visually resembles a table. Careful observation would reveal that tabular items are quite common in document pages; for, any table, table of content (TOC), index page, and most of the displayed math-zones exhibit the tabular nature. Moreover, a simple test based on spatial criteria is enough to detect these tabular structures. With this backdrop, here, a novel approach is presented, treating any page as tabular or non-tabular for reaping its benefit in subsequent DIA tasks.

The paper is organized as follows. The next section presents a brief review of the existing works in the related field. The basic stages of our algorithm have been explained in two subsequent sections: In the first one, we show how a document page is classified as a tabular or a non-tabular page using certain characteristic features of tables, index pages, and math-zones. The detailed procedures of finally classifying all sorts of tabular structures are explained with examples in the second one. The next section contains experimental results including performance-wise comparison with some existing methods. Finally, we conclude the paper with the future possibilities that evolve out of our work.

PAST WORK

The approach of treating the text as either tabular or non-tabular is not yet proposed by others. Hence, citation with similar approaches is not possible. However, there are various works available in the literature for detection/segmentation of tabular components, e.g., table, TOC, displayed math, etc. In this section we present a brief review of the past work under the following categories.

1. Table: Table detection and segmentation have been done in several ways (Chandran et al., 1996; Mandal et al., 2006b; Tsuruoka et al., 2001; Watanabe et al., 1995). The algorithms may be classified broadly into two types: one based on the presence of rule lines in the table and the other based on the knowledge of table layout. Watanabe et al. (1995) have proposed a tree for representation of the structure of various kinds of tables. In Chandran et al. (1996), the horizontal and vertical lines of the table are used to recognize the structure of the tabulated data. In Itonori (1993), a similar technique is found, which also uses row-column pairing and the relationship of cells and ruled lines. Zuyev (1997) has defined a table grid, and has described simple and compound cells of any table based on the table grid. Node property matrix has been used by Tanaka and Tsuruoka (1998) in the processing of irregular rule lines and generation of HTML files. Method of analysis for unknown table structure has been proposed by Belaid, Panchevre, and Belaid (1998). Tersteegen and Wenzel (1998) have proposed a system for extraction of tabular structure (table only) with the help of predefined reference table. Tsuruoka et al. (2001) have presented a segmentation method for complex tables with or without rule lines. A technique has been described by Das and Chanda (1998a) to separate out tables and headings in document images.