LOCATING TABLES IN SCANNED DOCUMENTS WITH HETEROGENEOUS LAYOUT

M.A.C. Akmal Jahan[1*], M.A.C. Jiffriya[1] and Roshan G. Ragel[2]
[1]*Post Graduate Institute of Science, University of Peradeniya*
[2]*Department of Computer Engineering, University of Peradeniya*
[*]*akmaljahan@fas.seu.ac.lk*

The pool of knowledge available to the mankind depends on the source of learning resources, which can vary from ancient printed documents to present electronic materials. The rapid conversion of material available in traditional libraries to digital form needs a significant amount of work for format preservation. Most of the printed documents contain not only characters and its formatting but also some associated non text objects such as tables, charts and graphical objects. Since most of the existing optical character recognition techniques face challenges in detecting such objects and do not concentrate on the format preservation of the contents while reproducing them, we attempt to locate all type of tables in scanned documents with heterogeneous layout. Generally all the documents with multi columns are not purely divided by the inter column space. Long headings, centered aligned page numbers, lengthy text in headers and footer and horizontal lines extremely interfere the inter column space which was commonly used in layout analysis. To address this issue, we propose an algorithm using specific threshold to eliminate the interfering parts in inter column space and using local thresholds for word space and line height to detect and extract all categories of tables from scanned documents. From the experiment performed in 50 documents, we conclude that our algorithm has an overall accuracy of about 73% in detecting tables from multi-column layout. Even though complex layout document still have some problem, the system could treat some of these kind of documents as well. Since the algorithm does not completely depend on number of columns, inter column spaces, rule lines which bound the tables, it can detect all categories of tables in a range of different layout scanned documents.

Keywords: Optical character recognition