Qualitative and Quantitative Methods in Libraries (QQML) 4: 811-–820, 2015

Quality Assessment Method for Warping and Cropping Error Detection in Digital Repositories

Roman Graf and Ross King and Martin Suda

AIT Austrian Institute of Technology, Vienna, Austria

Abstract. Two of the common challenges in the mass-digitisation of book collections are correctly cropping (removing unnecessary border from the digital image) and detection of the documents, which were warped during the automated scanning process. This paper presents a method that supports the analysis of digital collections (e.g. JPG files) for detecting common problems such as text shifted to the edge of the image, unwanted page borders, rotated text, unwanted text from a previous page on the image, or error detection in situations when the document is physically or optically warped. One contribution of this work is a definition of evaluation use cases assessing the extent of warping and cropping problems. A second contribution is the creation of a reliable expert tool for document warping and cropping error detection based on image processing techniques. This tool can be applied in quality assessment workflows for digital book collections. Our suggested method employs evaluation parameters that can be defined for each book. By means of these parameters a preservation expert can express institutional preferences for collection analysis. We are targeting the text position annotation but are not aiming at geometric undistortion of the text. 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 rates for both corrupted and correct images.
Keywords. Digital library, quality assessment, warping, cropping

1. Introduction

Many large-scale digitisation projects are running in digital libraries and archives and in public-private partnerships between cultural heritage institutions and industrial partners. The overall digitization outcome 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.

Received: 23.3.2015 / Accepted: 7.11.2015 ISSN 2241-1925

Two of the frequently encountered problems in digitised book collections are warping and incorrect cropping during the automated scan process. In mass digitisation the master-images are usually slightly bigger than the digitized original media. During scan-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 or warped pages. Quality assessment tools that aid the detection of possible quality issues are required. This necessity has been recognised and solutions are available. But most of them are targeting challenging goal of the de-warping or restoration of a document, whereas our goal is just to detect corrupted images.

To address this, the proposed method supports the analysis of digital collections (e.g. JPG files) for cropping and warping problems during the scanning by establishing of geometric bounding.

The main contribution of this paper is the development of a warping and cropping detection tool for the analysis of digital document collections and for decision support about analysed data. The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the warping and cropping detection process and also covers MSER features application. Section 4 presents the experimental setup, an automatic approach for error page detection 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 assessment of digital content by replacing of a human expert regarding the decision-making process in a particular domain. 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 the context of digital preservation in four major National Libraries in Europe.

The CROPDET (Graf et al (2014)) approach of cropping error detection employs computation of average luminance values along the width of the X axis. This approach transforms RGB image into a greyscale image using the perceptually weighted formula and subsequently creates a luminance projection. This method is highly dependent on scan quality and illumination conditions.

Our CWDET (cropping and warping detection) that is the proposed approach employs computation of Maximally Stable Extremal Regions (MSER) features (Matas et al (2002)) for detection of text regions. The MSER technique detects regions, which brightness or colour is different in comparison to surrounding regions. The Meyer (1992) algorithm uses a colour profile in form of its three component profiles for image segmentation based on the watershed transform.

The binary image de-warping approach is presented in (Gatas (2007)) and is used for text line and word detection. This approach is too complex for our goal since we are looking for a text block for further analysis. The Roli (2005) employs distorted text lines for error detection, which belongs to the algorithms type that make use of information derived from the source of the document distortion. Our approach belongs to another type of algorithms, which detect distortion by means of an analysis of the document image. The proposed MSER method appears to be the best choice for the targeted simple text error detection and its subsequent analysis because of its stability (only similar regions are selected), high repeatability score regarding illumination changes and invariance to affine transformation.

3. The proposed warping and cropping detection schema

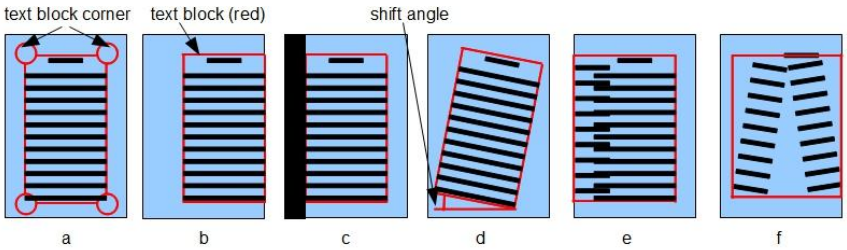

Figure 1. Expected correct document (a) and digitisation failure cases (b-f)

CWDET is a cropping and warping detection tool for quality assessment in document image collections. In order to detect the most frequent problems we regard five digitisation failure cases. The first failure case (b) detects a mis-cropped page where, due to the cropping, a text that is shifted to the one border of the page image and therefore one border is much wider than another border. In the second case (c), 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 (d), an image comprises part of the text from the adjacent page. Fourth case (e) handles images with rotated text. Finally, in the fifth case (f), image contains physically or optically warped text.

The proposed MSER feature based image warping and cropping detection employs parameters that can be defined for each book by an evaluation expert. 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 spatial layout rules. That is, usually a text block (a) 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 and Y axes and is employed for border analysis.

The second important parameter is the maximal allowed text area shift angles on X and Y axes in relation to the border line of the image. Usually for a correct geometrically aligned document these are of a small size due to the fact that text rows and columns should be parallel with the border lines.

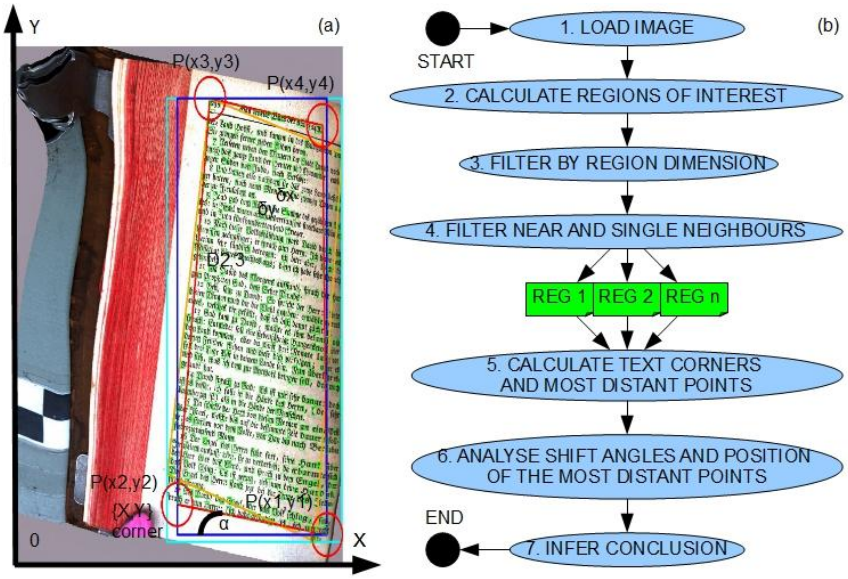

Figure 2. Warping and cropping detection algorithm: (a) computation parameter, (b) workflow

Collection analysis is conducted according to the quality assurance workflow shown in Figure 2b. In order to detect documents with warping and cropping errors we aggregate document specific data (text block position and dimension) and perform analysis on the spatial distribution of MSER features and combine it with expert prior knowledge on text block attributes. In the first workflow step the image is loaded in RGB colour format. In the next step the RGB image is converted to greyscale format. Alternatively, the MSER detection scheme can be carried out for all color channels as well. In the subsequent step we analyse the greyscale image and calculate hypothetical text block (region of interest) boundaries employing MSER algorithm and individual segmented characters. Detected characters are marked by the green rectangles (see Figure 2a). In the third step we filter character rectangles by their size (δx_i for width and δy_i for height in Formula 1), mean value μ (with minimal s_{min} and maximal coefficient s_{max} from Formula 2) and border relation b_i considering them for text regions. Additionally, in the fourth step we check that detected letters have neighbouring letters (D_{i12} in Formula 3) and are not single. In further steps we analyse detected letters and calculate corners of a text block and most distant points, marked by red and orange lines accordingly. For reasoning on aggregated data, using Formulas 4-6, we analyse the angles α_j between text block corner lines

and the position of the most distant points (e.g. *RU{X,Y}* for the right upper corner and *LMD{X,Y}* for the left most distant point from Formula 4 (see Figure 2a) are the coordinates of the lowest text region on the left side). *i* is a number of characters (green small rectangle elements):

$$b_i = \frac{\delta x_i}{\delta y_i}, (1)\ \mu s_{min} \le b_i \le \mu s_{max}, (2)\ D_{i12} = \sqrt{(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2}. (3)$$

Calculation of corner points $P(x_j,y_j)$ for the red quadrangle in Figure 2a:

$$LD\{X,Y\} = \{minx_i, miny_i\}, LU\{X,Y\} = \{minx_i, maxy_i\},$$
$$RU\{X,Y\} = \{maxx_i, maxy_i\}, RD\{X,Y\} = \{maxx_i, miny_i\},$$
$$LMD\{X,Y\} = \{minx_i, y_i\}, RMD\{X,Y\} = \{maxx_i, y_i\},$$
$$UMD\{X,Y\} = \{x_i, maxy_i\}, LMD\{X,Y\} = \{x_i, miny_i\}, (4)$$

$$r: \{l_1: y = k_1x + d_1; l_2: y = k_2x + d_2; l_3: y = k_3x + d_3; l_4: y = k_4x + d_4\} (5)$$

$$k_j = \frac{y_{j2} - y_{j1}}{x_{j2} - x_{j1}} = tg\ \propto_j, \propto_j = arctgk_j, \quad j = 1..4, \quad x_{j1}, x_{j2}, y_{j1}, y_{j2} \in x_i, y_i. (6)$$

Where red lines *r* from Figure 2a are described by 4 linear equations $l_{1\text{-}4}$. The $k_{1\text{-}4}$ and $d_{1\text{-}4}$ are designate constants that are dependent on particular text corner position, which are marked by the red circle. Formula 6 shows the calculation of an angle α_j between text corners. The number *j* is a particular linear equation from the Formula 5 describing the red quadrangle in Figure 2.

$$R = \{G(x,y) | x \in f(x) \cap y \in f(y)\}, (7) \quad S_{n,m} = \sum_{n,m} \delta_{n,m} d(G_{n,m}, R_{n,m}), \quad (8)$$

$$\delta_{n,m} = \begin{cases} 1\ if\ d(G_{n,m}, R_{n,m}) < d_{max} \\ 0\ else. \end{cases} \quad (9)$$

$$d(G_{n,m}, R_{n,m}) = \sqrt{(R_x - G_x)^2 + (R_y - G_y)^2}. \quad (10)$$

The coordinates for the text regions market by green rectangles are computed using MSER features extraction algorithm. Where *R* (Formulas 7-10) represents the rectangles set computed over the image dimension, $G_{(x,y)}$ represents the coordinate points set of green rectangles depending on $f_{(x)}$ and $f_{(y)}$ functions .

n and *m* represent the rectangle index for X and Y coordinate axis. A neighbouring rectangles count around a filtered rectangle $R_{n,m}$ is computed as a sum over all characters, which are located in acceptable for an expert distance to the $R_{n,m}$. Where $S_{n,m}$ represents the total number of matching detected characters $G_{n,m}$ with coordinates G_x and G_y computed over the dimensions n and m around correspondent point $R_{n,m}$ with coordinates R_x and R_y. *d* represents the distance between an $R_{n,m}$ point and evaluated current rectangle point coordinates. d_{max}

stands for maximal accepted distance where rectangle is considered as a correct value. $\delta_{n,m}$ is a coefficient with value 1 for rectangle range < d_{max} or 0 otherwise.

$$V = \begin{cases} 1 \; if \; (LMD_X > B) \cap (W - RMD_X < B) \\ \cap \, (H - UMD_Y < B) \cap (LMD_y > B), \\ 1 \; if \; (LD_X > B) \cap (W - RD_X < B) \\ \cap \, (RD_Y > B) \cap (LD_y > B), \\ 1 \; if \; (LU_X > B) \cap (W - RU_X < B) \\ \cap \, (H - RU_Y < B) \cap (H - LU_y < B), \\ 1 \; if \; \alpha_1 < A \cap \alpha_2 < A \cap \alpha_3 < A \cap \alpha_4 < A, \\ 0 \; else. \end{cases} \quad (11)$$

Where *V* (Formula 11) represents the valid document. Several expert constants were required for accurate computation: *B* stands for the border offset from X and Y axis with respect to the image width *W* and height *H*. *A* stands for the acceptable α_j threshold. The blue rectangle bounds the text region around the text corners and the bright blue rectangle bounds the text region around the most distant points. Subsequently, we employ expert parameters (expected border distances and maximal allowed shift angle of the text area on the X and Y axes) and associated expert thresholds in order to infer conclusion whether scan is corrupted or not. Figures 3-6 demonstrate calculated values visually. This supports the human expert to infer an informed evaluation of the scan quality of image candidates that could be mis-cropped or warped and evaluate whether these images are in fact affected by the indicated error and to perform rescanning if necessary.

The presented tool is working on a greyscale representation of the image data. 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.

4. Evaluation

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 its historical book holdings ranging from the 16th to 19th century with a scope of 600.000 books (Kaiser (2012)) online. The project includes aspects ranging from digitisation preparation and logistics to quality assessment and online-access of the digitized items. Especially the quality presents a challenge where automatic and semi-automatic tools are required to support the quality process for the vast range and amount of material (described in Kaiser & Majewski (2013)).

Our hypothesis is that image processing techniques could help to detect warping and cropping errors in the document collection. We have analysed two test collections. The documents from the first test collection, with 715 correct

images and 15 empty or text free images, are described in detail in this section and contain text. The analysis of the first collection marked 597 images as correct and 133 images as having warping or cropping error. The second collection comprises a selection of 340 images including a variety of the defined cases (as depicted in Figure 1) of warping and cropping errors. The analysis of the second collection marked 54 images as correct and 286 images as having warping or cropping error. The ground truth data created by human expert mostly (≈84%) confirms this result. Our tool correctly detected most corrupted images as corrupted and correct images as correct. Of course, the accuracy depends on the expert parameter settings and for a larger collection cannot be 100 percent due to the different and often controversial book layouts. Nevertheless, for standard text-pages the algorithm is expected to yield good (>80%) results. See samples for correct and corrupted images in Figures 3-6 with associated analysis results.

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 cropping and warping errors. Images were analysed for previously defined use cases: text shifted to the edge of the image, unwanted page borders, rotated page or unwanted text from previous page on the image and for physically or optically warped documents.

The proposed method has been tested with a variety of images from different origins using a default set of parameters (see Formula 11). Mis-classifications coming along with correct results are few and happen with complex text shapes or in low quality scans. Improvements in the algorithm and filtering of mis-classifications are subjects of a future work.

Figure 3. Sample true positive detections

As shown in Figure 3 even complex text shapes could be correctly detected by the algorithm. Some documents were correctly detected despite the fact that they comprise big images or specifically formatted text. Due to the corner calculation approach that describes a correct shape of a text block it was still possible for the algorithm to correctly detect them.

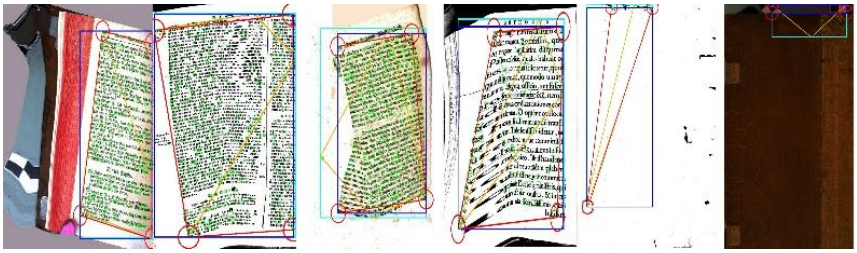

Figure 4. Sample true negative detections

Figure 4 shows true negative digital documents detected in both collections. The first four documents demonstrate cropping and warping errors. The last two address empty and cover cases from a digital book.

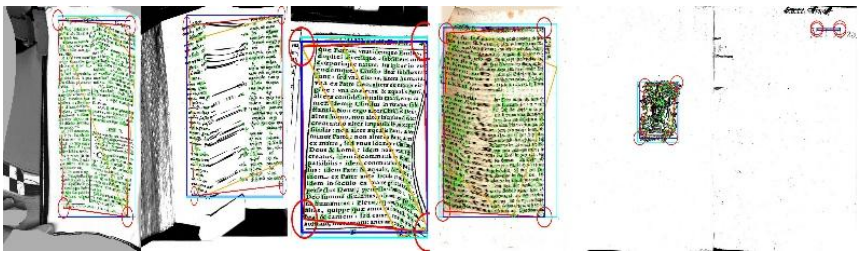

Figure 5. Sample false positive detections

Figure 5 illustrates false positive scans detected in both collections. The reason for these false positive detections is that the warping error exists in the middle of the text. Therefore, additional expert settings are required in order to detect such errors. In the last two documents images and staining and discoloration were mistakenly recognized as a text. Such cases are difficult to avoid and manual expertise is required to eliminate these false positives. This can be fixed by a better scan quality and by setting expected text dimensions, since detected regions in the last two cases could be too small for the text. Such a decision should be made by an institutional expert.

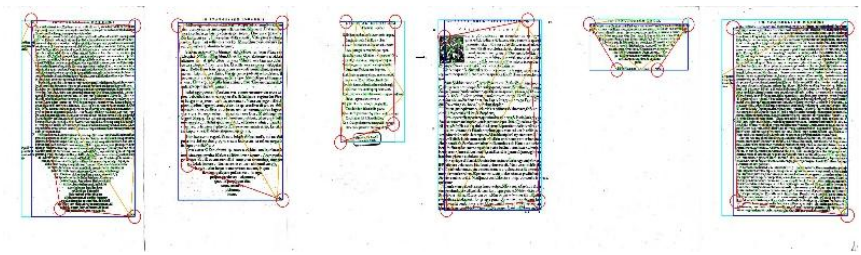

Figure 6. Sample false negative detections

The digital documents presented in Figure 6 were not detected as valid text documents. The reasons are that these text shapes are very complex and

unusual, local discoloration patches confuse the algorithm and lead to incorrect cropping in the last case. The calculation time for one document was about 10 seconds. This time is dependent on page content. The more text information the more calculation time is required. The accuracy was between 83,7% for the first collection and 84,12% for the second collection. The search effectiveness for warping and cropping error detection can be determined in terms of a Relative Operating Characteristic (ROC). For analysis we divided the given document collection in two groups "error pages" and “correct pages". Since the second collection has only error samples we combine both collections for ROC calculation. The tool detected 592 true positive TP documents 300 true negative TN documents, 59 false positive FP duplicates and 119 false negative FN documents. The main statistical performance metrics for ROC evaluation are sensitivity or true positive rate TPR and false positive rate FPR. The sensitivity TPR of the presented approach for the first collection is 0.8326, the FPR is 0.1643. The associated optimum operational point (0.1643, 0.8326) is located very close to the so called perfect classification point (0, 1). These results demonstrate that an automatic approach for error page detection enhances effectiveness of the collection analysis compared to manual analysis.

Therefore, the evaluation results have corroborated our initial hypothesis that image processing techniques based on calculation of MSER features, text corners and most distant text points could be a useful method for detecting cropping errors in document collections.

5. Conclusions

We have developed a quality assessment tool for warping and cropping error detection in document image collection handling. This tool detects images in a document collection where the cropping defined during the automated scan and image post-processing is incorrect and requires the removal of unnecessary border material from the digital image or rotation of the text into a correct position. Additionally this tool detects physically or optically warped documents that should be rescanned. The presented approach supports the analysis of digital collections (e.g. JPG, TIFF, PNG files) for warping and cropping problems e.g. warped text, text shifted to the edge of the image, unwanted page borders, rotated text, or unwanted text from adjacent page on the image. An important contribution of this work is a definition of the evaluation use cases for warping and cropping problems. Another contribution is the creation of a reliable semi-automated expert tool for document warping and cropping detection based on image processing techniques. Our proposed method employs evaluation parameters that are customizable and can be defined for each collection by an institutional expert of digital preservation. The tool works independently of the image size, format and colour. We have analysed two real world collections consisting of correct and corrupted images. Our tool has demonstrated good accuracy for both corrupted and correct images.

References

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

Graf, R., King, R.C., Majewski, S., (2014). Quality Assurance Method for Cropping Error Detection in Digital Repositories., *QQML Proceedings*, pp.551-559, 2014

Gatos, B., Ntirogiannis, K., (2007). Restoration of arbitrarily warped document images based on the text line and word detection. *Proc. of the Fourth IASTED Int. Conf. Sig. Proc., Pattern Recognition, and Applications*, Innsbruck, Austria. (2007), 203-208

Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. *In: Proc. 13th British Machine Vision Conference*, Cardiff, UK. (2002), 384–393

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

Kaiser, M., 'Putting 600,000 Books Online: The Large-Scale Digitisation Partnership between the Austrian National Library and Google', *Liber Quarterly*, 21 (2012), 213–225

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

Roli, F., and Vitulano, S. (Eds.), (2005). Document Image De-warping Based on Detection of Distorted Text Lines. *ICIAP 2005*, Springer-Verlag, Berlin, 1068-1075

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)

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)