Tool Suite for Quality Assurance of Digital Repositories

Roman Graf and Ross King
AIT Austrian Institute of Technology GmbH, Safety & Security Department, Vienna, Austria.
Roman.Graf@ait.ac.at and Ross.King@ait.ac.at

INTRODUCTION

Digitization workflows for automatic acquisition of image collections are susceptible to errors. Large document image collections in digital libraries require automatic quality assurance.

Fig 1. Evaluation results samples 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.

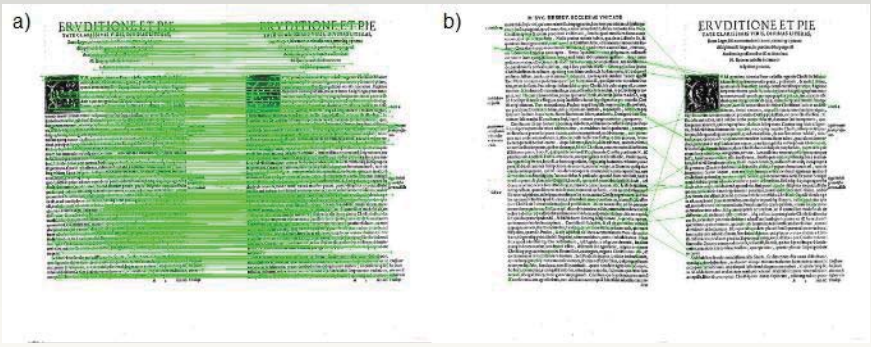


Fig 2. Selected samples of blank pages in digital collections from different sources with associated file name, file size, OCR and Scale-Invariant Feature Transform (SIFT) analysis result.

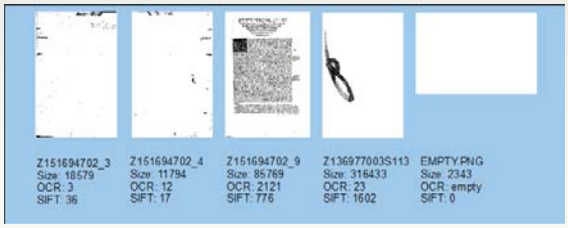


Fig 3. Positive detections of fingers on scans with associated edges where suspected areas are marked by green rectangles.

TOOLS

1.Matchbox tool [1] for accurate near duplicate detection in document image collections. A modern quality analysis tool based on SIFT feature extraction (see Figure 1).
2.Blank page detection tool [2] that employs different image processing techniques and OCR (see Figure 2).
3.Finger detection tool [3] 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).
4.Cropping error detection tool that 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).

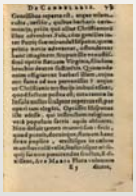
SOURCE CODE

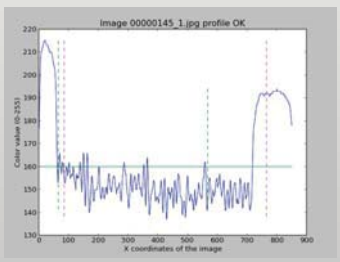
1.Matchbox tool [http://openplanets.github.io/matchbox/]
2.Finger detection tool [http://openplanets.github.io/finger-detection-tool/]
3.Cropping error detection tool [http://openplanets.github.io/crop-detection-tool/]

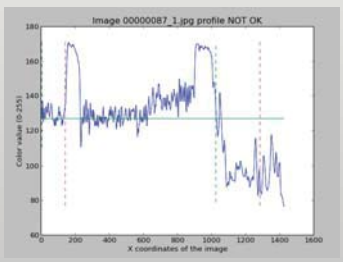
Fig 4. Cropping detection samples

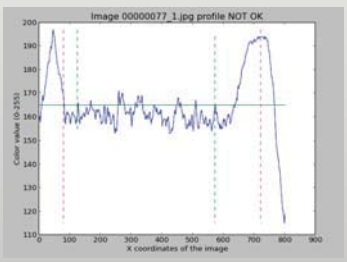
Use cases of cropping errors

- Text shifted to the edge of the image
- Unwanted page borders
- Unwanted text from previous page on the image







CONCLUSION

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.

REFERENCES

[1] Huber-Mörk, R., and Schindler, A. 2012. Quality assurance for document image collections in digital preservation. Proc. of the 14th Intl. Conf. on ACIVS (ACIVS 2012). LNCS, vol. 7517, pp. 108–119. Springer, Brno, Czech Republic.
[2] R. Graf, R. King, and S. Schlarb. Blank page and duplicate detection for quality assurance of document image collections. APA CDAC 2014, 2014:87-97, February 5-6 2014.
[3] Graf, R., and King, R. 2013. Finger Detection for Quality Assurance of Digitized Image Collections. Archiving conference. Volume 2013, April. 2013, 122-125.

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).