Qualitative and Quantitative Methods in Libraries (QQML) 2:551 –559, 2014

Quality Assurance Method for Cropping Error Detection in Digital Repositories

Roman Graf[1] and Ross King[1] and Stefan Majewski[2]

[1]AIT Austrian Institute of Technology, Vienna, Austria
[2]Austrian National Library, Vienna, Austria

Abstract: One of the common challenges in the mass-digitisation of book collections is correctly cropping (removing unnecessary border material from the digital image) during the automated image post-processing. This paper presents a method that supports the analysis of digital collections (e.g. JPG 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. One contribution of this work is a definition of the evaluation use cases for cropping problems. A second contribution is the creation of a reliable expert tool for document cropping error detection based on image profiling techniques. This tool can be applied in quality assurance workflows for digital book collections. Our suggested method employs evaluation parameters that can be defined for each book. The tool works independently of the image size, format and colour. We have analysed two real world collections with correct and corrupted images, and our tool has demonstrated good recall and precision for both corrupted image and correct images.

Keywords: digital library, quality assurance, cropping

1. Introduction

Within the last decade, significant effort has been invested in digitisation projects in libraries. 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.

Received: 30.3.2014 / Accepted: 12.5.2014 ISSN 2241-1925

The material used in our experimental setup has been digitized in the context of *Austrian Books Online* (see Austrian National Library (2014)), 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 (cf. Kaiser (2012)). 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 Kaiser & Majewski (2013)).
The main contribution of this paper is the development of a cropping detection tool for the analysis of digital document collections and for reasoning about analysed data. The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the cropping detection process and also covers image processing issues. Section 4 presents the experimental setup, applied methods and results. Section 5 concludes the paper and gives outlook on planned future work.

2. Related Work

Image processing techniques can be employed for quality assurance of digital content by replacing of a human expert regarding the decision-making process in a particular domain. The CROPDET (cropping detection) approach of cropping error detection employs image profile computation. The Meyer (1992) algorithm uses a colour profile in form of its three component profiles for image segmentation based on the watershed transform method. Our approach does not use the RGB image directly but transforms it into a greyscale image and subsequently creates a colour profile. In the context of digital preservation computer vision techniques are employed in different scenarios. Strodl et al (2007) present the Planets (Schlarb et al (2010)) preservation planning methodology by an empirical evaluation of image scenarios. They demonstrate specific cases of recommendations for image content in four major National Libraries in Europe.

3. Cropping detection process

CROPDET is a cropping detection tool for quality assurance in document image collections. One of the frequently encountered problems in digitised book collections is incorrect cropping during the automated scan process. In mass digitisation the master-images are usually slightly bigger than the digitized original media. During post-processing the correct cropping to the page size must be determined and applied. In most cases automated methods yield the expected and correct result. But, as the processing is performed in batches, a method is needed to identify potentially mis-cropped pages. To address this, the proposed method supports the analysis of digital collections (e.g. JPG files) for cropping problems during the scanning and post-processing. In order to detect the most frequent cropping problems we regard three use cases. The first use

case detects a mis-cropped page where, due to the cropping, a text that is shifted to the right border of the page image and therefore the left border is much wider than right border. In the second use case, the cropping is so close to the text that there is no gap remaining between text edge and image border. This case is particularly important as it potentially indicates possible text loss as the text may be cut by the cropping. In the third scenario, an image comprises part of the text from the adjacent page.

Image-profile-based image cropping detection employs evaluation parameters that can be defined for each book. The tool works independent of the image size and colour. This is particularly important as the evaluation is based on assumptions of commonly used printing and layout practices. That is, usually a text block in a book is surrounded by margins, which are governed by particular proportions. One important parameter is the border width, which describes an expected distance between page border and text edge on X axis and is employed for border analysis. The second important parameter is the relation between left and right border widths of the image. Usually these are of different width due to the binding at the spine.

Collection analysis is conducted according to the quality assurance workflow shown in Figure 1. The user triggers a complete collection analysis, the results of which are stored in a text file. In order to detect documents with cropping error we aggregate digital text document (see Figures 2-4) specific expert knowledge and analyse images using image profiling technique demonstrated in Figures 5-7. In the first workflow step the image is loaded in RGB colour format. In the next step RGB image is converted to the greyscale format using the perceptually weighted formula (see Guojun (1998)). We used formula (1):

$$\mathbf{Pres = 0{,}299 \times Pred + 0{,}587 \times Pgreen + 0{,}114 \times Pblue} \quad \mathbf{(1)}$$

Where Pres is a pixel value in a greyscale image, Pred, Pgreen and Pblue are pixel values in an RGB image.

In subsequent steps we analyse greyscale image and calculate left and right border distances in order to apply different parameters like minimal and maximal border distance, left to right border relation and average colour (Pres value) for the image. Additionally the average colour for the whole collection can be calculated and used as a parameter in the analysis. The evaluated image profiles (see Figures 5-7) demonstrate calculated values visually. This supports the human expert to infer an informed evaluation of the cropping quality of image candidates that could be mis-cropped and evaluate whether these images are in fact affected by the indicated error and to perform necessary actions.

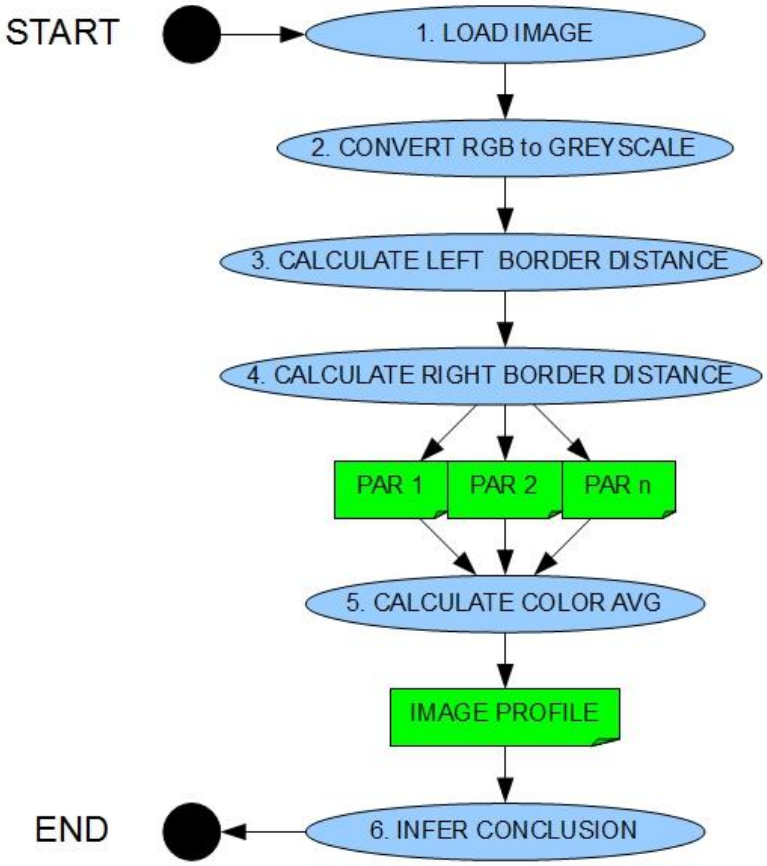


Figure 1. Workflow for cropping error detection

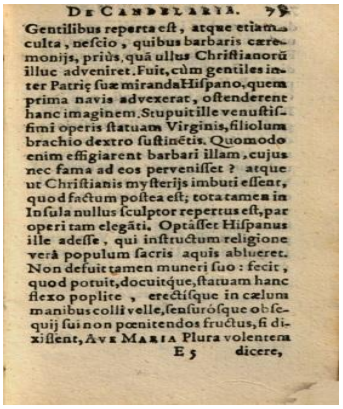
The presented tool is working on a greyscale representation of the image data and is configured with relative measures. The initial configurations should be set by a digital preservation expert who is familiar with the material of a particular institution and the type of document collection.

3. Evaluation

Our hypothesis is that image profiling could help to detect cropping errors in document collection. We have analysed two test collections. The documents from the first test collection, with two correct images and five corrupted images, are described in detail in this section. The second collection comprises a selection of 40 images including a variety of the defined cases of cropping errors. The analysis of the second collection marked 13 images as correct and 27

images as having cropping error. The ground truth data created by human expert confirms this result. Our tool correctly detected all corrupted images as corrupted and correct images as correct. Of course, the accuracy depends on the expert parameter settings and for larger collection cannot be 100 percent. Nevertheless, for standard text-pages the algorithm is expected to yield good results. See samples for correct and corrupted images in Figures 2-4 with associated analysis results in Figures 5-7.

The evaluation has been performed on an Intel Core i73520M 2.66GHz computer using Python 2.7 language on Windows OS. We evaluate images with corrupted cropping and calculation accuracy for each image. Images were analysed for previously defined cropping use cases like text shifted to the edge of the image, unwanted page borders, or unwanted text from previous page on the image.


DE CANDELARIA. 75

Gentilibus reperta eſt, atque etiam culta, neſcio, quibus barbaris cæremonijs, priùs, quã ullus Chriſtianorũ illuc adveniret. Fuit, cùm gentiles inter Patriç ſuæ miranda Hiſpano, quem prima navis advexerat, oſtenderent hanc imaginem. Stupuit ille venuſtiſſimi operis ſtatuam Virginis, filiolum brachio dextro ſuſtinẽtis. Quomodo enim effigiarent barbari illam, cujus nec fama ad eos perveniſſet? atque ut Chriſtianis myſterijs imbuti eſſent, quod factum poſtea eſt; tota tamen in Inſula nullus ſculptor repertus eſt, par operi tam elegãti. Optãſſet Hiſpanus ille adeſſe, qui inſtructum religione verâ populum ſacris aquis ablueret. Non defuit tamen muneri ſuo: fecit, quod potuit, docuitque, ſtatuam hanc flexo poplite, erectiſque in cælum manibus colli velle, ſenſuróſque obſequij ſui non pœnitendos fructus, ſi dixiſſent, AVE MARIA Plura volentem

E 5 dicere,

Figure 2. A correct image 000000145_1.jpg from Austrian National Library collection with associated profile


MARIA MAIOR. 25

Filij ſui Vicario; ſed & ipſi parricidæ: etenim alia pœna Virgo hominem non mulctavit, quàm quòd quæſitum diu Pontificem invenire non potuerit, quin ab ipſo etiam Pontifice, ubi ſe ſponte reum egit, gratiã pro vindicta retulit.

Hæc ipſiſſima eſt imago, quam Magnus Gregorius publica ſupplicatione pelliferam luem averſurus, feſtis Paſchalibus ad S. Petri templum deportavit, quámq; è mole Adriani, cùm pontem tranſiret pompa, SS. Angeli cantu, quem nemo non audiret, venerati ſunt, ſolenne illud concinentes: *Regina cæli lætare, alleluia: Quia, quem meruiſti portare, alleluia, Reſurrexit, ſicut dixit, alleluia:* Namque, quod ſequitur, *Ora pro nobis DEVM, alleluia,* indubitata fides habet, adjunxiſſe Gregorium. Quin ipſo in cacumine

Figure 3. A corrupted image 00000087_1.jpg from Austrian National Library collection

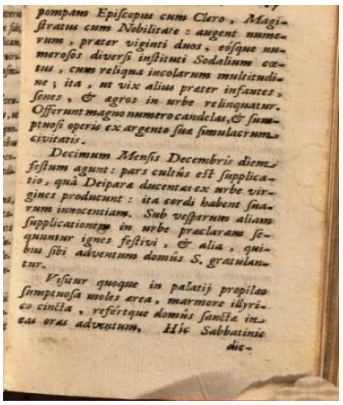


Figure 4. A corrupted image 00000077_1.jpg from Austrian National Library collection

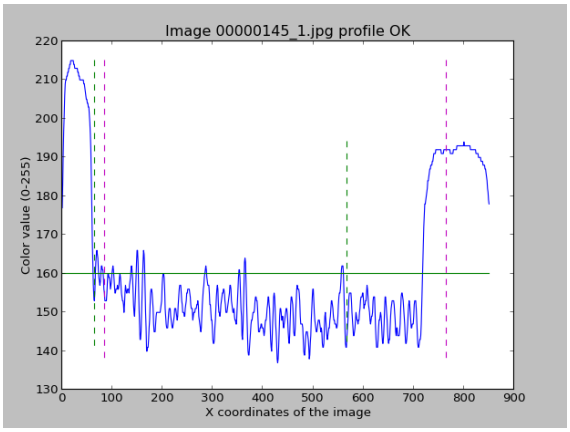


Figure 5. A correct image 000000145_1.jpg profile

Figure 5 presents the image profile for document 00000145_1.jpg. This document has correct cropping for both borders. The left border distance from border to text edge is about 60 pixels and the larger right border distance is about 270 pixels from the total image width of 900 pixels. The average colour (`Pres` value) for this document is 160 on an axis from 0 to 255, where 0 is completely black and 255 is completely white. Figure 5 demonstrates that both borders are white and have correct border relation significant for text document. The text space from 60 to 730 on the X axis is dark what is correct and expected for pages printed in black font.

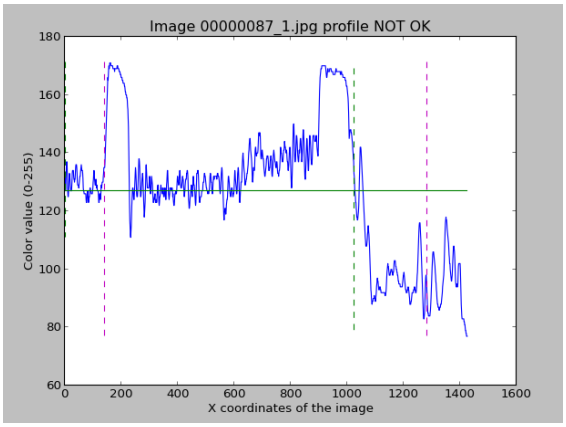


Figure 6. A corrupted image 00000087_1.jpg profile

Figure 6 presents the image profile for document 00000087_1.jpg. This document has corrupted cropping on both borders. The left border includes unwanted text from the adjacent previous page. This is represented in the diagram between pixels 0 to 120 on the X axis. Then we can observe a left border to text edge from 120 to 270 pixels. On the right side, the image includes the unwanted representation of the edges of the following pages of the book. The width of this unwanted edge reaches from 1040 to 1450 pixels on the X axis. The right border distance for the given page is about 160 pixels from the total image width of 1450 pixels and is located between 880 and 1040. The average colour value for this document is 126. The profile in Figure 6 shows that both borders have a cropping error and this digitised page should be considered as corrupted.

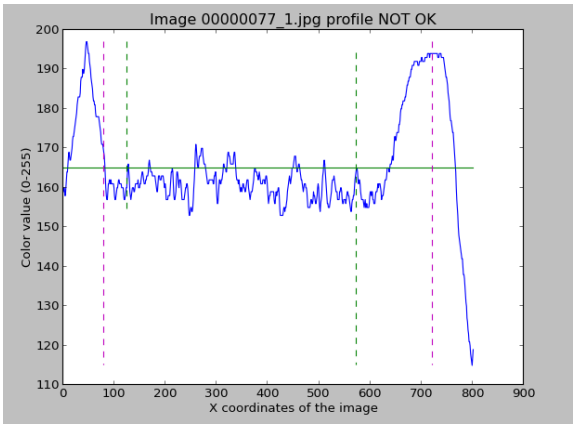


Figure 7. A corrupted image 00000077_1.jpg profile

Figure 7 shows the image profile for document 00000077_1.jpg. This document has corrupted cropping for the left border. The left border includes unwanted text from previous page what is depicted by distance from 0 to 85 pixels on X axis, which leads to the sharp shape on the plot in contrast to the steep rise and fall of colour values observed in correctly cut images such as in Figure 5. This shape demonstrates that document includes text left of the page margin which is an indicator for text from the adjacent left page. On the right side the image is correct. The correct right border distance for the given page is about 170 pixels from the total image width of 820 pixels and is located between 650 and 820. The average colour value for this document is 165. Figure 7 depicts that left border has a cropping error and this text document should be considered as corrupted.

Therefore, the image profiles have corroborated our initial hypothesis that image profiling could be a useful method for detecting cropping errors in document collections.

4. Conclusions

We have developed the CROPDET QA tool for cropping error detection in document image collection handling. This tool detects images in a document collection where cropping defined during the automated scan and image post-processing is incorrect and requires the removal of unnecessary border material from the digital image. The presented approach supports the analysis of digital collections (e.g. JPG, TIFF, PNG files) for cropping problems e.g. text shifted to the edge of the image, unwanted page borders, or unwanted text from adjacent page on the image. Important contribution of this work is a definition of the evaluation use cases for cropping problems. Another contribution is the creation of a reliable expert tool for document cropping detection based on image profiling techniques. Our suggested method employs evaluation parameters that are customizable and can be defined for each collection by institutional expert of digital preservation. The tool works independently of the image size, format and colour. We have analysed two real world collections with correct and corrupted images. Our tool has demonstrated good recall and precision for both corrupted image and correct images.

5. Acknowledgments

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)

References

Guojun Lu; Phillips, J., (1998). Using perceptually weighted histograms for colour-based image retrieval, *Signal Processing Proceedings, 1998. ICSP '98. 1998 Fourth International Conference on Signal Processing*, vol.2, no., pp.1150 - 1153, 1998 doi: 10.1109/ICOSP.1998.770820

Meyer, F., (1992). Color image segmentation, *Image Processing and its Applications, 1992., International Conference on* , vol., no., pp.303,306, 1992

Strodl, S., Becker, C., Neumayer, R., Rauber, A., (2007). How to choose a digital preservation strategy: evaluating a preservation planning procedure. *In: JCDL '07: Proceedings of the 2007 conference on digital libraries.* pp. 29–38. ACM, New York, NY, USA (2007)

Schlarb, S., Michaelar, E., Kaiser, M., Lindley, A., Aitken, B., Ross, S., Jackson, A., (2010). A case study on performing a complex file-format migration experiment using the planets testbed. *IS&T Archiving Conference 7*, 58–63 (2010)

Austrian National Library. Austrian Books Online. http://www.onb.ac.at/austrianbooksonline (accessed 20.3.2014)

Kaiser, M., 'Putting 600,000 Books Online: The Large-Scale Digitisation Partnership between the Austrian National Library and Google', *Liber Quarterly*, 21 (2012), 213–25

Kaiser, Max, and Stefan Majewski, 'Austrian Books Online: Die Public Private Partnership Der Österreichischen Nationalbibliothek Mit Google', *Bibliothek Forschung und Praxis*, 37 (2013), 197–208