Duplicate Detection for Quality Assurance of Document Image Collections *

Reinhold Huber-Mörk
Intelligent Vision Systems
Safety & Security Department
AIT Austrian Institute of Technology GmbH
reinhold.huber@ait.ac.at

Alexander Schindler
Department of Software Technology and Interactive Systems
Vienna University of Technology
schindler@ifs.tuwien.ac.at

Sven Schlarb
Department for Research and Development
Austrian National Library
sven.schlarb@onb.ac.at

ABSTRACT

Digital preservation workflows for image collections involving automatic and semi-automatic image acquisition and processing are prone to reduced quality. We present a method for quality assurance of scanned content based on computer vision. A visual dictionary derived from local image descriptors enables efficient perceptual image fingerprinting in order to compare scanned book pages and detect duplicated pages. A spatial verification step involving descriptor matching provides further robustness of the approach. Results for a digitized book collection of approximately 35.000 pages are presented. Duplicated pages are identified with high reliability and well in accordance with results obtained independently by human visual inspection.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System issues; I.5.5 [**Pattern Recognition**]: Interactive systems

General Terms

Algorithms

Keywords

digital preservation, information retrieval, image processing

1. INTRODUCTION

During the last decade, libraries have been carrying out large-scale digitisation projects, many of them in public-private partnerships with companies like Google or Microsoft, for example, and new digital collections comprising millions of books, newspaper, and journals have been created. Given that each of the single collection items contains up to several hundreds of document images, OCR result files, and other information entities, libraries are facing a paradigm shift in the way how preservation, maintenance, and quality assurance of these collections have to be addressed. Libraries need (semi-)automated solutions that are able to operate on large parts or even on the collections as a whole. Additionally, there are special requirements regarding performance and throughput of the solutions which can be reached by either optimising the time-critical parts of software components or by taking advantage of a distributed software architecture and parallel computing.

In this article, a new approach of document image duplicate detection is presented as a basis for quality assurance in digital library preservation workflows where different versions or derivatives of digital objects have to be maintained and compared to each other. When comparing book pairs, for example, the differences between versions range from variations on the document image level, like additional noise, artefacts, black borders, and more apparent differences due to cropping, page skew, etc., to differences on the object level, like missing or duplicate pages.

Starting with the algorithmic part, there are different aspects of similarity related to document images, including

1. pixel-based similarity, i.e. identity at each pixel, e.g. lossless format conversion, or similarity under lossy compression or radiometrical modifications, e.g. color to greyvalue conversion, color profile adjustment, etc.,
2. similarity under geometrical postprocessing, i.e. scaling, cropping and warping transforms,
3. general similarity induced by independent acquisition under different viewpoint and/or acquisition device and settings.

Figure 1 shows the start of a 730 pages image sequence corresponding to a single book. Starting with the second image a run of eight pages is duplicated from images 10 to 17. Note, that the duplicated images are acquired and post-processed independently. Therefore, the images are geometrically and radiometrically different although showing the same page content.

In general image content comparison is related to visual perception. Perceptual hashing [14, 22], image fingerprinting

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

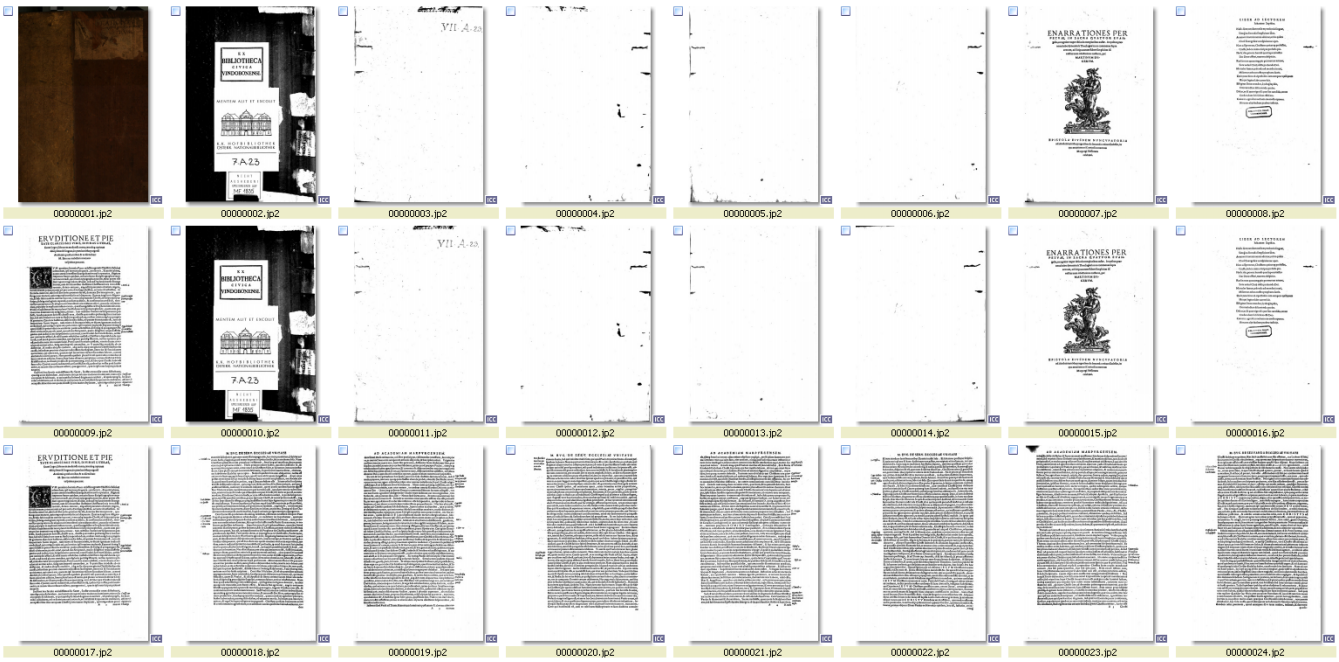

Figure 1: Sample of book scan sequence with a run of eight duplicated pages: images 10 to 17 are duplicates of images 2 to 9 (book identifier is 151694702).

[21] and near-duplicate detection [12, 28] algorithms are related fields. Perceptual similarity, namely structural similarity [25], becomes especially important for comparison of visual content in document images.

Hashing or fingerprinting of images using standard hash functions, like MD5 [18], for example, does only make sense in the narrow domain of bit level image preservation, i.e. if the bitwise representation of the image including all header and formatting information is to be preserved.

The challenges to image processing algorithms can be categorized according to the intensity of preservation actions:

1. The least invasive preservation action for image collections are file format conversions or modifications of the image header information.

2. Preservation actions of moderate intensity are lossy image compression, noise reduction, cropping, scaling and warping transformations, e.g. deskewing.

3. The most invasive modification is completely replacing the representation of an intellectual entity, like the reacquisition of a book in a new scan workflow, for example, possibly involving a different hardware environment and producing differences on the image and/or object level.

Perceptual hashing is interesting especially when significant modifications have been applied to images. Typically, the global characterization of an image, e.g. an individual book page, is obtained to fingerprint the image with respect to its content. The hashing or fingerprinting function has to be designed in a way that equal or similar fingerprints are obtained for perceptual similar images, e.g. cropped, denoised or deskewed images, while significantly different fingerprints should be obtained for images with different content, while having similar global characteristics, e.g. color distribution, image dimensions etc.

Global and structural page comparison commonly relies on information or feature extraction from page images. Optical character recognition (OCR) is an established method for information extraction from document images. OCR heavily relies on appropriate page segmentation and adequate font descriptions. Extraordinary page layout, multilingual texts, archaic language and pages containing graphical representations may lead to practical difficulties when taking an OCR based approach. In contrast to web page analysis, where background information regarding the layout can be derived from the document object model (DOM) of the HTML documents, in the case of scanned document images layout information is only achieved by page segmentation. However, good page segmentation results can only be expected if the text regions are clearly structured and the page layout is generally not too complex. Especially for these difficult cases, where reliable background information is not available we suggest a purely image based approach.

Our approach incorporates and extends state-of-the-art computer vision methods for fast object recognition based on the bag of words (BoW) model for condensed representation of image content. We will present a two-stage workflow for image duplicate detection in the context of book preservation which is basically an image fingerprinting approach creating a shortlist of possible duplicates. Spatial verification based on geometrical matching of images followed by structural comparison is then applied to potential duplicates from the

shortlist.

This paper is organized as follows. In Sect. 2 we review related work in document image analysis and computer vision domain. Section 3 presents our approach along with details on the workflow and algorithms. The experimental setup to evaluate our approach and results are presented in Sect. 4. Conclusions are drawn in Sect. 5.

2. RELATED WORK

Image comparison is an applied research area ranging from the inspection of specific objects in machine vision to very general object identification, classification and categorization tasks. Several approaches for the identification of individual objects in large image collections have been proposed in the literature. Typically, approaches in this area make use of local image descriptors to match or index visual information. Near-duplicate detection of keyframes using one-to-one matching of local descriptors was described for video data [28]. A bag of visual keywords [6], derived from local descriptors, was described as an efficient approach to near-duplicate video keyframe retrieval [26]. For detection of near-duplicates in images and sub-images local descriptors were also employed [12].

Image quality assessment can be dived into reference-based (non-blind) [23, 25, 27] and no reference-based (blind) [9, 15] evaluation. It is well known that image difference measures such as taking the mean squared pixel difference does not correspond to the human perception of image difference [24]. To overcome those limitations the structural similarity image (SSIM) non-blind quality assessment was suggested [25]. SSIM basically considers luminance, contrast and structure terms to provide a measure of similarity for overlaid images.

Related work in the field of analysis of document image collections include tasks such as indexing, revision detection, duplicate and near-duplicate detection. Several authors mention that the use of optical character recognition, which is an obvious approach to extract relevant information from text documents, is quite limited with respect to accuracy and flexibility [1, 7, 17].

An approach combining page segmentation and Optical Character Recognition (OCR) for newspaper digitization, indexing and search was described recently [5], where a moderate overall OCR accuracy on the order of magnitude of 80 percent was reported. Page Segmentation is prerequisite for the document image retrieval approach suggested in [1] where document matching is based on the earth mover's distance measured between layout blocks. The PaperDiff system [17] finds text differences between document images by processing small image blocks which typically correspond to words. PaperDiff can deal with reformatting of documents but is restricted as it is not able to deal with documents with mixed content such as pages containing images, blank pages or graphical art. A revision detection approach for printed historical documents [2] where connected components are extracted from document images and Recognition using Adaptive Subdivisions of Transformation (RAST) [3] was applied to overlay images and highlight differences without providing details on the comparison strategy.

The most similar work, compared to our paper is a method for duplicate detection in scanned documents based on shape descriptions for single characters [7]. Similarly to our approach, this approach does not make use of OCR, but, contrarily to our approach, it is based on some sort of page segmentation, i.e. text line extraction.

3. SUGGESTED METHOD

The suggested workflow is shown in Fig. 2, where the digital image collection refers to the set of scanned book images. Note, our analysis is basically applied to individual books independently and what is usually called a document in text image processing refers to an individual page in our setting. The suggested workflow, for which details will be given below, comprises of

1. Detection of salient regions and extraction of most discriminative descriptors using standard SIFT detector and descriptors [13].
2. A visual dictionary following a Bag of Word approach [6] is created from a set of spatially distinctive descriptors.
3. Once the dictionary is set up, fingerprints - visual histograms expressing the term frequency (tf) for each visual work in the corresponding image - are extracted for each image.
4. Comparison of images becomes matching of visual fingerprints and results in a ranked shortlist of possible duplicates.
5. Taking the top-most ranking image gives a fast result for manual post-processing. If one is interested in a more reliable guess the possible duplicate candidates are subject to spatial verification. Spatial verification is realized by descriptor matching, affine homography estimation, overlaying of images and calculation of structural similarity.

3.1 Algorithmic details

In cases of geometric modifications filtering, color or tone modifications the information at the image pixel level might differ significantly, although the image content is well preserved. Therefore, we suggest to use interest point detection and derivation of local feature descriptors, which have proven highly invariant to geometrical and radiometrical distortions [13, 20] and were successful applied to a variety of problems in computer vision. To detect and describe interest regions in document images we used the SIFT keypoint extraction and description approach. The keypoint locations are identified from a scale space image representation. SIFT selects an orientation by determining the peak of the histogram of local image gradient orientations at each keypoint location. Subpixel image location, scale and orientation are associated with each SIFT descriptor (a 4×4 location grid and 8 gradient orientation bins in each grid cell).

Learning of the visual dictionary is performed using a clustering method applied to all SIFT descriptors of all images, which could become computationally very demanding. As

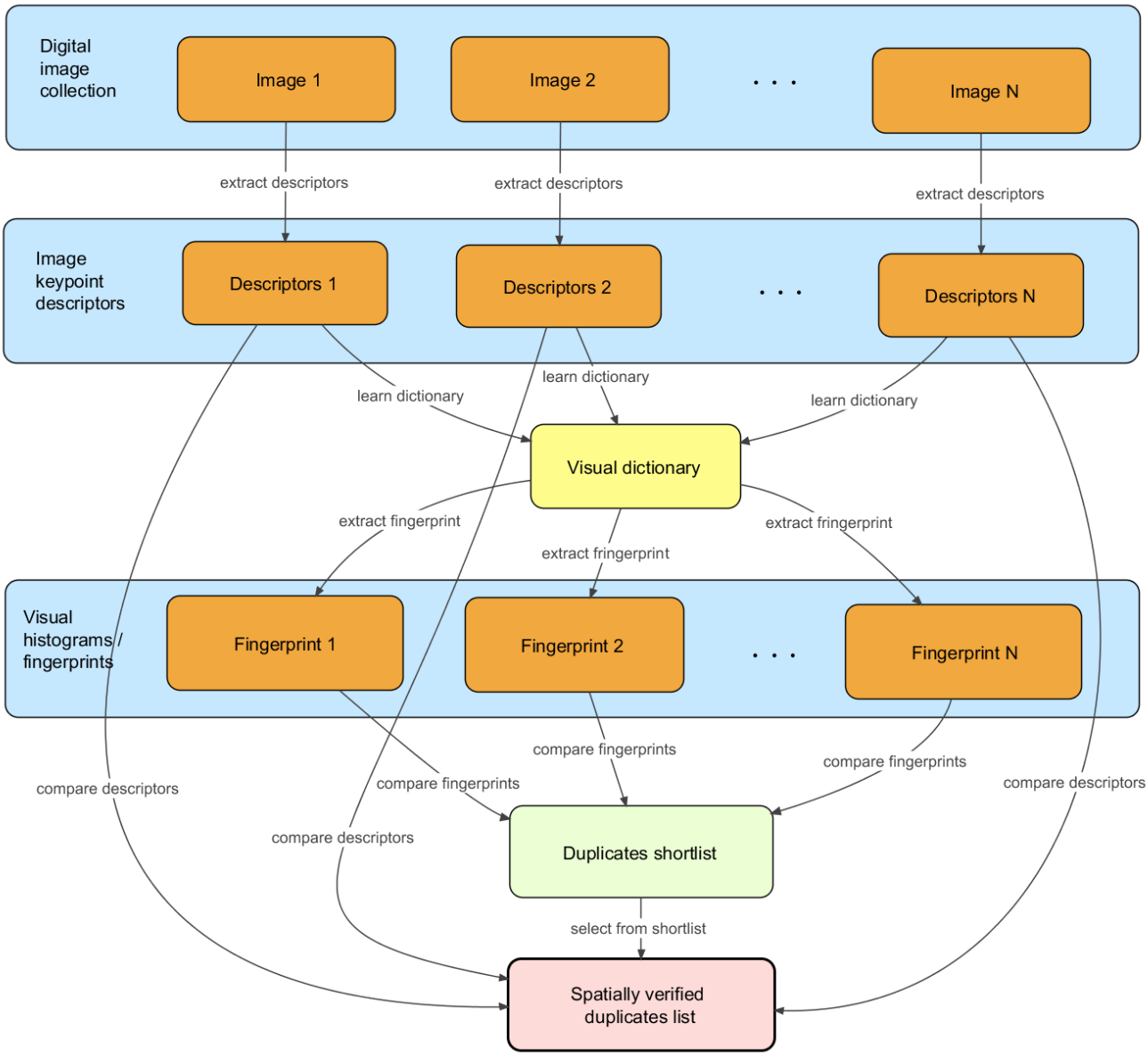

Figure 2: Duplicate detection workflow involving BoW learning, image fingerprinting and spatial verification.

a single scanned book page already contains a large number of descriptors we applied preclustering of descriptors to each image. In contrast to a similar procedure, where all descriptors for all images of the same category are clustered independently and subsequently appended to the BoW [11], we construct a list of clustered descriptors and cluster this list in a second step in order to obtain a dictionary for the whole book. We used k-means for preclustering and final clustering of the BoW. Similar approaches include approximate and hierarchical k-means schemes [16].

Individual terms, or visual keywords, i occur on each page with varying frequency t_i. The visual histogram of term frequencies t_i for an individual book is derived from the BoW representation by counting the indices of the closest descriptors with respect to the BoW. The term frequencies t_i are represented in its normalized form, i.e. $\sum_{i=1\ldots|V|} t_i = 1$, where V is the set of visual words contained in the visual vocabulary for an individual book.

In order to down-weight the influence of terms occurring in a large number of images and up-weight terms occurring only in some specific images the inverse document frequencies (idf) are optionally combined with the term frequencies [19]. The inverse document frequency idf, in our case better called inverse page frequency, reweights the occurrence of individual visual words on single document image page. We used the common definition of idf for an individual visual word t_i given by

$$t_i^{\text{idf}} = \log \frac{|V| + 1}{(v \in V : t_i \in v) + 1}. \quad (1)$$

The combines tf/idf becomes

$$t_i^{\text{tfidf}} = t_i \cdot t_i^{\text{idf}}. \quad (2)$$

Matching of two visual words t^a and t^b is based on histogram intersection $S_{ab} \in [0,1]$ given by

$$S_{ab} = \sum_{i=1}^{|V|} \min(t_i^a, t_i^b). \quad (3)$$

At current each page fingerprint is matched against all other page fingerprints. E.g. for an book containing 1000 pages this results in approx. $5 \cdot 10^5$ calculations of vector intersection distances, which could take several minutes on a single core computer.

Spatial verification is based on the established robust matching method called Random Sample Consensus (RANSAC) [8], where corresponding points are randomly drawn from the set of spatially distinctive keypoints and the consensus test is constrained on an affine fundamental matrix describing the transformation between image pairs. The obtained affine transformation parameters are used to overlay corresponding images by warping one image to the other in order to calculate the structural similarity index SSIM.

Spatial verification is computationally very demanding. It was observed that each document image contains 40.000 descriptors on the average. Matching two such images is a bipartite graph matching task requiring $1.6 \cdot 10^9$ computations of the distance between descriptor pairs. On the other hand, spatial matching of images is the most reliable and detailed approach for quality assurance in image preservation. In order to reduce the computational cost on one hand and get access to a more detailed quality assurance method we suggest the following two algorithmic steps:

1. The number of descriptors is reduced in each image by selecting distinctive local keypoints.
2. Descriptor matching is applied to image pairs extracted from the shortlist obtained by image fingerprint matching.

Spatially distinctive local keypoints are obtained by overlaying a regular grid onto each image and selecting the most salient keypoints from local influence regions centered at each grid point. This approach is related to adaptive non-maximal suppression [4], with main the difference that a regular grid and the measure of saliency as used in the Harris corner detector approach [10] is used in our approach. We found that using a grid point number of 2000 delivers sufficiently matching accuracy. Thus, the required number of vector distance computations in spatially matching a pair of images is reduced to $4 \cdot 10^6$. Using a shortlist of moderate size a combined fingerprinting and spatial matching approach becomes feasible.

Combination of fingerprint matching is combined with spatial verification by

$$S_{ab}^{\text{comb}} = S_{ab} \cdot \text{MSSIM}_{ab}, \quad (4)$$

where $\text{MSSIM}_{ab} \in [0,1]$ is the mean structural similarity index [25].

4. EVALUATION

We evaluated the proposed workflow on a collection of 59 books containing 34.805 high-resolution scans of book pages. Thus, the average number of page images contained in a single book scan was 590. Ground truth data indicating duplicated pages for each book was obtained manually in advance.

The main parameters for the results presented below are summarized as follows. We used standard SIFT features as proposed by [13] providing 128-element vectors. The vocabulary size of the visual BoW was set to 1500 visual words. The number of spatially distinctive keypoints was chosen equal to 2000. The length of the shortlist for spatial verification was 10. All processing was done on greyscale images.

4.1 Comparison of different matching schemes

Using the book with identifier 151694702 and the starting sequence shown in Fig. 1 we compared three query combinations involving

1. visual term frequency histograms only (tf),
2. combined with inverse document frequency (tf/idf),
3. combined with spatial verification (sv).

We calculated the similarity S_{ab} between all image pairs in a book containing N digitized pages. Naturally, there is some level of similarity between text pages due to similar layout, same font etc. Book cover pages have lower similarity to text pages. Finally, duplicated pages show a high similarity. We calculated the maximum similarity found for each image fingerprint when compared to all of the remaining fingerprints

$$S_a^{\max} = \max(S_{ab}), \quad (a,b) \in [1,\ldots,N], a \neq b. \quad (5)$$

The considered book shows two runs of duplicates in the scan sequence: page images $2--9$ are duplicated into page images $10--17$ and there are nested occurrences of duplicates around page images $108--125$. We look for local maxim with respect to the scan sequence of Equ. 5 to identify those runs.

Figure 3 shows the $S_a^{\max}$ versus for each image a in the book scan sequence. A sequence of duplicated images starting approximately from image $a = 100$ is visible in Fig. 3 (a) for matching based on tf only. Contrarily, to expected improvement, the tf/idf matching scheme shown in Fig. 3 (b) shows less discrimination for duplicated images. Both methods, tf and tf/idf are not able to identify the duplicated sequence at the start. The reason for this are empty or nearly empty pages, where only a small number of descriptors could be extracted. Finally, Fig. 3 (c) presents tf matching combined with spatial verification applied to a shortlist of length 10.

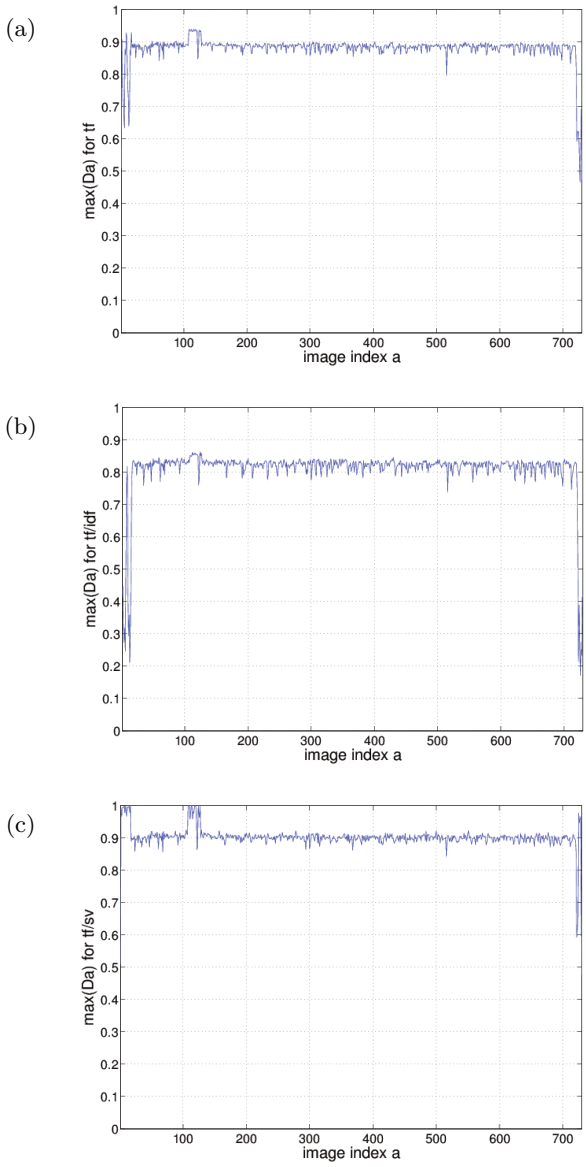

Figure 3: Maximum similarity in duplicate detection using (a) term frequency (tf) only, (b) tf combined with inverse document frequency (idf), (c) tf combined with spatial verification.

Both runs of duplicated sequences of images are visible in this plot.

Remarkably, we observed no advantage using tf/idf compared to the tf matching scheme. The book scan data is characterized by low inter-page variation and the combination with the global idf term seems to lower discriminability for the majority of pages. Therefore, we did not consider the idf term in further experiments. Deeper investigation of this behavior could be topic of future work.

Table 1: Detected duplicates by manual verification and using different image fingerprinting schemes for book 119529601.

Manual detection		Automatic detection tf		tf/sv	
page	dup.	page	dup.	page	dup.
		142	158	142	158
		-157	-173	-157	-173
242	252	242	252	242	252
-251	-261	-251	-261	-251	-261

Table 2: Detected duplicates by manual verification and using different image fingerprinting schemes for book 137274000.

Manual detection		Automatic detection tf		tf/sv	
page	dup.	page	dup.	page	dup.
26	36	26	36	142	158
-35	-45	-35	-45	-157	-173
		264	274	242	252
		-272	-282	-251	-261

4.2 Detailed results for a sample of books

We give a detailed analysis on duplicate detection for a sample of three books. To decide whether an image is a duplicate of another image we applied the following thresholding operation

$$\mathrm{DUP}_a = S_a^{\max} > \Big(\mathrm{median}(S_i^{\max}) + n \cdot \mathrm{mad}(S_a^{\max})\Big), \quad (6)$$

where mad() denotes the median absolute deviation, a robust estimator for the standard deviation

$$\begin{aligned} \mathrm{mad}(S_a^{\max}) &= \mathrm{median}(|S_i^{\max} - \mathrm{median}(S_i^{\max})|), \\ i &= 1 \ldots, N, \end{aligned} \quad (7)$$

The parameter $n = 3$ was found experimentally.

We start with analysis of the book with identifier 119528906. Tab 1 shows that both automatic schemes detected two runs of duplicates. The missing first sequence in manual detection was verified to be a real run of duplicate images.

Tab 2 shows the results for the book with identifier 137274000. The tf and the combined scheme detected tow runs of duplicates. The ground truth did not contain the second run of duplicates, which was verified to be a real run. In the second sequence there is a gap of a singe page image, which caused by the poor quality of the version of the image duplicated at the end of the sequence.

The book with identifier 151694702, also investigated in the last subsection, contains page images occurring three times and even one missing page image. Missing pages could not be detected using our approach. This complicated sequence was identified by both automatic approaches, although it was not found by manual inspection. The tf/sv approach involving spatial verification also detected the duplicate sequence at the begin of the book. The tf approach was not

Table 3: Detected duplicates by manual verification and using different image fingerprinting schemes for book 151694702.

Manual detection		Automatic detection tf		tf/sv	
page	dup.	page	dup.	page	dup.
2-9	10-17			2-9	10-17
		108 -111	118 -121	108 -111	118 -121
		112 -115	124 -127	112 -115	124 -127
		116 -117	124 -125	116 -117	124 -125
				725	11
				726	3
				727	12
				728	6

able to detect this sequence as is mostly consists of nearly empty pages. Additionally, there were four nearly empty pages at the end of the book which were incorrectly identified as duplicates of the empty pages at the beginning of the book. Table 3 list all sequences of duplicates with their location for different matching approaches.

We will present an heuristics to eliminate the four false detections in the next subsection.

4.3 Results for the whole test corpora

We compare the fast tf matching scheme to ground truth obtained by manual page image inspection. Due to computational complexity, we did not include the tf/sv scheme in this experiment. The decision whether a run of pages is detected by counting the detections DUP_i from Equ. 6 of duplicates locally with respect to the sequence number i. In our case, we used a sequence search range of 10 and threshold on the number of locally detected duplicates of 4. The obtained results are shown in Tab. 4. Interestingly, if there are 2 runs all 2 runs are always detected. In total 53 out of 59 books are correctly treated. There remaining 6 books, which are not correctly classified, are characterized by single runs and atypical image content, e.g. graphical art, high portion of blank pages. The simple thresholding strategy given in Equ. 6 derived from global books statistics seems not appropriate for mixed content.

At current, the ground truth contains only books with runs of duplicates, i.e. there is a detection rate of $53/59 \approx 0.9$. Looking at the number of runs of duplicates, i.e. a total number of duplicate runs of 75 was obtained by manual inspection. Automatic inspection delivered 69 duplicate runs, which results in an accuracy for automatic detection of $69/75 = 0.92$.

Actually, using the automatic method more runs of duplicated images are correctly detected, as already shown in the previous subsection. These additional detection are not shown in Tab. 4.

Further investigation concerning adaptive methods to deal with mixed content and computing strategies to involve spatial verification should further improve the results. Additionally, improved ground truth including the duplicate detection correctly indicated by the automatic method could be derived for future experiments.

Table 4: Detected runs of duplicates by manual verification and using fast fingerprinting scheme.

Book identifier	Runs M.	Runs A.	Res	Book identifier	Runs M.	Runs A.	Res
119528906	2	2	ok	119529601	1	1	ok
119565605	1	1	ok	119566804	2	2	ok
119567602	1	1	ok	119572300	2	2	ok
119575003	2	2	ok	119586608	1	1	ok
136403308	2	2	ok	136417009	1	1	ok
136424403	2	2	ok	136432308	2	2	ok
136432400	1	1	ok	136436600	1	1	ok
13646520X	1	1	ok	136465508	1	1	ok
136466203	1	1	ok	136905909	1	1	ok
136975602	1	0	nok	137114501	1	1	ok
137141103	2	2	ok	137141206	1	1	ok
137193905	1	0	nok	137196001	1	0	nok
137203807	1	1	ok	137205804	1	1	ok
137205907	1	1	ok	13721930X	1	1	ok
137220404	1	1	ok	137237301	1	1	ok
137239607	2	2	ok	137247707	1	1	ok
13727100X	2	2	ok	137274000	1	1	ok
150450702	2	2	ok	150709801	2	2	ok
150711807	1	1	ok	150800701	1	1	ok
150803805	1	0	nok	150816800	1	1	ok
150836306	2	2	ok	150920408	1	1	ok
150930402	2	2	ok	150964102	1	1	ok
150976104	1	1	ok	150976207	1	1	ok
151616508	1	1	ok	151638401	1	1	ok
151671106	1	1	ok	151685609	1	1	ok
151687606	1	0	nok	151694209	1	1	ok
151694702	1	1	ok	151698604	1	1	ok
151699207	1	1	ok	152200609	2	2	ok
152213008	2	2	ok	153936506	1	1	ok
162507508	1	0	nok				

4.4 Evaluation in a productive environment

To give an overview on future plans, it is planned to perform an evaluation in a productive environment. First, the accuracy of the book pair comparison is evaluated using an evaluation data set of 50 randomly selected book pairs that will be annotated for that purpose. Second, a large-scale evaluation will be done in order to determine performance and throughput on a distributed system (Hadoop[1]). In this context, we compare the runtime of the data preparation and quality assurance workflows on one machine compared to a Hadoop Map/Reduce job running on a cluster with increasing sample size (50, 500, 5000 books) in various steps up to a very large data set (50000 books).

5. CONCLUSION

We have presented an approach for duplicate detection based on perceptual image comparison using image fingerprinting and descriptor matching. The approach reliably indicates positions in the scanned image sequence containing duplicated images for typical text content. We have shown its capabilities on a complicated multilingual and historical book scan data set. Atypical image content, i.e. non-text content,

[1]http://hadoop.apache.org/

is still an issue to be resolved. Combination with meta-data, such as OCR files and document structure, as well as heuristics incorporating the digitization process, e.g. more detailed information of the scanner operation, through a rule-based system are topics of future research. First heuristics into this direction, i.e. local pooling of duplicate detection events during the scan sequence, were already presented in this work. Further research also includes optimization and deployment of concurrent and parallel computation on the SCAPE platform, especially using the Hadoop Map/Reduce scheme.

6. REFERENCES

[1] van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.: Distance measures for layout-based document image retrieval. In: Proc. of Conf. on Document Image Analysis for Libraries. pp. 231–242 (April 2006)

[2] van Beusekom, J., Shafait, F., Breuel, T.: Image-matching for revision detection in printed historical documents. In: Proc. of Symposium of the German Association for Pattern Recognition. LNCS, vol. 4713, pp. 507–516. Springer (Sep 2007)

[3] Breuel, T.: Fast recognition using adaptive subdivisions of transformation space. In: Proc. of Conf. on Computer Vision and Pattern Recognition. pp. 445 –451 (Jun 1992)

[4] Brown, M., Szeliski, R., Winder, S.: Multi-image matching using multi-scale oriented patches. pp. 510–517. San Diego (June 2005)

[5] Chaudhury, K., Jain, A., Thirthala, S., Sahasranaman, V., Saxena, S., Mahalingam, S.: Google newspaper search - image processing and analysis pipeline. In: Proc. of Intl. Conf. on Document Analysis and Recognition. pp. 621 –625 (July 2009)

[6] Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)

[7] Doermann, D., Li, H., Kia, O.: The detection of duplicates in document image databases. Image and Vision Computing 16(12-13), 907 – 920 (1998)

[8] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395 (June 1981)

[9] Gabarda, S., Cristóbal, G.: Blind image quality assessment through anisotropy. J. Opt. Soc. Am. A 24(12), B42–B51 (Dec 2007)

[10] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of ALVEY Vision Conf. pp. 147–152 (1988)

[11] Hazelhoff, L., Creusen, I., van de Wouw, D., de With, P.H.N.: Large-scale classification of traffic signs under real-world conditions. In: Proc. SPIE Electronic Imaging, Conference 8304W Multimedia on Mobile Devices 2012; and Multimedia Content Access: Algorithms and Systems VI (2012)

[12] Ke, Y., Sukthankar, R., Huston, L.: An efficient parts-based near-duplicate and sub-image retrieval system. In: Proc. of ACM Intl. Conf. on Multimedia. pp. 869–876. ACM, New York, NY, USA (2004)

[13] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Comput. Vision 60(2), 91–110 (2004)

[14] Monga, V., Evans, B.L.: Perceptual image hashing via feature points: Performance evaluation and trade-offs. IEEE Transactions on Image Processing 15(11), 3452–3465 (Nov 2006)

[15] Moorthy, A., Bovik, A.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing 20(12), 3350 –3364 (dec 2011)

[16] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. of Conf. on Computer Vision and Pattern Recognition (2007)

[17] Ramachandrula, S., Joshi, G., Noushath, S., Parikh, P., Gupta, V.: PaperDiff: A script independent automatic method for finding the text differences between two document images. In: Proc. of Intl. Workshop on Document Analysis Systems. pp. 585 –590 (Sep 2008)

[18] Rivest, R.: The MD5 Message-Digest Algorithm. RFC 1321 (Informational) (Apr 1992), `http://www.ietf.org/rfc/rfc1321.txt`, updated by RFC 6151

[19] Robertson, S.: Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation 60, 503–520 (2004)

[20] Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. of Computer Vision 37(2), 151–172 (2000)

[21] Seo, J.S., Huitsmu, J., Kulke, T., Yoo, C.D.: Affine transform resilient image fingerprinting. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 61–64 (2003)

[22] Venkatesan, R., Koon, S.M., Jakubowski, M.H., Moulin, P.: Robust image hashing. In: Proc of Intl. Conf. on Image Processing (2000)

[23] Wang, Z., Bovik, A.: A universal image quality index. Signal Processing Letters, IEEE 9(3), 81 –84 (mar 2002)

[24] Wang, Z., Bovik, A.: Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE Signal Processing Magazine 26(1), 98 –117 (Jan 2009)

[25] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600 –612 (April 2004)

[26] Wu, X., Zhao, W.L., Ngo, C.W.: Near-duplicate keyframe retrieval with visual keywords and semantic context. In: Proc. of ACM Intl. Conf. on Image and Video Retrieval. pp. 162–169. New York, NY, USA (2007)

[27] Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing 20(8), 2378 –2386 (aug 2011)

[28] Zhao, W.L., Ngo, C.W., Tan, H.K., Wu, X.: Near-duplicate keyframe identification with interest point matching and pattern learning. IEEE Transactions on Multimedia 9(5), 1037 –1048 (Aug 2007)