

Chapter 9

CONCLUSIONS AND FUTURE WORK

With the large number of existing documents and the increasing speed in the production of multitude new documents, finding efficient methods to process these documents for their content retrieval and storage becomes critical. For the last three decades, document image analysis researchers have successfully developed many outstanding methods for character recognition, page segmentation and understand of text-based documents. Most of these methods were not designed to handle documents containing complex objects, such as tables. We developed a table structure understanding system which can detect and decompose table structures form document images. Our algorithm is probability based, where the probabilities are estimated from an extensive training set of various kinds of measurements of distances between the terminal and non-terminal entities with which the algorithm works. The off-line probabilities estimated in the training then drive all decisions in the on-line table structure understand modules.

Although there are constant interests on table structure understanding problem, there are no publicly available table ground truth data sets. Nonetheless, large data sets with ground truth are essential in assessing the performance of computer vision algorithms. Manually generating document ground truth proved to be very costly. For table ground truth, the problems brought by the operators' bias are difficult to be eliminated. We developed a software package that can simulate any given table ground truth with additional controlled variety. We demonstrated the feasibility of our algorithm on a real image data set and used the synthetic table data to aid our table structure understanding research. The software package is publicly available.

We presented a table structure understanding performance evaluation protocol. A large quantity of table structure ground-truth data, varying in quality, are manually or automatically generated. Using this dataset and the proposed performance evaluation protocol, several table structure understanding algorithms are evaluated and compared.

The increasing ubiquity of the Internet has brought about a constantly increasing amount of online publications. Tables are used frequently in web documents. The automatic understanding of tables has many applications including knowledge management, information retrieval, web mining, summarization, and content delivery to mobile devices. We propose a new machine learning based approach for table detection from generic web documents. We introduce a set of novel features which reflect the layout as well as content characteristics of tables. These features are then used in a tree classifier and a support vector machine classifier trained on thousands of examples. To facilitate the training and evaluation of the table classifier, we designed a novel web document table ground truthing protocol and used it to build a large table ground truth database. Experiments on this database using the cross validation method demonstrate a significant performance improvement over the previously developed rule-based system.

Given segmented zone entities and document image, zone content classification determines the zone types. In a complete document image understanding system, the zone classification technique plays the key role. In the design of a zone classifier, a set of measurements are first done and calculated along different directions. We employ a decision tree classifier for the classification. Two methods are used to optimize the decision tree classifier by eliminating the data over-fitting problem. To enrich our model, we incorporated the context constraints to classification for some zones. We propose a performance evaluation for zone content classification experiments. Our zone content classification algorithms are evaluated on the University of Washington English Document Image Database-III, which contains 1,600 document images and

24,177 zones.

A document structure analysis system converts a scanned document page or a document encoded by a Page Description Language (PDL), such as PostScript and Portable Document Format (PDF), into a well partitioned hierarchical representation that reliably identifies the basic document components – text words, text lines, and text blocks. Thus, extracting words (word segmentation) from a scanned document page or a PDF is an important and basic step in document structure analysis and understanding system. We present a text word extraction algorithm that takes a set of bounding boxes of glyphs and their associated text lines of a given document and partitions the glyphs into a set of text words, using only the geometric information of the input glyphs. The algorithm is statistical based. The performance evaluation of our algorithm and two other algorithms are done on the University of Washington English Document Image Database-III. There is strong evidence that the detection performance of the statistical-based algorithm is significantly better than the other two algorithms.

The future extension of the current work includes:

- Functional labeling of text entities;
- Mathematical equation/formula detection and analysis;
- Functional and structural analysis and table interpretation of table structure in web documents.