DIAL 2004 Working Group Report on Acquisition Quality Control

Elisa H. Barney Smith[1], Henry Baird[2], William Barrett[3], Frank Le Bourgeois[4], Xiaofan Lin[5], George Nagy[6] and Steve Simske[5]

1 Boise State University, Boise, Idaho, USA

2 Lehigh University, Bethlehem, Pennsylvania, USA

3 Brigham Young University, Provo, Utah, USA

4 LIRIS I.N.S.A. de Lyon, Villeurbanne Cedex, France

5 Hewlett Packard Labs, Palo Alto, California, USA

6 Rensselaer Polytechnic Institute, Troy, New York, USA

Abstract

This report summarizes the discussions of the Working Group on Acquisition Quality at the International Workshop on Document Image Analysis for Libraries, Palo Alto, CA, 23-24 January 2004. Acquisition of the image is one of the most time intensive components of forming a digital library, and the quality of the acquisition will affect all later stages of the digital library project. The current state of the art in acquisition is analyzed. Problems and suggested improvements for image acquisition and storage formats and the special problems associated with acquisition from microfilm follows. A list of general suggestions was developed which was complemented by a wish list of things the Working Group would like to see followed in acquisition discussions in the future.

1. Introduction

Acquisition of the image is one of the more time intensive components of forming a digital library, and the quality of the acquisition will affect all later stages of the digital library project. Far too many people believe that "Quality is only display and printing" and that the quality of Optical Character Recognition (OCR) isn't significant. High quality acquisition will increase the ability to do high quality OCR. But even with degraded text images, under certain circumstances, OCR can be done with high success. Baird has shown that given knowledge of the typeface and a degradation model, he can automatically off line generate a classifier with extremely high recognition rates [1]. It has been shown that most of the OCR errors come from broken and touching characters [4]. Hence there is a need for good segmentation, or Baird-like models that compensate for such noise sources.

If the image is poorly acquired, there is a cascading effect of errors downstream. While that might seem obvious, there is a perception by some that linguistic models will fix all the problems of bad OCR so the input quality is not important [7]. But, as the error rate in OCR increases, the level of effort of contextual processes must increase, and eventually the effort level approaches infinity. This is the point at which the errors can no longer be compensated for.

Many digitization operations are a function of cost. According to Baird about one third of the cost of digitizing a book is cataloging, description, and indexing [2]. The second third is scanning, OCR, correction, and markup. The final third is quality control, file maintenance, and administration. This division however will change, and can affect the total cost, if any part is poorly done. A decision must be made on whether it is cost effective to do high quality scanning and analysis. The committee contemplated whether we can convince people it is worthy to spend the money on quality control. Good quality control will involve operators and necessary supervision, which makes it more expensive. Therefore we need to minimize user interaction. During acquisition, the quality must be increased and the variability must be decreased. Currently the most common method is to eyeball the document to check acquisition quality. This leads to both a high level of variability and an increased cost from user interaction.

There are several places where the Document Image Analysis community could contribute tools. It would be useful to count page numbers during acquisition to automatically warn operators if pages are missing. Many times if the data is improperly acquired, it can be adjusted or filtered to compensate for some of the degradations. Under certain circumstances, the degradation results in a loss that is not recoverable. Some examples of this include misalignment of a page within the image acquisition window such that part of the document is not captured, as shown in Figure 1a. Irrecoverable loss of data will also happen if the range of intensity values falls outside the sensor range as shown in Figure 1b. In circumstances like

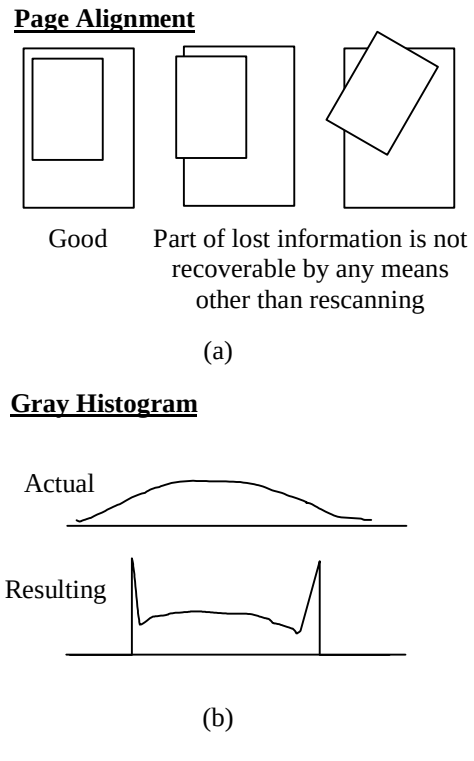

Figure 1: Examples of where information is irretrievably lost during acquisition.

these the system needs a red light to alert the operator when the acquisition is below par, similar to the low oil light on an automobile. Can we not do more with auto-calibration, as is done routinely with desktop scanners? Of course, sometimes this means two passes, but that could be factored into the economics and weighed against the cost of unrecoverable losses.

2. Image Acquisition and Storage Formats

When doing acquisition, the project leaders must make decisions on whether to acquire in bi-level, gray scale or color. They must also decide on the resolution at which to make the acquisition. Then once the images are acquired, and processed, the fate of the original scans must be decided. Each of these decisions will affect the ability to do processing for the current project as well as to take advantage of future growth in technology. Since a large portion of the cost of a digital library project is in the acquisition, these decisions should not be taken lightly.

With gray scale, acquisition can be done at a lower resolution and still get the same OCR accuracy as acquisition in bi-level. In the 1980's Pavlidis suggested that grey-depth could be traded off for spatial sampling rate [5]. In an independent study Barrett and Barzee [3] showed that results comparable to 300-400 dpi bitonal could be achieved with 200 dpi grayscale. Based on the current state of the art of OCR, over the next 10 years for projects where scanning is being done just for content, and not for historical fiber analysis, it is felt that scanning with 400dpi color will be sufficient. Given these recommendations, one should recall there is a difference between acquiring at 100dpi and interpolating 400dpi and acquiring initially at 400dpi optical resolution. Also for such specifications to be valid, the acquisition has to be carefully calibrated. However, many acquisition project directors put a size limit on acquisition. What quality control can we suggest to best meet this?

Another issue with digitally stored images and content extracted text is the longevity of file formats. While storage of data in digital form may be a method to make data more widely available, because of the rapid changes in storage media and formats, it may not be the best way to ensure longevity. However, storage of document images in digital form does not preclude keeping the original paper document. The feeling of this committee was that in all major digitization efforts the original scan should always be saved, because later technology may be available to better process the data for better results, or for an as yet un-thought of purpose. Essentially bits on disk are free, but human labor has a cost. However, often the direction from management is to save only a processed image. If users don't save the original scan, what amount of filtering and reduction is acceptable?

With the constraints imposed on acquisition by the project directors, there will be consequences down the road for the digitization project. This is a place where the Document Image Analysis (DIA) community can assist by clearly documenting for the non-researcher why our recommendations have been made, and listing the consequences when deviations from these suggestions are made. Another large decision was that the DIA community should open the door to hardware designers. Currently there are no benchmarks for designers (like old CCITT images). The copier repair industry has all sorts of quality control test targets and specifications. Do we need something like JBIG2 for documents?

3. Microfilm

Document image acquisition with home or commercial desk top scanners has a high degree of uniformity built into it by the scanner manufacturer. The lighting is designed to be uniform across the entire acquisition region. Most software that accompanies these units will automatically select brightness and contrast values to increase quality.

Acquisition with microfilm scanners is a bigger

problem. Auto-calibration software also needs to be integrated into microfilm scanners to lend efficiency to the scanning of large collections. Two-pass calibration can certainly be done as with conventional flat-bed scanners, but this would have to be weighed against the cost of loss of data as described in Figure 1. Currently no calibration methods come with microfilm scanners, and this is a serious impediment to efficient scanning of large collections.

OCR from microfilm presents a great problem. Microfilm itself gets scratched very easily and the image quality on the film degrades with time requiring the film to be copied every few years. After a certain number of years, under certain environmental conditions, the old acetate films begin to smell of vinegar at which point deterioration accelerates. Charts exist for acetate film that explain the relationship between temperature, relative humidity (RH), and the "vinegar syndrome" (the slow chemical decomposition of acetate plastics leading to loss of their value in a film collection). For example, if a fresh acetate film is kept in an air-conditioned room where it is comfortable for people: 70°F (21°C), with an RH of 50%, and these conditions are maintained year round (not always an easy thing to do), the film will last for 40 years before the film hits the "vinegar syndrome." While the film is still useful at this point, deterioration begins to accelerate, putting it on the "hit list" for copying. Under ideal conditions, films may last 100's of years. Under less ideal conditions, it may take only a few years to a few 10's of years for the "vinegar syndrome" to emerge. Paper documents also deteriorate in storage, but not rapidly, so a different mindset is needed when working with microfilm.

Acquisition from microfilm is either done by parsing the film into individual images as it is acquired using interactive cropping software or the entire roll is read in to one large image file containing the image from the start of the roll to the end. This then needs to be segmented into pages. The light is usually set on the first page and then maintained for scanning of all following pages. The images stored on the film often have inconsistent lighting page to page, and within a given page, because the lighting used during acquisition is rarely uniform. Getting the right level and uniformity of lighting is difficult. Often the lighting is extremely poor, demonstrating all the more the need for adaptable calibration procedures.

4. General Suggestions

The committee through its discussions did come to agreement on a few general guidelines that can make an immediate impact on acquisition image quality. There is a need to develop and distribute test targets and software to analyze the signal to noise ratio (SNR) and the drift of calibration. A test chart should be scanned at the start of an acquisition process and every N-times to track calibration drift. For some processes this has been developed and is available. For instance, Hewlett Packard has introduced multi-pass quality assurance in their Digital Content Re-Mastering Project [6]. A Modulation Transfer Function (MTF) test pattern is used at the start of their work. The first several passes are fully automatic. Human operators will get involved in the final pass on less than 1% of the pages. Barney Smith has used several test targets to analyze the calibration of a scanner when doing bi-level acquisition and has developed methods to estimate the scanner calibration from textual documents when the scanner is no longer present [8].

5. Wish list

So while the committee recognizes that there is quite a bit of technology already out there to create high quality digital libraries, there are a few places where either technology or human willingness falls short. The working group ended its discussion by creating a 'wish list' of the top things we would like to see either developed or implemented in the near future for digital library acquisition projects:

1. Development of automatic tools for test target analysis and measurement of MTF, SNR as part of all digitizers.
2. All digitizers should have calibration of density/optical range. Since microfilm scanners don't include this, we would like to see development of the equivalent technology for microfilm.
3. All acquisition of machine printed documents should be done at 400dpi color (minimum). Some of the fine print that exists on Census records, for example may require scanning at a resolution of 500 dpi, or greater, if we are to have any chance at successful OCR.
4. Preserve the original scanned image, without the clean-up and post-processing improvements.
5. Minimize user interaction. This can increase quality and decrease variability. Put human intervention at critical points in the system process. The system needs to do more of the bookkeeping. With this, extraordinary things can be done by ordinary people.

6. References:

[1] Henry S. Baird, "Document Image Defect Models," *Proc. IAPR Workshop on Syntactic and Structural Pattern Recognition*, Murry Hill, NJ, June 1990, pp. 13-15. Reprinted in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer Verlag: New York, 1992, pp. 546-556.

[2] Henry S. Baird, “Difficult and Urgent Open Problems in Document Image Analysis for Libraries,” Proc. Document Image Analysis for Libraries, Palo Alto, CA, 23-24 January 2004, pp. 25-32.
[3] William Barrett and Rex Barzee, “Posting Paper on the Web,” Proc. Vision Interface ‘98, pp. 381-388.
[4] Luis R. Blando, Junichi Kanai, Thomas Nartker, “Prediction of OCR Accuracy Using Simple Features,” Proc. of the Third International Confernce on Document Analysis and Recognition, Montreal, Canada, 14-16 August 1995, pp. 319-322.
[5] Theo Pavlidis, “Effects of Distortions on the Recognition Rate of a Structural OCR System,” Proc. IEEE Computer Vision and Pattern Recogn. Conf. Washington, D.C., June 21-23, 1983, pp. 303-309.
[6] Steven Simske and Xiaofan Lin, “Creating Digital Libraries: Content Generation and Re-Mastering,” Proc. Document Image Analysis for Libraries, Palo Alto, CA, 23-24 January 2004, pp. 33-45.
[7] Kazem Taghva, J. Borsack, and A. Condit, “Evaluation of model-based retrieval effectiveness with OCR text,” ACM Transactions on Information Systems, vol 14, pp. 64-93, January 1996.
[8] HokSum Yam and Elisa H. Barney Smith, “Estimating Degradation Model Parameters from Character Images,” Proc. International Conference on Document Analysis and Recognition 2003, Edinburgh, Scotland, 3-6 August 2003, pp. 710-714.