Report on Web Document Analysis Discussion Group

Robert Haralick [1], Dimosthenis Karatzas [2]

[1] Department of Electrical Engineering
University of Washington, Seattle, WA 98195 U.S.A.
haralick@george.ee.washington.edu
[2] Department of Computer Science, University of Liverpool,
Peach Street, Liverpool, L69 7ZF, United Kingdom
karatzas@csc.liv.ac.uk

Abstract. During DAS'2000 the participants of the workshop had the opportunity to form three discussion groups, in order to address issues that they consider to be important for Document Analysis nowadays. What follows is a synopsis of the main points for the Web Document Analysis discussion group. A list of the participants and their details can be found at the end of this sort paper.

Introduction

The importance of the exploding expansion of the World Wide Web for the Document Analysis field, was early identified either as a possible new area for applications or as a threat for the field, since the classical definition of what a document is has to change in order to include the electronic document as well as the paper document. The main concern of this discussion group, was to assess the current state of the World Wide Web and Web Documents, and comment on the possibilities of using current Document Analysis methods on Web Documents and on the future of Document Analysis in conjunction to the wide spread of World Wide Web.

In order to properly address the issue of Web Document Analysis, we tried to identify the basic characteristics of Web Documents and the World Wide Web itself and the rules that control the creation of Web pages at the present time. The first chapter of this paper lists the common characteristics of Web Documents and the common rules followed in Web Document creation that we consider being global throughout the Web Community. The second part of this paper focuses on the possibility to identify and categorize different types of Web Documents, and the importance of ground truth data and suggests a way to create a first Web Document Analysis database of documents and their ground truth data. Finally, a summary of all the suggestions is given in chapter three.

1. Web Document Characteristics

There is a slight confusion about what should be defined as a Web Document. There are two representations that are essentially different views of the same document, the first is the actual HTML code of a Web page and the second is the output we finally view on screen when the HTML code is parsed by a Web browser. In a perfect world it should not make a difference whether we define the Web document as the HTML description or the final output. However, it seems like there are different pieces of information stored in each representation. For instance, the HTML code could provide information about the filenames of the images included, but the actual content of images is not going to be added to the document until the Web browser parses the HTML code and presents some output. Furthermore, there is a great amount of unused information stored in the HTML code that never gets shown in the final output. As an example we could mention a special type of spamming, where the designer writes key-words in the same colour with the background in order to achieve a higher ranking in search engines, this text is never visible at the final output and should be ignored. Finally, the output document can change slightly depending on the browser

we use to view the Web page. The Web Document should be defined as a combination of the HTML code and the final output, containing all the useful information from both sides.

The determinant difference between Web Documents and all the other types of documents is that the first ones are specifically designed to be viewed on a computer monitor (or any other type of electronic display), whereas the classically defined documents are designed to be printed on and read from a piece of paper. This fundamental divergence is the origin of most of the differences that we identified between Web documents and paper documents.

In contrast to other types of documents, there isn't any widely accepted standard format for Web pages. The use of WWW is so manifold, and the creators of Web pages so many (potentially everyone) that no specific format could possibly be adapted by all. It is a fact that people do not follow rules when creating Web pages - following rules would be contrary to the definition of creativity anyway -. There is an attempt to introduce specific tags in order to store structural information, but until now there is no standard use of tags. In fact, most of the tags get filled in automatically by the authoring tool being used.

Another key issue is the fact that Web documents can be dynamic. This means that human do not always directly create the Web documents we see. Instead they might write code in some scripting language that creates a Web document per request according to certain info, possibly different for each user, or even worse, they might choose to save as HTML code a document they create in totally irrelevant software (such as Microsoft's Word or Excel). The HTML code of these automatically created Web pages usually contains a big amount of irrelevant information. Furthermore, a Web document can be dynamic not only as far as it concerns its creation but also its viewing. It is possible to have interactive documents that the user can change during view time.

As mentioned before, Web Documents are designed to be viewed on monitors. That makes easy the use of colour in Web Documents, something that every designer seems to employ in order to create impact to the reader of his documents. This fact could prove useful since we can use this colour information during Document Analysis. However, because a Web Document contains many different entities, colour information is defined in many different places throughout the Web Document and in many different ways. For example colour information for a paragraph of text can be defined directly in HTML code, but for an image like the one containing the title of that page, is defined inside the image file itself.

Finally, a very important observation about Web documents is that just because they are designed with the use of computers for displaying in computers, they do not suffer from any type of image degradation like the scanned documents.

2. Key Issues and Proposals

Trying to find ways to initiate document analysis for Web Documents, Henry Baird posed a question to the discussion group. Are there any techniques used in classical document analysis, that can be reliably applicable to Web Document Analysis, and how? The only techniques that we identified as reliable and easily applicable to the new problem were three: table recognition, logical layout analysis and maybe extraction of images (if we consider the final output as our only information).

Not having that many easily applicable techniques to Web Documents, it would be an advantage if could derive some structural information about the Web document beforehand. Although there is no standard format for all the Web documents, it seems that there are some basic categories of Web pages that share common characteristics as far as it concerns their layout and their content. The discussion group decided that it would be a benefit if we manage to identify different categories - like news sites or shopping sites - and find what common characteristics Web Documents of each category share.

At this point, we should stress the importance of ground truth data in Document Analysis. No successful Document Analysis can be conducted without having a measure to compare our results to, and unfortunately ground truth data for Web Documents do not exist, and it is rather difficult to be created, since we are not really sure yet about what kind of information these ground truth data should contain. The suggestion that came from Robert Haralick was that each one of the participants of the discussion group could try to create some ground truth data for a small number of Web Documents. The format of these ground truth data should not be strictly specified in advance, instead each one should make note of the things that he considers important for each of the Web Documents we use and ways to include them in the ground truth format specification. These first set of ground truth data can be circulated between the

members of this discussion group towards a first attempt to create a specification for Web Documents' ground truth data and possibly a small database. Daniel Lopresti raised the copyright issue, which is somewhat complicated due to the nature of WWW, but for the extent of this experiment we decided that this should not pose a problem.

3. Outcomes

What follows is a summary of the outcomes of this discussion, in terms of suggestions and decisions we made.

1. **Categorization of Web Documents.** The discussion group agreed that there are some categories of Web Documents that do share similar layouts and content. We decided to try to identify these categories and their unique characteristics and report our thoughts on-line for further discussion.
2. **Ground truth data collection.** After Robert Haralick's suggestion, we decided to start an effort to ground truth a few Web Documents with logical labels each, and pool the results into a first small database of ground truth data for Web Documents.
3. **Continuation of this discussion on-line.** We all found the discussion we had in the workshop fruitful, and unanimously agreed to continue the conversation on-line. Towards this direction Oliver Hitz offered to create an on-line forum, which is now available at http://www.egroups.com/group/web-das.
4. **Organization of activities.** It was agreed that the Document Analysis community has to come closer to the Web Content Extraction community. Towards this direction Robert Haralick proposed the following activities:
   - We will organize a workshop attached to ICDAR'2001. This will be aimed at bringing together researchers of both communities to discuss issues of common interest and discover areas of future collaboration. The workshop is going to have an educational function, with many tutorial presentations and will focus on discussion more than presentations. *Dr. Jianying Hu* from Avaya Labs, USA and *Dr. Apostolos Antonacopoulos* from the Univ. of Liveprool, UK will be the co-chairs of the workshop.
   - We will examine the possibility to co-organize a conference serving the same function together with NATO Advanced Studies Institute (ASI). We thought it good to have two co-chairs, one from Europe and one from USA. Currently *Dr. Apostolos Antonacopoulos* from the Univ. of Liverpool, UK offered to act as the chairperson from Europe, we still have not decided about the chair from USA.

The Web Document Analysis discussion group would like to invite anyone interested to join the ongoing discussion, and the activities of the group. Details about joining the discussion can be found at http://www.egroups.com/group/web-das.

4. List Of Participants

Participants at the DAS'2000 Web Document Analysis discuss group:

Antonacopoulos, Apostolos	aa@csc.liv.ac.uk	Univ. of Liverpool, UK
Baird, Henry	baird@parc.xerox.com	Xerox Palo Alto Research Center
Hitz, Oliver	oliver.hitz@unifr.ch	Univ. of Frebourg, Switzerland
Haralick, Robert	haralick@gc.cuny.edu	City University of New York
Hu, Jianying	jianhu@avaya.com	Avaya Labs Research
Ingold, Rolf	rolf.ingold@unifr.ch	Univ. of Frebourg, Switzerland
Karatzas, Dimosthenis	karatzas@csc.liv.ac.uk	Univ. of Liverpool, UK
Klink, Stefan	stefan.klink@dfki.de	German Research Center for Artificial Intelligence Germany
Lopresti, Dan	dlopresti@lucent.com	Lucent Technologies Bell Labs, US
Tan, Chew Lim	tancl@comp.nus.edu.sg	National Univ. of Singapore, Singapore
Tereschenko, Vadim	vadim_t@abbyy.ru	
Sugawara, Kazuhide	sugawara@jp.ibm.com	IBM Japan, Japan