Quality Assurance Tools for Digital Repositories

Roman Graf
Ross King
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
{roman.graf,ross.king}@ait.ac.at

ABSTRACT

Digitization workflows for automatic acquisition of image collections are susceptible to errors and require quality assurance. This paper presents a quality assurance tool suite for long term preservation. These tools support decision making for blank pages, cropping errors, mistakenly appearing fingers in scans and accurate duplicate detection in document image collections. The important contribution of this work is a definition of the quality assurance workflow and its automatic computation. The goal is to create a reliable tool suite that is based on image processing techniques.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System issues

General Terms

preservation strategies and workflows

Keywords

digital preservation, quality assurance, image processing, information integration

1. INTRODUCTION

Within the last decade, significant effort has been invested in digitisation projects. Many large-scale digitization projects are running in digital libraries and archives and in public-private partnerships between cultural heritage institutions and industrial partners. The overall production in these projects has reached a level where a comprehensive manual audit of image quality of all digitized material would be neither feasible nor affordable. Nevertheless, cultural heritage institutions are facing the challenge of assuring adequate quality of document image collections that may comprise millions of books, newspapers and journals with hundreds of documents in each book. Quality assurance tools that aid the detection of possible quality issues are required. The material used in our experimental setup has been digitized in the context of Austrian Books Online, a public private partnership of the Austrian National Library with Google. In this partnership the Austrian National Library digitises and puts online its historical book holdings ranging from the 16th to 19th century with a scope of 600,000 books (see [1]). The project includes aspects ranging from digitisation preparation and logistics to quality assurance and online-access of the digitized items. Especially the quality assurance presents a challenge where automatic and semi-automatic tools are required to facilitate the quality assurance processes for the vast range and amount of material (described in [2]). The main contribution of this paper is the development of a DIGLIB QA Suite for the analysis of digital document collections and for reasoning about analysed data.

2. QUALITY ASSURANCE TOOLS

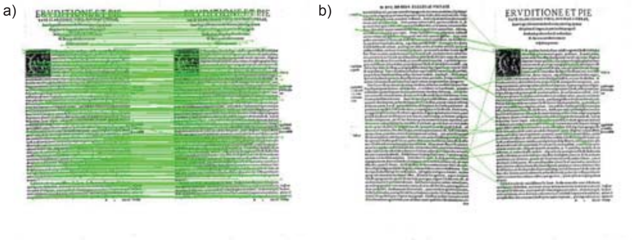

Figure 1: Samples of evaluation results from book identifier 151694702 (Austrian National Library) for duplicate detection with SIFT feature matching approach: (a) similar pages with 419 matches, (b) different pages with 19 matches.

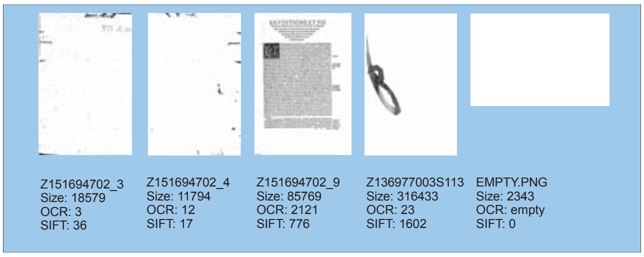

Figure 2: Selected samples of blank pages in digital collections from different sources with associated file name, file size, OCR and scale-invariant feature transform (SIFT) analysis result.

The suite includes four tools. The *matchbox* tool [3] for accurate duplicate detection in document image collections is a modern quality analysis tool based on Scale-Invariant Feature Transform (SIFT) feature extraction (see Figure 1). The *blank page detection* tool [4] that employs different image processing techniques and Optical Character Recognition (OCR) (see Figure 2). The *finger detection* tool [5] for automatic detection of fingers that mistakenly appear in scans from digitized image collections. This tool uses modern image processing techniques for edge detection, local image information extraction and its analysis for reasoning on scan quality (see Figure 3). The *cropping error detection* tool supports the analysis of digital collections (e.g. JPG, PNG files) for detecting common cropping problems such as text shifted to the edge of the image, unwanted page borders, or unwanted text from a previous page on the image (see Figure 4).

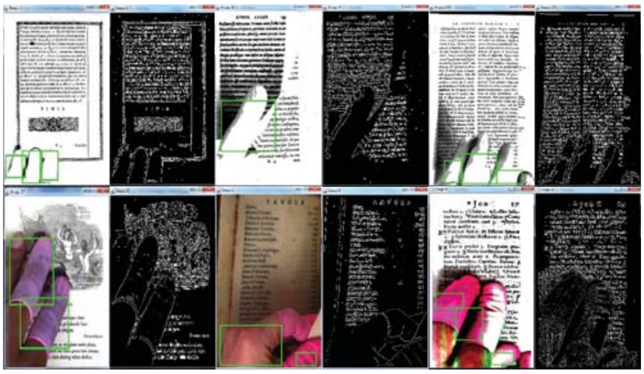

Figure 3: Positive detections of fingers on scans with associated edges where suspected areas are marked by green rectangles.

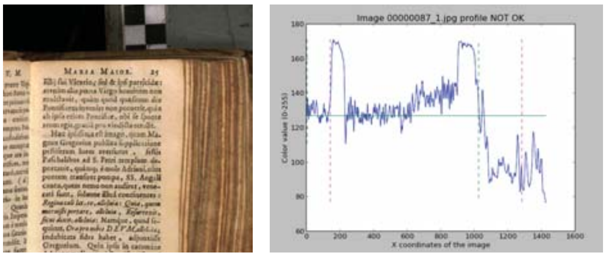

Figure 4: Cropping detection sample.

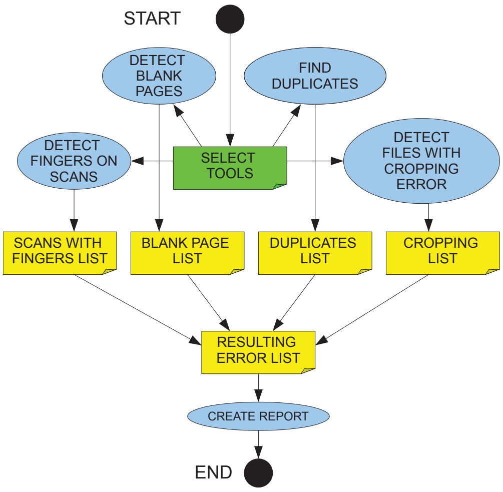

Figure 5: The workflow for the DIGLIB QA tool suite.

3. THE ERROR DETECTION PROCESS

The presented tools cover multiple error scenarios. The main use case for *matchbox* tool is a detection of the duplicated documents. Blank pages in a collection may address failure in a scanning process. Fingers should not be visible on the scans. The use cases for cropping errors are: text shifted to the edge of the image; unwanted page borders and unwanted text from the previous page on the image. Figure 5 presents the quality analysis workflow that employs different image processing tools for detection of errors and inaccuracies in digital document collections. This workflow includes the acquisition of local and global image descriptors, its analysis and an aggregation of resulting data in a single report for collection. The metadata and the selection criteria of digitization for preservation should be defined by an institutional expert for digital preservation. Selection criteria are dependent on particular collection types. Evaluation took place on an Intel Core i73520M 2.66GHz computer using Java 6.0 and Python 2.7 languages on Windows OS. The Relative Operating Characteristic (ROC) values for duplicate detection, cropping errors, blank pages and fingers on scans detection are represented by (0.013, 0.7), (0.001, 0.666), (0.007, 1.0) and (0.04, 0.844) points respectively. All these points are located very close to the so called perfect classification point (0, 1).

4. CONCLUSIONS

In this work we presented an approach for bringing together information automatically aggregated from different quality assurance tools regarding possible errors or inaccuracies in digital collection. The quality assurance tools for digital collections can help to ensure the quality of digitized collections and support managers of libraries and archives with regard to long-term digital preservation. As future work we plan to perform a statistical analysis of the automatically extracted information from the quality assurance tool and the qualitative analysis of the aggregated knowledge.

5. ACKNOWLEDGMENTS

This work was partially supported in part by the EU FP7 Project SCAPE (GA#270137) `www.scape-project.eu`.

6. REFERENCES

[1] M. Kaiser, "Putting 600,000 Books Online: The Large-Scale Digitisation Partnership between the Austrian National Library and Google," *Liber Quarterly*, vol. 21, pp. 213–225, 2012.

[2] M. Kaiser and S. Majewski, "Austrian Books Online: Die Public Private Partnership Der Österreichischen Nationalbibliothek Mit Google," *Bibliothek Forschung und Praxis*, vol. 37, pp. 197–208, 2013.

[3] R. Huber-Mörk and A. Schindler, "Quality assurance for document image collections in digital preservation," in *Proc. of the 14th Intl. Conf. on ACIVS*, LNCS, (Brno, Czech Republic), Springer, September 4-7 2012.

[4] R. Graf, R. King, and S. Schlarb, "Blank page and duplicate detection for quality assurance of document image collections," *APA CDAC 2014*, vol. 2014, pp. 87–97, February 5-6 2014.

[5] R. Graf and R. King, "Finger detection for quality assurance of digitized image collections," *Archiving conference*, vol. 2013, pp. 122–125, 2013.