Duplicate Detection Approaches for Quality Assurance of Document Image Collections*

Roman Graf
Research Area Future Networks and Services
Department Safety & Security
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
roman.graf@ait.ac.at

Reinhold Huber-Mörk
Research Area Intelligent Vision Systems
Department Safety & Security
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
reinhold.huber-moerk@ait.ac.at

Alexander Schindler
Research Area Intelligent Vision Systems
Department Safety & Security
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
alexander.schindler@ait.ac.at

Sven Schlarb
Austrian National Library
Vienna, Austria
sven.schlarb@onb.ac.at

ABSTRACT

This paper presents an evaluation of different methods for automatic duplicate detection in digitized collections. These approaches are meant to support quality assurance and decision making for long term preservation of digital content in libraries and archives. In this paper we demonstrate advantages and drawbacks of different approaches. Our goal is to select the most efficient method which satisfies the digital preservation requirements for duplicate detection in digital document image collections. Workflows of different complexity were designed in order to demonstrate possible duplicate detection approaches. Assessment of individual approaches is based on workflow simplicity, detection accuracy and acceptable performance, since image processing methods typically require significant computation. Applied image processing methods create expert knowledge that facilitates decision making for long term preservation. We employ AI technologies like expert rules and clustering for inferring explicit knowledge on the content of the digital collection. A statistical analysis of the aggregated information and the qualitative analysis of the aggregated knowledge are presented in the evaluation part of the paper.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Collection, Dissemination; K.6.4 [**System Management**]: Quality assurance

General Terms

Measurement, Reliability, Experimentation, Verification

Keywords

digital preservation, quality assurance, image processing

*(Produces the permission block, and copyright information). For use with SIG-ALTERNATE.CLS. Supported by ACM.

MEDES'13 October 29-31, 2013, Neumünster Abbey, Luxembourg

1. INTRODUCTION

Quality assurance of document image collections plays an increasingly important role for libraries, archives and companies. These organizations have carried out large-scale digitization projects. New digital collections comprising millions of books, newspapers and journals have been created. Such collections contain hundreds of document images and other information entities. Since manual maintenance and quality assurance of these collections are very time consuming and require high personal and storage costs, institutions are facing a paradigm shift in the manner in which digital preservation and quality assurance of digital collections have to be addressed. There is a strong need for automated solutions that are able to operate on these collections. A typical task in libraries quality assurance is an update of digitized books collections. One such project runs at the Austrian National Library. This digitization project produces digital images from books through an automatic scanning process without involvement of human interaction. The resulting digital collections are stored in the long term digital documents repository. Stored collections are maintained and constantly merged with new versions applying image enhancement and OCR tools. Quality assurance is required in order to select between the old and the new version (see Figure 1) of the associated documents due to the high cost of storage space. A system should be able to automatically make a decision about whether the documents should be overwritten or if human inspection is required. The data currently stored in digital collections is not structured, which additionally com-

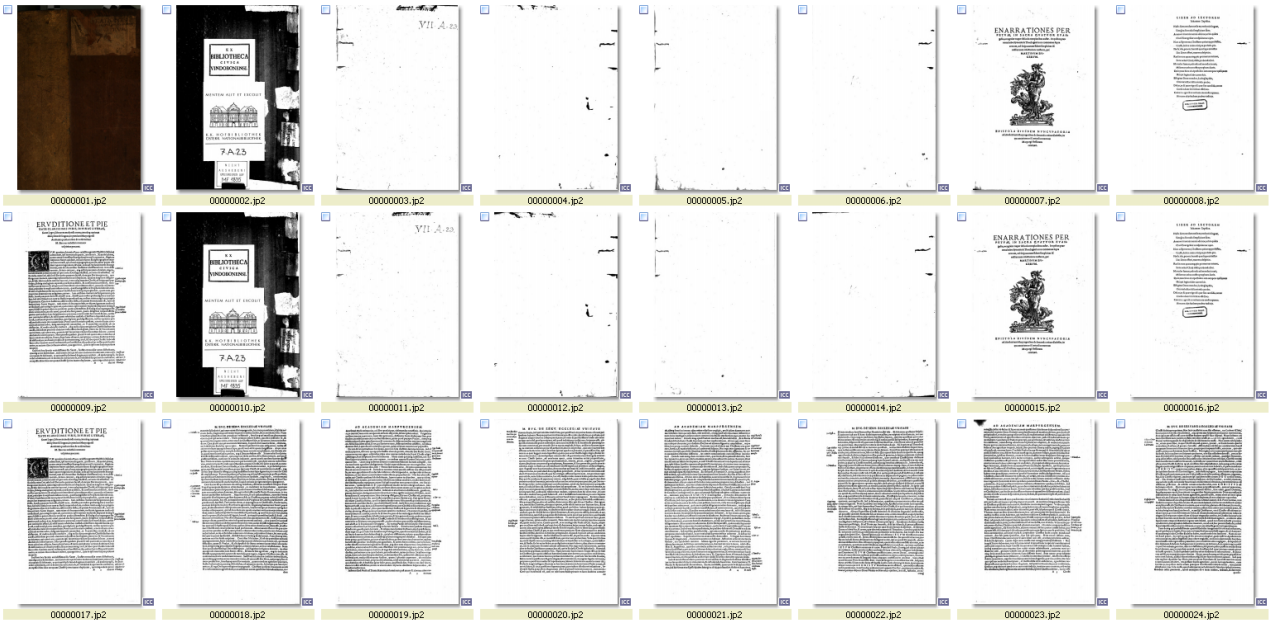

Figure 1: Sample of book scan sequence with a run of eight duplicated pages: images 10 to 17 are duplicates of images 2 to 9 (book identifier is 151694702).

plicates the inspection process. At this time institutions do not have an automatic method to detect duplicates and to remove them. A decision support system is required since users often lack expertise and efficient methods for finding particular images with suspect on duplication in a huge collection. The main proposed method for digital collection analysis is based on the *matchbox* tool [5, 6] that implements image comparison for digitized text documents.

The *matchbox* algorithm is a new innovative method that was published in previous authors work. In this paper this method is compared to another duplicate detection methods. The main contribution of this paper is an evaluation of the *matchbox* tool for the analysis of digital document collections in comparison to three alternative tools. The output of these methods is used for reasoning about analyzed data and for assessment regarding duplicate detection and preservation risks. We aim at identifying the most efficient duplicate detection method that could be used for decision making support for quality assurance of document image collections. For the assessment of evaluation results we use ground data truth, manually created by experts from the Austrian National Library.

The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the duplicate detection methods workflows 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

Methods based on image processing could support the techniques of quality assurance for digital content and replace a human expert regarding the decision-making process in a particular domain. In our duplicate detection approach the similarity computation task is provided by the image processing techniques. One of these tools is a *matchbox* tool, which is based on SIFT [8] feature extraction. In contrast to other evaluated approaches like the one chosen by the *OpenIMAJ* tool [4], SIFT descriptor matching, ORB [11] descriptor matching the *matchbox* tool makes use of a bag of visual words (BoW) algorithm. The SIFT and ORB approaches are very similar, with the difference that we use a different feature descriptions. Typically, approaches in the area of image retrieval and comparison in large image collections make use of local image descriptors to match or index visual information. Near duplicate detection of key frames using one-to-one matching of local descriptors was described for video data [18]. A BoW [2] derived from local descriptors was described as an efficient approach to near-duplicate video key frame retrieval [17]. Local descriptors were employed for the detection of near-duplicates [7]. Several authors mention that the use of optical character recognition, which is an obvious approach for the extraction of relevant information from text documents, is quite limited with respect to accuracy and flexibility [15], [10]. A state of the art with respect to technical requirements and standards in digital preservation is given by Becker et al [1]. Strodl et al [14] present the Planets [12] preservation planning methodology by an empirical evaluation of image scenarios and demonstrate specific cases of recommendations for image content in four major National Libraries in Europe.

3. DUPLICATE DETECTION PROCESS

With increasing amount of digitized data quality assurance plays an important role. Decision making process for quality assurance in digital preservation requires deep knowledge about image processing, file formats and regular library processes. The search for such knowledge is very time consuming, requires an expertise in the domain of digital preservation and image processing skills. A consistent collection

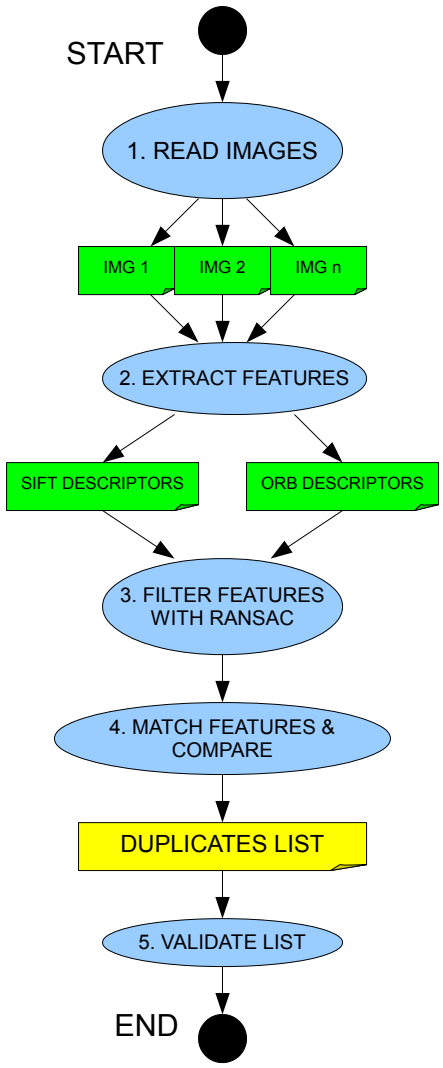

Figure 2: Workflows for OpenIMAJ, SIFT and ORB duplicate detection approaches.

should not contain duplicates or ambiguous entries. Due to huge number of images and text documents we provide automatic image duplicate identification. An additional challenge for manual analysis is that existing information often is either not structured or is only partly structured. Therefore manual search is not possible and is very time consuming. We aim at automatic duplicate detection, verification and support in decision making regarding collection cleaning.

3.1 Image Processing

Application of different digitization methods for the same document might result in information significantly differing at the image pixel level. This depends on performed geometric modifications as well as filtering, color or tone modifications. Therefore, we used interest point detection along with local feature descriptors, which have proven highly invariant to geometrical and radiometrical distortions [8][13] 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 approach [1]. The keypoint locations are identified from a scale space image representation.

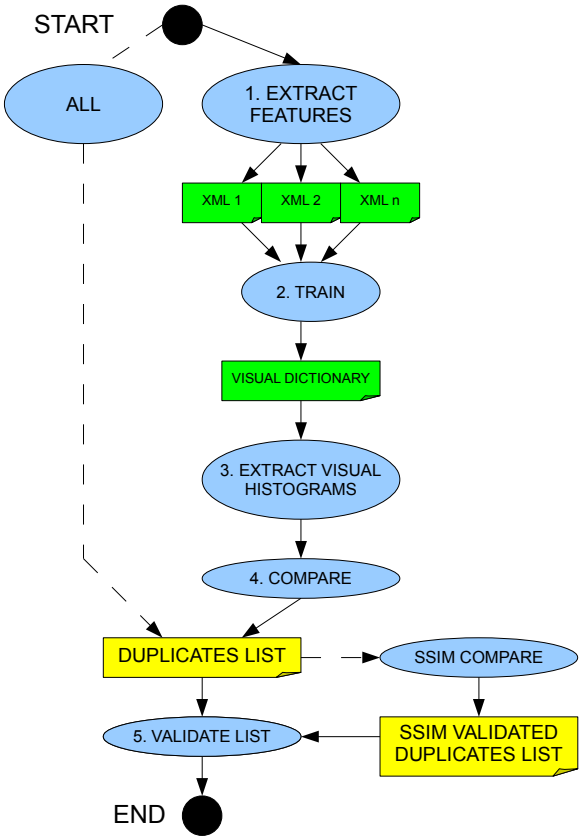

Figure 3: Workflow for matchbox duplicate detection approach.

In our approach we make use of a BoW of about 1000 visual words created using a clustering method applied to all SIFT descriptors of all images in given collection. This can become computationally very demanding. As 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 [9], where all descriptors for all images of the same category are clustered independently and subsequently appended to the BoW, 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. The similarity score between two documents is obtained from the comparison of corresponding keyword frequency histograms followed by structural similarity comparison.

3.2 Duplicate Detection Workflows

In order to detect duplicates we aggregate collection specific knowledge and analyze collections using ORB and SIFT feature matching, as well as the *OpenIMAJ* and *matchbox* tool. Figure 2 demonstrates duplicate detection workflows using SIFT or ORB feature extraction, filtering and matching. Local feature descriptors are extracted from SIFT or ORB keypoints. The main difference of duplicate detection methods is a definition of feature descriptor. The more descriptors per document we have and the more accurate for particular use case they are, the more is a calculation result quality. Robust descriptor matching employs the RANSAC

[3] algorithm which is conditioned on an affine transformation between keypoints locations. In the next step we compare images by matching consistent local features with each other. Finally human expert should validate the list of duplicate candidates. Collection analysis with the *matchbox* tool is conducted according the quality assurance workflow shown in Figure 3. A user triggers a complete collection analysis and the results of which is stored in a text file. In order to handle collection analysis in separate steps the user can extract document features, create a visual dictionary according to the BoW method, create visual histograms based on the visual dictionary for each document and finally perform a pair-wise comparison of all documents. A validation by a human expert might end the process. The "all" path in the workflow means that all workflow steps will be executed per default. Additionally, duplicate candidates contained in a shortlist can be validated by SSIM [16] structural similarity comparison, which requires additional computation time.

4. EVALUATION

The goal of evaluation is an application of different methods for collection analysis for duplicates resulting in its cleaning, i.e. a collection with no duplicates. Additionally, a statistical overview of evaluated data and methods characteristics like performance and accuracy is delivered. The considered collection with identifier Z151694702 is provided by then Austrian National Library and contains 730 documents corresponding to a single book. Manually created ground truth was available.

4.1 Hypothesis and Evaluation Methods of the Collection Analysis

The presented four evaluation methods find duplicate pairs and present them for additional manual analysis and collection cleaning. Our hypothesis is that all automatic approaches would be able to detect duplicates with reliable quality. Then these methods would be a significant improvement over a manual analysis and could be used depending on time and accuracy requirements. We assume that calculation with an OpenCV based python workflow and ORB feature comparison will demonstrate the best performance but lower accuracy. Employing of SIFT feature comparison written in python will have the next best performance with quality comparable to ORB approach. The *OpenIMAJ* tool is written in Java and also makes use of SIFT features with relative simple workflow. We expect that *OpenIMAJ* application will deliver similar accuracy but will be slower when compared to SIFT feature matching implemented in python. The *matchbox* tool implements the most sophisticated image processing workflow and should demonstrate the best accuracy but will require additional time for building of BoW dictionary. Evaluation takes place on an Intel Core i7-3520M 2.66GHz computer using Java, Python and C++ languages on Linux OS. We evaluate duplicate candidate pairs, calculation time and calculation accuracy for each evaluation method.

4.2 Experimental Results and its Interpretation

Table 1 lists all duplicate pairs (Manual1; Manual2) that were discovered by an expert. Duplicated images automatically inferred by the *matchbox* tool written in C++ and python are denoted with (Matchbox1; Matchbox2), whereas 1 and 2 stand for the original and new version of the document. The results of the analysis with *OpenIMAJ* tool written in Java are presented with (OpenIMAJ1; OpenIMAJ2). Application of the OpenCV library based on SIFT features and written in python is depicted by (SIFT1; SIFT2) columns and for ORB features by (ORB1; ORB2) columns. The number of pages between the original and the new version of the duplicated documents in the collection is an additional help to find duplicates, since duplicates often appear in a sequence. The manual analysis of the test collection (M1; M2) shows eight duplicate pairs. The *matchbox* algorithm (Matchbox1, Matchbox2) lasted about 7264 seconds and has detected six duplicate pairs correctly and 10 false positive duplicate pairs. The distribution of the different workflow steps for *matchbox* is 3814 seconds for SIFT descriptors extraction, 2031 seconds for BoW learning and matching and 1419 seconds for concluding spatial verification. The automatic approach of duplicate search did not find two duplicated pages (5 and 6) which were identified as duplicates by manual analysis. The reason for that is the computed average similarity score was higher than the scores of pages 5 and 6. In this specific case we have to deal with nearly empty pages with dominating white color, which makes it difficult to identify these pages as a pair of duplicates. The pages in the range 108 to 115 and page 116 are detected as false positives by the automatic analysis. In contrast to the dominating color case similarity scores are in range here. Manual checking of mentioned pages reveals that there are no duplicates. The reason for detecting false positive is a high structural similarity of digital image data. But this high similarity doesn't always mean semantically text similarity that can be validated only by human expert. The SIFT features method scores with five true positives and ORB feature methods with four true positives. The calculation times of SIFT, ORB and *OpenIMAJ* methods are 95940, 38422 and 3650000 seconds, respectively. We suspect, the reason for the high calculation time using *OpenIMAJ* libraries could be memory leaks slowing down the total calculation time. Therefore, this method is only feasible for small digital collections. Table 1 shows that most of detected false positives are shared by all evaluation methods, whereas *matchbox* demonstrates the highest accuracy (10 false positives) and ORB feature matching the lowest accuracy (23 false positives) with SIFT feature matching (13 false positives) in between. The *OpenIMAJ* method demonstrates low accuracy with only three correct detections among eight possible and 11 false positive results. The relatively high number of false positives for ORB method could be explained with relatively low number of its descriptors per page. A typical text document image in *matchbox* workflow contains up to d=40.000 descriptors. In contrast 2000, 1000 and 400 descriptors on average for *OpenIMAJ* , SIFT and ORB methods, respectively. Matching two images based on the BoW representation in *matchbox* tool requires a single vector comparison. For a sample book with n=730 pages $n*(n-1) = 532.170$ vector comparisons are necessary. In contrast, direct matching of feature descriptors requires between $d^2 = 1.6 * 10^5$ (ORB) and $d^2 = 1.6 * 10^9$ (*matchbox*) vector comparisons for a single pair of images. Therefore, direct feature matching is much more computationally intensive but its workflow is simpler then *matchbox* implementation. The average relative computational costs for *matchbox*

Table 1: Manually (Manual1,2) and automatically (Matchbox1,2, OpenIMAJ1,2, SIFT1,2, ORB1,2) detected duplicates. Numbers 1 and 2 in the table header present duplicate pairs (e.g. 2 and 10 in the first row is a duplicate pair). The relatively large number of duplicates detected by ORB method often is a false positive detection.

Manual1	Manual2	Matchbox1	Matchbox2	OpenIMAJ1	OpenIMAJ2	SIFT1	SIFT2	ORB1	ORB2
2	10	2	10	2	10	2	10	2	10
3	11	3	11					3	37
4	12	4	12			4	12	4	363
5	13							5	127
6	14							6	127
7	15	7	15	7	15	7	15	7	15
8	16	8	16			8	16	8	16
9	17	9	17	9	17	9	17	9	17
								11	37
								12	363
						13	14	13	127
								14	127
		108	118			108	118	108	118
		109	119	109	119	109	119	109	119
		110	120	110	120	110	120	110	120
		111	121	111	121	111	121	111	121
		112	116	112	124	112	124	112	116
		113	125	113	125	113	125	113	125
		114	126	114	126	114	126	114	126
		115	127	115	127	115	127	115	127
				116	112	116	124	116	112
		117	125	117	125	117	113	117	125
		124	112	124	112	124	116	124	116
								723	37
				725	729			725	37
								726	37
						728	703	728	37

Table 2: Experimental analysis results. Manual verification detected eight duplicates in the collection of 730 documents. AVG descriptors/page means the number of feature descriptors in a document for associated algorithm.

Algorithm	AVG descriptors/page	Calculation time (sec.)	TP	FP	TN	FN	Sensitivity (ROC)	FPR (ROC)
Matchbox	40000	7264	6	10	712	2	0,75	0,012
SIFT	1000	95940	5	13	709	3	0,625	0,018
ORB	400	38422	4	23	708	4	0,5	0,032
OpenIMAJ	2000	3650000	3	11	711	5	0,375	0,015

workflow are 53 percent for feature extraction, 28 percent for BoW construction and 19 percent for actual comparison. All of presented methods demonstrate ability to detect duplicate documents and can be applied for quality assurance of digital collections. All of these approaches help to automatically find out duplicate candidates in a huge collection. Following this, manual analysis of duplicate candidates separates real duplicates from structural similar documents and evaluates resulting duplicate list. Presented methods save time and therefore costs associated with human expert involvement in quality assurance process. Therefore the choice of the correct duplicate detection method is a trade-off between performance and high accuracy. The advantage of using *OpenIMAJ* tool is that user does not need deep knowledge in image processing domain. The *matchbox* tool demonstrates the best detection accuracy combined with relative good performance. Therefore our initial hypothesis is verified. But further research is required to improve performance and accuracy metrics of mentioned methods.

4.3 Effectiveness of the Duplicates Search

The duplicates search effectiveness can be determined in terms of a Relative Operating Characteristic (ROC). Similarity analysis divided the given document collection (book identifier is Z151694702) in two groups “duplicates” and “single” by similarity threshold. For the *matchbox* method we detected 6 true positive TP duplicates, 712 true negative TN documents, 10 false positive FP duplicates and 2 false negative FN documents. The main statistical performance metrics for ROC evaluation are sensitivity or true positive rate TPR and false positive rate FPR (see Equation 1). Table 2 summarizes evaluation results.

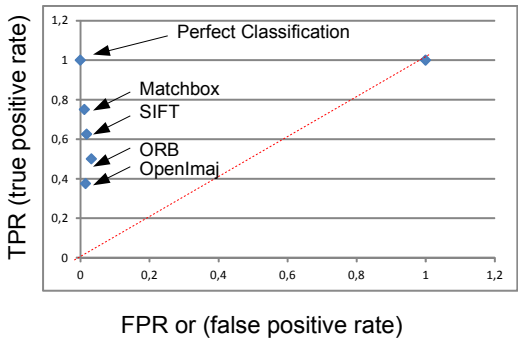

Figure 4: ROC space plot for different duplicate detection approaches.

$$TPR = \frac{TP}{(TP + FN)}, FPR = \frac{FP}{(FP + TN)}. \quad (1)$$

Therefore the sensitivity of the presented *matchbox* approach is 0.75, the FPR is 0.012. ORB, SIFT and *OpenIMAJ* are represented by (0.032, 0.5), (0.018, 0.625) and (0.015, 0, 375) points respectively. The ROC space demonstrates that the calculated FPR and TPR values form all these points are located very close to the so called perfect classification point (0, 1). These results demonstrate that the automatic approaches for duplicates detection are effective and it is a significant improvement compared to manual analysis. We evaluated the ROC space plot (Figure 4) on a set of duplicate detection method containing four approaches. Each point is defined by FPR and TPR rates of associated method. The best possible classification is represented by the point (0, 1). The distribution of collection points above the red dashed diagonal demonstrates quite good classification results and justifies using of these methods for duplicate search. The *matchbox* tool demonstrates the best classification results.

5. CONCLUSIONS

In this paper we presented an evaluation of different methods for automatic duplicate detection in digital image collections. These approaches support quality assurance and decision making for long term preservation of digital content in libraries and archives. In this paper we evaluated advantages and drawbacks of different approaches. Important contribution of this work is the selection the most efficient method which satisfies the digital preservation requirements in terms of accuracy and performance for duplicate detection in digital document image collections. In designed workflows of different complexity we applied image processing techniques in order to demonstrate possible duplicate detection approaches. Applied methods make use of expert knowledge that facilitates decision making for long term preservation. We employed AI technologies like expert rules, and clustering for inferring explicit knowledge on the content of the digital collection. A statistical analysis of the aggregated information and the qualitative analysis of the aggregated knowledge were performed in the evaluation part of the paper. The evaluation results demonstrate that the *matchbox* tool has the best results in terms of detection accuracy and performance. The ORB method is simple and the fastest but does not show high accuracy and reliability. The SIFT approach demonstrates high accuracy and acceptable calculation time but is less efficient in comparison to the *matchbox* regarding features count per image, architectural efficiency and computation time. The *OpenIMAJ* tool demonstrates moderate results regarding accuracy and acceptable results for small collections but for larger collections it requires a lot of time for computation due to architectural drawbacks. As future work we plan to extend an automatic quality assurance approach of image analysis for mixed (image/text) documents to other image processing techniques and digital preservation scenarios. We also plan an evaluation with a large set of digital books. That is required to improve analysis quality.

6. 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).

7. REFERENCES

[1] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. In *Int. Journal on Digital Libraries*, volume 10, pages 133 – 157, 2009.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

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

[4] J. S. Hare, S. Samangooei, and D. P. Dupplaw. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 691–694, Scottsdale, Arizona, USA, Nobember 28 - December 1 2011.

[5] R. Huber-Mörk and A. Schindler. Quality assurance for document image collections in digital preservation. In *Proc. of the 14th Intl. Conf. on ACIVS (ACIVS 2012)*, volume 7517 of *LNCS*, pages 108–119, Brno, Czech Republic, September 4-7 2012. Springer.

[6] R. Huber-Mörk, A. Schindler, and S. Schlarb. Duplicate deterction for quality assurcance of document image collections. In *In iPRES 2012 - Proceedings of the 9th International Conference on Preservation of Digital Objects*, pages 136–143, Toronto, Canada, October 1-5 2012.

[7] Y. Ke, R. Sukthankar, and L. Huston. An efficient parts-based near-duplicate and sub-image retrieval system. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pages 869–876, New York, NY, USA, 2004. ACM.

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

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *In: Proc. of the IEEE CCVPR*, 2007.

[10] S. Ramachandrula, G. Joshi, S. Noushath, P. Parikh, and V. Gupta. Paperdiff: A script independent automatic method for finding the text differences between two document images. In *The Eighth IAPR Intl. Workshop on DAS*, pages 585 –590, Sep 2008.

[11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571, 2011.

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

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

[14] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. How to choose a digital preservation strategy: evaluating a preservation planning procedure. In *In: JCDL 2007: Proceedings of the 2007 conference on digital libraries*, pages 29 – 38, New York, NY, USA, 2007. ACM.

[15] J. van Beusekom, D. Keysers, F. Shafait, and T. Breuel. Distance measures for layout-based document image retrieval. In *2nd ICDIAL, 2006. DIAL '06*, pages 231–242, April 2006.

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

[17] X. Wu, W.-L. Zhao, and C.-W. Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 162–169, New York, NY, USA, 2007. ACM.

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