

Open Source, Open Content, Social Aspects and Future of DAS

Moderator: Marcus Liwicki

Scribes: Phillip Pearson and Pingping Xiu

Participants: Adam Behringer,
Apostolos Antonacopoulos,
Bertin Klein, Daniel Lopresti, Henry Baird,
Hideaki Goto, Jim Fruchterman, Jim Yaghi,
Rolf Ingold and Sargur Srihari

Abstract. In this group, we discussed the trends of Document Analysis Systems. Mainly, we focused on the topics **Open Source, Open Content, Future Suggestions** and **The Revolution**. This report will summarize the main results of the discussion group.

1 Open Source

Existing open source document analysis software only focuses on small parts of the document analysis process. There are several open source OCR projects (GOCR¹, SOCR [2], Open Mind²), but they are incomplete and none perform particularly well. Everybody doing OCR research needs to have certain functions, so effort is wasted reimplementing these for each project.

What is holding back the development of good open source OCR? It is possible that much of the work is performed by students, who lose interest after completing their courses. A possible process for developing good open source OCR inside the academic community would be to split the work into small modules, each of which could be completed in a relatively short period of time, i.e., the time available to a student in a course. On a modularized OCR only small parts would have to be changed to apply it to a special recognition task.

2 Open Content

Popularity and accessibility are orthogonal concepts. Generally, researchers work with documents that have passed into the public domain, as there are no copyright issues to worry about. However, these documents are not necessarily representative of *all* documents. The group identified four different types of documents:

- Old/Closed: Historical documents: These are often held by libraries, which are reluctant to permit reproduction, in order to maintain their roles as sole providers of access to the documents.

¹ GOCR open source project - <http://jocr.sourceforge.net>

² Open Mind Project - <http://www.openmind.org>

- Old/Open: Accessible public domain documents, e.g. Project Gutenberg³: These are freely accessible.
- New/Closed: Commercial documents: These are popular but not accessible.
- New/Open: Wikipedia⁴ etc: These are freely accessible.

3 Suggestions for the Future

3.1 Context is Important - Frame Problem

For humans, context is an extremely important part of the recognition process. The information can be gathered easily if the context is known, even if the data is very noisy. However, it is hard to be used in OCR, because the algorithm frame is not fit for integrating various forms of context information. Future technology should pay more attention on the frame building. Maybe with some innovative format, context information can be stored in context database, which can be re-used for recognition.

3.2 Take Advantages from Speech Recognition

The speech recognition community has made good use of language modeling, both for speech recognition and for speech synthesis. They talk about recognition in restricted domains (e.g., broadcast news). As a result, they have been able to make what appears to be faster progress than we have in OCR and handwriting recognition. There is potential to further develop language modeling in our field. Some first steps into this direction have already been taken [1].

4 When will the OCR Revolution Happen?

What happened to the revolution? It is said that OCR is a "solved problem", however this is an exaggeration: OCR is only "solved" for clearly machine-printed English text, scanned at high resolution. There is still far to go before all OCR problems are solved. We talked about possible "components of the revolution":

- Anytime Algorithm

Currently, there is an emphasis on developing algorithms that run in as little time and use as little resources as possible. However, it is also sensible to develop algorithms that run over a long period of time and use large quantities of resources in order to provide better results. The concept of "anytime" algorithms is interesting. An anytime algorithm will run indefinitely, producing results which can be retrieved at any time, but steadily improve over time. It would be appropriate for anytime algorithms to also scale to use as much input data, memory and processing power as is available.

³ Projekt Gutenberg - <http://gutenberg.spiegel.de/>

⁴ Wikipedia - <http://www.wikipedia.org>

- Beyond recognition: Understanding
This seems to be the post-step of recognition, however, humans achieve recognition and understanding simultaneously. How can understanding be more tightly coupled with recognition? A few research groups already focus on this topic [3].
- Recognizer for all languages
Different languages require different segmentation algorithm. It is often difficult to adapt an algorithm for language A to work for language B . Is a universal segmentation model, like the popular bi-gram or tri-gram model in post-processing, possible?
- Device independent recognizer
Cameras are becoming ubiquitous, and are likely to become dominant for document image fetching, however the device-specific parameters limit the application scope of algorithms. Is there any algorithm that is robust against variation between devices? A device independent recognizer would be valuable.
- Hardware implementation
Human vision is a parallel system. Due to the parallel nature, the computation cost is tremendous. Future technology may use parallel hardware architectures to perform OCR tasks.

5 The Future of Paper Documents

In this topic, we reached the consensus that paper documents will be alive for a long time. In particular we spoke about the preferences of the community, the preservation, the use in specific domains and documents in general.

5.1 Preferences of the Community

Paper documents are the most direct and friendly interface for communication. Even when digital media is accessible everywhere and anytime, people may prefer paper documents for convenience of reading. There is a group of "digital natives" (people for whom digital media is very natural), but at the same time, companies are producing reproductions of very old books, for people who like how they feel. The Exbiblio system⁵ sets up a good example to augment but not replace paper documents, facilitating fast searching for scanned keywords from physical pages so that you can "click on" text in a book. In a word, paper documents will not disappear, but adapt its role to survive in a digital world.

5.2 Preservation

Paper can survive for thousands of years, whereas digital media starts to lose data within decades. So maybe in the future, though most paper documents may be converted to a digital form, the most important ones must be retained in hardcopy for safe preservation. Whenever hardcopies exist, the gap between digital media and physical copy exists, making Document Analysis necessary.

⁵ Exbiblio system - <http://www.exbiblio.com/>

5.3 Specific Domains

In specific domains or specific areas, paper documents are still needed for a long time. For example, when lawyers want to locate facts buried in a huge pile of paper documents, computer aid is urgently needed. In poor areas, such as developing countries, fewer people have access to digital media, so communication is mostly based on paper documents, which need Document Analysis Systems to improve efficiency.

5.4 Document Analysis

Considering the question “Will document analysis be useful in the future?”, it is important to understand that paper documents are only a subset of “all documents”. Document image analysis applies to all document images, which may never even have been printed on paper. Documents embody personal communication, so documents exist forever, and Document Analysis Systems exist and develop forever.

References

1. Marti, U.V., Bunke, H.: Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence* **15** (2001) 65–90
2. Peng, H., Gan, Q.: SOCR 1.03: a handwritten data form producing and reading system. (2000) 197–202
3. Toselli, A.H., Juan, A., Gonzalez, J., Salvador, I., Vidal, E., Casacuberta, F., Keysers, D., Ney, H.: Integrated handwriting and interpretation using finite-state models. *Int. Journal of Pattern Recognition and Artificial Intelligence* **18** (2004) 519–539