Blank Page and Duplicate Detection for Quality Assurance of Document Image Collections *

Roman Graf[1], Ross King[1], and Sven Schlarb[2]

[1] Research Area Future Networks and Services
Department Safety & Security
Austrian Institute of Technology
{roman.graf,ross.king}@ait.ac.at
[2] Austrian National Library
sven.schlarb@onb.ac.at

Abstract. Digitization workflows for automatic acquisition of image collections are susceptible to errors and require quality assurance. This paper presents an automatic expert system for long term preservation that supports decision making for blank page and accurate duplicate detection in document image collections. The important contribution of this work is a definition of the expert rules with associated severity level and its automatic computation. Our goal is to create a reliable inference engine and a solid knowledge base from the output of an image processing tool that detects blank pages and duplicates based on methods of computer vision. We employ artificial intelligence technologies (i.e. knowledge base, expert rules) to emulate reasoning about the knowledge base similar to a human expert. In order to improve analysis accuracy we use OCR tool for blank page and duplicate detection. The novelty of this approach is an application of OCR method for this task. A statistical analysis of the automatically extracted information from the image comparison tool and the qualitative analysis of the aggregated knowledge are presented.

Keywords: expert system, digital libraries, image processing

1 Introduction

During the last decades, libraries, archives and museums have been carrying out large-scale digitization projects at different scales. New digital collections comprising millions of books, newspapers, journals, and other digital object representations have been created, and internet-based access to a wide range of cultural heritage resources that so far have only been available in analogue form is now possible. Depending on the type of digital object, a typical book, newspaper or journal collection item can contain hundreds or even more document images and other related information entities. Due to the scale of digital information that has to be managed, memory institutions are facing a paradigm shift in the way how preservation, maintenance, and quality assurance of

* 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)

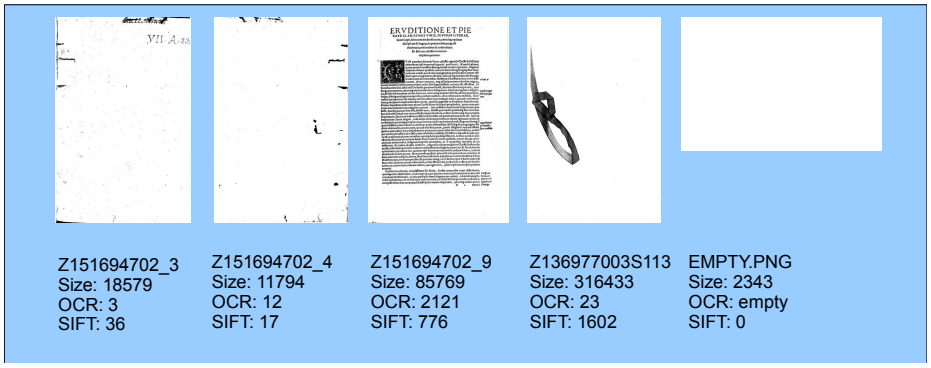

Fig. 1. Selected samples of blank pages in digital collections from different sources with associated file name, file size, OCR and scale-invariant feature transform (SIFT) analysis result.

these collections have to be addressed. For that reason, automated solutions for data management and digital preservation are absolutely necessary.

In a typical book digitisation workflow, the ability to update a digital copy is a frequent requirement. Either a new digital copy is created by scanning the original analogue resource again, or a new digital derivative based on the raw digital object is produced. The new derivatives are either created to get an improved representation of the digital image by removing page borders or skew, or to enrich the digital object by adding related information like full text, layout or semantic representations, etc.

In this context, image analysis and comparison technology can help to align and document changes that occur from one version to another. The detection of blank pages (see Figure 1) in a digital book and the selection between the old and the new version of the associated documents (see Figure 2) are basic operations in this regard. Based on this information a decision support system can automatically make a recommendation if a digital object can be safely overwritten or if human inspection is required. The Expert System proposed in this article is based on image processing and OCR methods, which implement image analysis and comparison for digitized text documents.

The main contribution of this paper is the development of a rule-based Expert System for the analysis of digital document collections, and for reasoning about analyzed data and for assessment regarding blank pages and duplicated images.

The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the duplicate and blank page detection process and also covers image processing and expert rules definition 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

In artificial intelligence, rule-based Expert Systems support the techniques of quality assurance for digital content and replace a human expert regarding the decision-making process in a particular domain. An expert system comprises the inference engine that is employed for reasoning about the knowledge base. The knowledge base contains expert knowledge in form of data or rules.

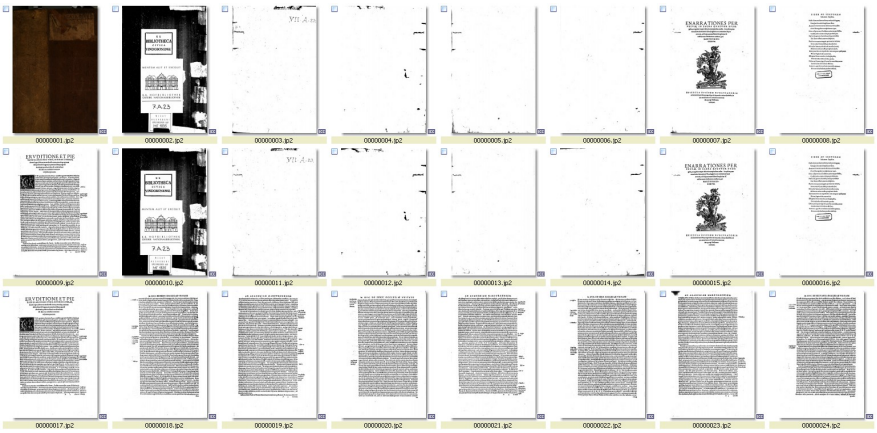

Fig. 2. Sample of book scan sequence with a run of eight duplicated pages: images 10 to 17 are duplicates of images 2 to 9 (book identifier is 151694702).

The implementation of an expert system for color retrieval described by Yoo et al [12] proposes an image retrieval system using color-spatial information from the content-based image retrieval applications. In contrast to our Expert System approach, the described system does not implement rules, although it does perform similarity computation. In our duplicate detection approach the similarity computation task is provided by the image processing techniques. One of the possible tools is a *matchbox* tool [5], which is a modern quality analysis tool based on SIFT [7] feature extraction. In contrast to SIFT descriptor matching, the matchbox tool makes use of a bag of visual words (BoW) [3] algorithm. The SIFT descriptor approach is very similar to the *matchbox* tool, with the difference that we do not use BoW and structural similarity matching in order to increase performance, whereas the *matchbox* tool guarantees better accuracy employing these technics. 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 [13]. A BoW derived from local descriptors was described as an efficient approach to near-duplicate video key frame retrieval [11]. Local descriptors were employed for the detection of near-duplicates [6].

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 [2], [9]. But employing of OCR methods in our approach meant just as a complimentary support for image processing methods and is just one expert rule in a pool of other duplicate detection rules. And innovative OCR methods application for blank page detection do not require high accuracy, since we employ a threshold metric for result estimation.

The rule-based system presented by Bernard [1] is designed for process and power control in a power plant. In order to evaluate a control action the relevant parameters

are measured. Actions are specified in rules. Given the current state of the plant and the desired objectives, the fuzzy logic and inference engine is used to search through the knowledge base in order to identify those rules that are applicable. This approach is very similar to our Expert System organization, with the difference that we have a different application field and another input parameters. Compared to existing systems the proposed system is more efficient due to the use of SIFT features instead of color signatures and filtering, and it is more simple without the use of linguistic variables for fuzzy logic. The proposed system is unique for the given domain.

3 Duplicate and Blank Page Detection Process

Due to huge number of images and text documents in modern digital collections the quality assurance plays an increasingly important role. Decision making process for quality assurance in digital preservation requires deep knowledge about image processing, file formats and regular library processes. The manual search for such knowledge is very time consuming, requires an expertise in the domain of digital preservation and image processing skills. A consistent collection should not contain duplicates or blank pages. Therefore we aim at providing automatic image duplicate and blank page identification and verification methods in order to support decision making regarding the collection cleaning. An additional challenge for manual analysis is that existing information often is either not structured or is only partly structured.

The Knowledge Base shown in Figure 3(a) is required in order to collect information and to perform automatic document assessment and duplicates detection. A basis for accurate reasoning is information aggregated from digital documents in image collection and from knowledge provided by human experts.

3.1 Expert Rules Identification

To organize the Knowledge Base we must structure the information that has been obtained from the domain experts of digital preservation and from conducted experiments. We define typical scenarios and identify the parameters used by library experts for collection handling. Then we define the linguistic labels to classify measured values of each parameter and associated ranges. Finally, we determine the conditional rules that relate these linguistic labels to specific consequences. Information retrieved from the image collection is processed by the customized domain model. This model enables structured and maintainable handling of analyzed data. If necessary, the data could be stored in a database for further treatment. A user communicates with the Expert System by sending a request query and receives an advice in response. A user could leverage these rules according the requirements and circumstances for a particular book or collection for example if a file name has a semantic meaning or if the file size is of interest for analysis. Possible actions according to the advice provided by the Expert System include removing of a document, ignoring the advice and performing a new scan for the particular image or a collection including similarity analysis.

The previously defined rules should be organized in order to process input statements (assertions) and to infer appropriate advice and conclusions. Forward rule chaining for duplicate or blank page detection is presented in Figure 3(b). Forward chaining

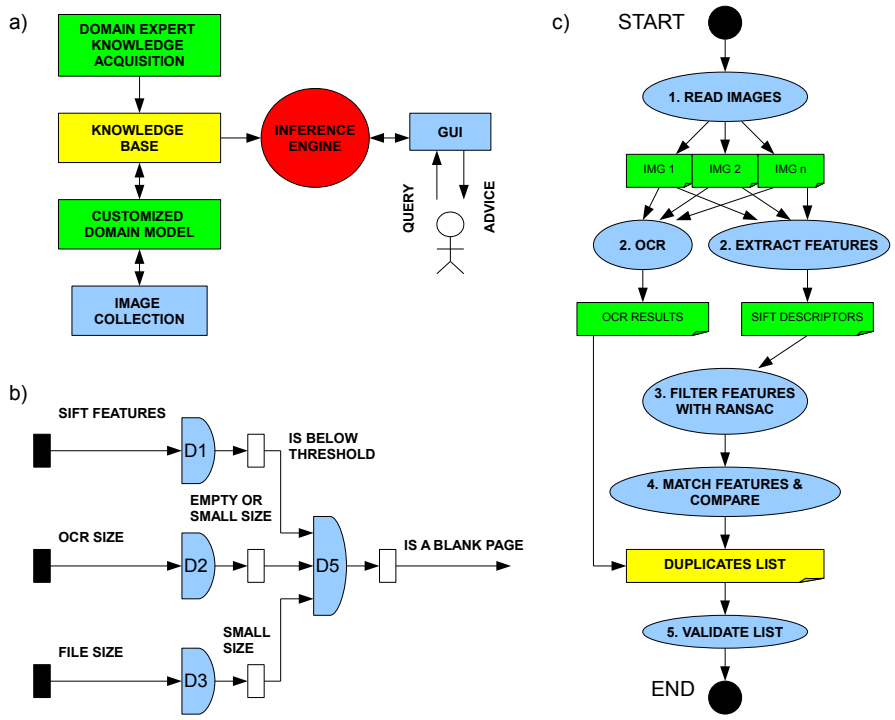

Fig. 3. Expert system: (a) overview, (b) forward rule chaining for duplicate and blank page detection and (c) duplicate detection workflow.

is the process of moving from the "if" patterns (antecedents) to the "then" patterns (consequents) in a rule-based system. We consider the antecedent as satisfied when the "if" pattern matches the assertion. Assertions are depicted by black rectangles on the input side and by the white rectangles on the output side, respectively. The rules are presented by blue half-spheres. A specific rule is triggered if all of its antecedents are satisfied. A triggered rule is considered as fired if it produces a new assertion or performs an action on the output (white rectangle). Since our Expert System is focused on duplicate and blank page detection there is no need for any conflict-resolution procedure to resolving possible rule conflicts. The rules are weighted according to their severity and can be extended. The weights are customizable and can be adjusted according to user requirements.

In Figure 3(b) we present rules distinguishing blank pages from non-blanked ones. The rule-base system starts blank page identification with the rule D1. Suppose that SIFT features score of particular document is over particular adaptive threshold. Then if the antecedent pattern matches that assertion, the value x becomes "is a blank page candidate" and the rule D1 fires. Because the document contains text (OCR size) and file size is not null, rule D5 fires, establishing that the document "is a blank page". Similarly we go through remaining rules. The final conclusion of the rule-based system is whether there is a blank page observed. The inference engine performs conditional rules and blank page analysis, infers appropriate action and formulates advice using relation of linguistic labels to specific consequences. The OCR score values could be

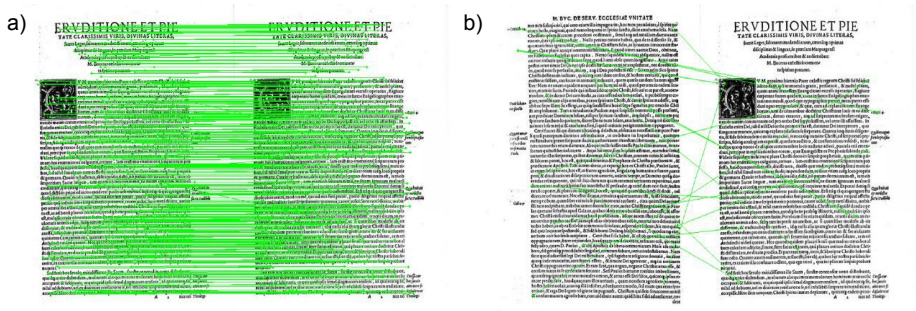

Fig. 4. Evaluation results samples from book identifier 151694702 for duplicate detection with SIFT feature matching approach: (a) similar pages with 419 matches, (b) different pages with 19 matches.

leveraged for duplicate detection 3(c) as additional method to the image processing methods and metadata analysis.

3.2 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 [7][10]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. The keypoint locations are identified from a scale space image representation. In our approach we make use of a direct matching of SIFT descriptors in contrast to the procedure [8], where all descriptors for all images of the same category are clustered independently and subsequently appended to the BoW. Our approach differs also from the *matchbox* approach, where a list of clustered descriptors is constructed and in the second step this list is clustered in order to obtain a dictionary for the whole book. The similarity score between two documents is obtained from the comparison of corresponding SIFT features followed by OCR output comparison.

3.3 Duplicate Detection Workflow

Collection analysis is conducted according the quality assurance workflow shown in Figure 3(c). The user triggers a complete collection analysis, the results of which are stored in a text file. In order to detect duplicates we aggregate collection specific knowledge and analyze collections using SIFT feature extraction demonstrated in Figure 4, filtering and matching, as well as the OCR analysis. Local feature descriptors are extracted from SIFT keypoints. Robust descriptor matching employs the RANSAC [4] 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. Additionally, duplicate candidates contained in a shortlist can be validated by OCR comparison, which requires additional computation time and also is limited for documents with printed text written in supported language.

4 Evaluation

The goal of evaluation is an application of different methods for collection analysis for duplicates and blank page detection resulting in its cleaning, i.e. a collection with no duplicates and blank pages. Additionally, a statistical overview of evaluated data and characteristics like performance and accuracy is delivered. The suggested Expert System processes reasons on found blank pages and duplicates and generates advice on how to clean up the collection.

4.1 Hypothesis and Evaluation Methods of the Collection Analysis

The presented two evaluation use cases find duplicate pairs and blank pages and present them for additional manual analysis and collection cleaning. Our hypothesis is that automatic approach should be able to detect blank pages and duplicates with reliable quality. We consider two use cases. First one is a duplicate detection. The OCR analysis should prove the results of image processing methods and OCR scores for similar files should have similar OCR scores. The second use case is a blank page detection. For this use case OCR scores should be null or near to null as well as SIFT descriptors score should be very low. We also aim to evaluate whether file size of blank page could be a reliable parameter for blank page analysis. If described hypothesis is true then this methods would be a significant improvement over a manual analysis. 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.

We assume that employing of SIFT feature comparison and calculation with an OpenCV 2.4.3 based python workflow and OCR analysis will demonstrate good performance by sufficient good accuracy. Evaluation takes place on an Intel Core i73520M 2.66GHz computer using Java 6.0 and Python 2.7 languages on Linux OS. We evaluate duplicate candidate pairs, calculation time and calculation accuracy for each evaluation method. For OCR analysis we use Tesseract 3.02 tool.

4.2 Experimental Results and its Interpretation

The threshold value 0.9 was determined using statistical approach and robust estimators. The threshold can be adapted dependent from the content. In the first instance we apply "SIFT features" rule. For this rule we compute average similarity score over the all similarity scores provided as an output of the SIFT descriptors matching analysis. In conjunction with similarity threshold rule we are able to isolate most of the duplicate pairs. 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. Some of the detected duplicates have a dominating color and relative high similarity score like documents 2 - 9. These documents should be verified manually and independent from the average similarity score and offsets. OCR scores comparison provides additional help for working with these documents.

The manual analysis of the test collection shows eight duplicate pairs. The automatic approach of duplicate search did not find three duplicated pages (3, 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 3, 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 pages 117, 124 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 does not always mean semantically text similarity that can be validated only by human expert. The SIFT features method scores with five true positives. The calculation times of SIFT method is 95940 seconds. Experiment shows that SIFT feature matching method detected 13 false positives, whereas OCR validation method demonstrates three false positives. The OCR method for duplicate detection used for validation of SIFT analysis results demonstrates sufficient accuracy with seven correct detections among eight possible and three false positive results. The total calculation time for OCR analysis is 11418 seconds. The results of OCR analysis are very dependent on printed text and image quality, threshold setting and OCR tool quality. Texts of duplicate files extracted by OCR method can differ and require further analysis. The manual search for blank pages in the test collection shows 18 blank pages with four cover pages among them that are not fully blank and are brown colored. The automatic approach of blank pages search successfully detected all blank pages and one false positive. The OCR output score for blank pages mostly is 0 or three. Manually we also detected 13 pages with large empty areas that takes approximately the half of the document. Seven of them we were able to detect automatically but this analysis is not very reliable, since the definition of such pages is very difficult. One page (index 634) was mistakenly tagged as a blank page, whereas it is a normal text page. The reason for that could be that the quality of the text was not sufficient and OCR output size was 0. A typical text document image in SIFT analysis workflow contains up to $d = 2.000$ descriptors. Direct matching of feature descriptors requires $d^2 = 4 \cdot 10^6$ descriptor comparisons for a single pair of images. For a sample book with $n = 730$ pages $n \times (n-1) \approx 532.170$ OCR score comparisons are necessary. The text, resulting from the OCR evaluation could not be regarded as a reliable evaluation parameter for duplicate detection, due to strong dependency on image quality. But the size of this text can be successfully employed for blank page detection. The advantage of the OCR method in comparison to SIFT method is that we analyse each file only once. Therefore the OCR method presents more reliable results for blank page analysis and can be applied for quality assurance of digital collections. All of these approaches help to automatically find out duplicate and blank candidates in a huge collection. Following this, manual analysis of duplicate candidates separates real duplicates and blank pages from structural similar documents and evaluates resulting duplicate or blank pages list. Presented methods save time and therefore costs associated with human expert involvement in quality assurance process. Therefore our initial hypothesis is true. But further research is required to improve performance and accuracy metrics of mentioned methods.

The duplicates and blank pages search effectiveness can be determined in terms of a Relative Operating Characteristic (ROC). Similarity analysis divided the given docu-

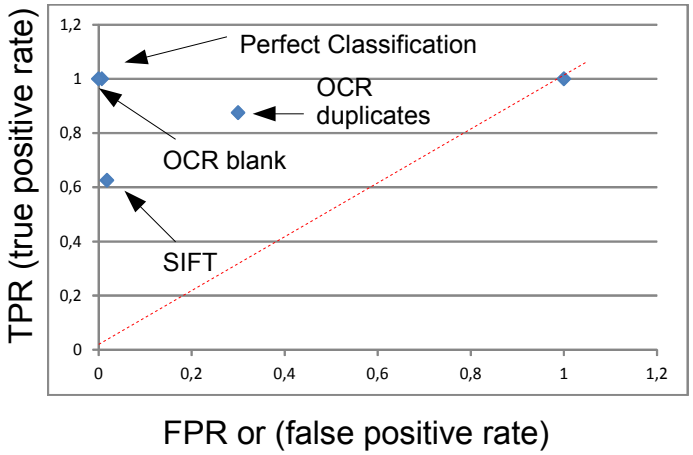

Fig. 5. ROC space plot.

ment collection (book identifier is Z151694702) in two groups "duplicates and "single or "blank" and "text/picture" by associated thresholds. The inference engine detected five true positive TP duplicates, 712 true negative TN documents, 13 false positive FP duplicates and three 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).

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

Therefore the sensitivity of the presented approach is 0.625, the FPR is 0.018. The associated ROC values for OCR blank page and duplicates detection are represented by (0.007, 1.0) and (0.3, 0.875) 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 (see Figure 5) that an automatic approach for blank page detection is very effective and it is a significant improvement compared to manual analysis. The best possible classification is represented by the point (0,1). The distribution of collection points above the red diagonal demonstrates quite good classification results that could be improved by refining of rules. Therefore, OCR analysis can be suggested as an effective method for blank page detection and as a verification step for duplicate detection.

5 Conclusion

We have presented an automatic expert system that supports decision making for blank page and accurate duplicate detection in document image collections. This system uses automatic information extraction from the image processing tools, performs analysis and aggregates knowledge that supports quality assurance process for preservation planning.

An important contribution of this paper is the definition of expert rules and creation of reliable inference engine with the solid knowledge base from the output of the image processing tools that detects blank pages and duplicates based on methods of computer

vision and OCR. We employed AI technologies (i.e. knowledge base, expert rules) to emulate reasoning about the knowledge base like a human expert.

The experimental evaluation presented in this paper demonstrates the effectiveness of employing the artificial intelligence techniques for knowledge base design and for generating reasoned suggestions. The Expert system reliably detects image sequences containing duplicated or blank images for typical text content. An automatic approach delivers a significant improvement when compared to manual analysis.

The expert system for document image collections presented in this paper ensures quality of the digitized content and supports managers of libraries and archives with regard to long term digital preservation.

As future work we plan to extend an automatic quality assurance approach of image analysis to other digital preservation scenarios. The rules could be combined with different subject categories in order to meet requirements for different use cases.

References

1. Bernard, J.: Use of a rule-based system for process control. Control Systems Magazine, IEEE 8(5), 3 –13 (oct 1988)
2. van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.: Distance measures for layout-based document image retrieval. In: 2nd ICDIAL, 2006. DIAL '06. pp. 231–242 (April 2006)
3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on SLCV, ECCV. pp. 1–22 (2004)
4. 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(6), 381–395 (1981)
5. Huber-Mörk, R., Schindler, A.: Quality assurance for document image collections in digital preservation. In: Proc. of the 14th Intl. Conf. on ACIVS (ACIVS 2012). LNCS, vol. 7517, pp. 108–119. Springer, Brno, Czech Republic (September 4-7 2012)
6. Ke, Y., Sukthankar, R., Huston, L.: An efficient parts-based near-duplicate and sub-image retrieval system. In: Proceedings of the 12th annual ACM international conference on Multimedia. pp. 869–876. MULTIMEDIA '04, ACM, New York, NY, USA (2004)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Comput. Vision 60(2), 91–110 (2004)
8. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. of the IEEE CCVPR (2007)
9. 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: The Eighth IAPR Intl. Workshop on DAS, 2008. DAS '08. pp. 585 –590 (Sep 2008)
10. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. of Computer Vision 37(2), 151–172 (2000)
11. Wu, X., Zhao, W.L., Ngo, C.W.: Near-duplicate keyframe retrieval with visual keywords and semantic context. In: Proc. of the 6th ACM ICIVR. pp. 162–169. CIVR '07, ACM, New York, NY, USA (2007)
12. Yoo, H.W., Park, H.S., Jang, D.S.: Expert system for color image retrieval. Expert Syst. Appl. 28(2), 347–357 (2005)
13. Zhao, W.L., Ngo, C.W., Tan, H.K., Wu, X.: Near-duplicate keyframe identification with interest point matching and pattern learning. Multimedia, IEEE Transactions on 9(5), 1037 –1048 (Aug 2007)