GREC 2013
10th IAPR International Workshop on
Graphics RECognition
Lehigh University, Bethlehem, PA

Aug 22 2013

# Post-GREC Meeting on Lehigh Mountaintop Data Collection

## Context

This meeting was held at the Lehigh University CSE Department, after a nearly 2hr visit of newly acquired office buildings that formerly had been part of Bethlehem Steel's research facilities. Many of the offices contain extensive collections of documents that Prof. Lopresti is in the process of retrieving and digitizing in order to preserve them to some extent and make them available as experimental "matter" and resource for document image analysis research.

## Attendees

(starred names are graduate students)

- Frédéric Baillard, Valconum, France
- Barri Bruno, Lehigh University, USA
- Jean-Christophe Burie, Université de La Rochelle, France (IAPR TC10 webmaster)
- Mickaël Coustaty, Université de La Rochelle, France
- Benjamin Duthil(*), Université de La Rochelle, France
- Lluis-Pere de las Heras(*), Computer Vision Center, Barcelona, Spain
- Bart Lamiroy, Université de Lorraine, France
- Josep Llados, Computer Vision Center, Barcelona, Spain
- Rafael Duire Lins, University of Pernambuco, Brazil (IAPR TC10 vice chair)
- Jean-Marc Ogier, Université de La Rochelle, France (IAPR TC10 chair)
- Christophe Rigaud(*), Université de La Rochelle, France
- Anjan Dutta(*), Computer Vision Center, Barcelona, Spain
- Klaus Broelemann(*), University of Muenster, Germany
- Nibal Nayef, Université de la Rochelle, France

## Overal Goal

The opportunity for the general document image analysis research community to have access to a freely available extensive, heterogenous and diverse data collection that also is sufficiently rich to contain coherent and consistent sub-collections is something that has been called for by the research community for a long time. The current document collection at Lehigh seems to be very useful to both the validation and development of document image analysis methods. However, the large quantity, diversity and complexity of the available data, combined with a hard deadline (Sep. 10 2013) at which the offices containing the document will be emptied of their contents for renovation work to be started requires some urgent preservation and contingency solutions to be conceived.

The overall goal of this meeting is to retrieve the opinion of part of the international research community on:

- how they perceive the interest of this data collection;
- what use they could imagine for it, once collected, curated and made available;
- how they could contribute to curate, digitize, promote or otherwise make the data collection available.

## Synoptic Account of Discussions

The discussions were lively and constructive and the following issues were raised:

- The data collection is potentially extremely useful and interesting on many accounts. It is something the community has been wanting for a long time. Curating, arranging and annotating the data is necessarily going to be a long and time-costly process, some of which can be done on the digitized documents, others may be requiring access to the physical documents, most of them will probably arise or be defined over time.
- It is questionable whether all collected data actually has a (market) value where either their contents or their physical supports are concerned. Although this may not necessarily impact their usefulness for research, it may make quite a difference when applying for grants and financing operations related to the collection.
- Related to the previous point, this is also unique opportunity for the research community to get involved in previously unsuspected uses of document collections. To that avail, it would be very interesting to get input from all other stakeholders to the document collection (librarians, archivists, historians) and their interest in specific document image analysis tools. Their context of use may also help identifying “valuable” documents and provide directions worthwhile investigating for funding purposes.
- The cost of scanning the “unconventional” size documents (blueprints, listings) is considerable and current available financial resources are almost inexistent. These documents definitely have a significant potential value for graphics-centered research. Given the cost related to their digitization, there is an higher need for curating these document than there is for more conventional sized documents, the digitization of which can be done more easily.
- The quality of the currently used scanning process is questioned. A first, quick and incomplete browsing though the digitized data seems to hint to JPEG compression, which is usually considered not an appropriate compression method for graphics image analysis.
- It is likely that digitizing costs are lower in Europe than in the USA.
- What are the Copyright issues we need to feel/be concerned with?
- This data collection is also a tremendous opportunity to create and setup a showcase of what the community is capable of doing in terms of document image analysis which could be used for marketing and reeling in industrial research contracts and collaborations.

  An interesting economic model was laid out for financing costs related to maintain and develop the dataset (*cf*. next section).
- The scanning process itself can potentially open to interesting image analysis techniques like alignment, color constancy, stitching … especially if multiple location sites, using different hardware or techniques end up being involved in the process.

## Conclusions and possible actions

The overall consensus is that, given the very short timespan left, the international community cannot (to its regret) contribute to anything tangible before Sep. 10. The main message to convey to Lehigh University would consist in trying to grab as many documents as possible and provide temporary storage (6 months, 1 year ?) until one of the following (or other solutions) can be worked out:

- Ship boxes to international partner labs to do part of the digitizing. This should be fairly cost and time effective, and share the efforts of digitizing among the community. This would mainly concern auto-feed compatible document formats.
- There are currently no obvious financing resources for handling large document formats (mainly graphical documents) although these documents have a potentially interesting role to play for the graphics recognition community.

One of the main requests here is *time*. It seems fairly clear that funding will come by bits and pieces, and resources made available progressively. Distributing the digitation over different locations, as previously mentioned, still is relevant, but may be temporary (longer term) storage would also be.

The key concern being financing, a couple of suggestions have been put forward.

1. If specific grant support is to be found, it is very likely it needs to be attached to some easily defined application context. Input from the other stakeholders can prove extremely valuable.
2. We may also try and look into self-financing by developing a business model using the available parts of the data collection for promoting and validating existing methods, that can be used for attracting private financing sources for specific task on their own data, in relation with comparable data in the data collection. Benefits can, in their turn, be partially used for digitizing the rest of the data collection, and iteratively get through all remaining available documents which can then also be used for promoting and validating other methods.
3. Call upon funding by IAPR. If we can come up with a well-defined scientifically sound activity that may promote TC-related activities and spur international collaboration or cross TC involvement, there may be a possibility to get funding. One idea that has been around for some time is to create a GREC "Grand Challenge", for instance.

- Making the data available and ensure long-term archiving also needs to be considered. Distributing the hosting over multiple locations/institutions may be a solution. Some recent advances and evolutions of the DAE platform using NoSQL storage may provide appropriate solutions.