WDA01 Discussion Group Summary
Web as 2D Documents: Tables and Images

Scribe: Yalin Wang
University of Washington
ylwang@u.washington.edu

Chairman: Larry Spitz
Document Recognition Technologies, Inc.
spitz@docrec.com

Abstract

There were 12 people participating this discussion. The discussion was led by Larry Spitz and scribed by Yalin Wang. First we addressed open problems related to the discussion topic: Web as 2D Documents: Tables and Images. We concentrated on three items from the list of problems. During the discussion, research progress was reported and new research ideas were shared by the participants.

1 Participants

The following people contributed to the group discussion: Henry Baird, Dov Dori, Karim Hadjar, Rolf Ingold, Ram Kashi, David Kennedy, Jisheng Liang, Minoro Mori, Ihsin Phillips, Larry Spitiz, Yalin Wang, Minoru Yoshida.

2 Open Problems Related to the Discussion Topic

In the beginning of the discussion, Larry Spitz asked everyone to think of some open problems related to the discussion topic. Nine open problems were considered by all of us.

1. Table location and understanding in web documents;
2. Logical labeling in web documents;
3. Web content easily viewing;
4. Techniques for hand-hold devices;
5. Comparing/contrasting WWW vs. scanned images;
6. Ground truth for table understanding research;
7. Web document pseudo rendering;
8. Engineering drawing recognition in web documents;

9. WWW specific OCR.

Later, problems 2, 5, and 6 were discussed in detail.

3 Table Ground Truth

To develop and compare different table understanding algorithms, a common large table ground truth database is required. Among existing web table ground truth databases is one developed by the University of Tokyo. The database contains tens of thousands of table tags of which approximately 200 have had manually annotated ground truth added. However, it is probably not publicly available. To build such a database, a table ground truth standard is necessary. A group led by Ihsin Phillips in Queens College, CUNY is working on this problem. We were also advised that copyright problems might be avoided if the final database comprised a list of URLs instead of the documents themselves.

Ihsin Phillips asked what domain(s) we should consider in building a table ground truth data set. Although we had some discussion on this topic, we did not reach a clear answer for this question.

We also talked about the reasons table understanding in web documents is a difficult problem. We concluded that a principal reason is that the table tag is a much abused term in many HTML files.

4 Logical Labeling

To attack this problem in web documents, Henry Baird suggested one possible solution is to build a personal library. He also suggested token-based searching as a potential research direction.

Jisheng Liang suggested that the research can start with some small problems. For example, finding titles of tables or captions of images in web documents. Rolf Ingold predicted that the possible solution is to use global ontology but specific strategy for each application.

Dov Dori said that electrical indexing would be a very useful tool. In his experience, he spent a painfully long time with indexing when he wrote his books.

5 Compare/Contrast WWW vs. Scanned Images

Henry Baird said that for most problems, research on the grey level is better than that on the binary level. Also it seems that nobody is using continuous tune analysis.

Since viewing the web pages with some text whose color is very similar to the background color is really difficult, Larry Spitz thought it would be an interesting problem for WWW rendering.

6 Conclusion

All of us actively participated in the discussion. From the discussion, some common interests might lead to future research cooperations. After the discussion, Yalin Wang reported the discussion summary to the whole workshop.