An Expert System for Quality Assurance of Document Image Collections*

Roman Graf[1], Reinhold Huber-Mörk[2], Alexander Schindler[2,3], and Sven Schlarb[4]

[1] Research Area Future Networks and Services,
Department Safety & Security,
Austrian Institute of Technology
roman.graf@ait.ac.at

[2] Research Area Intelligent Vision Systems,
Department Safety & Security,
Austrian Institute of Technology
reinhold.huber-moerk@ait.ac.at

[3] Department of Software Technology and Interactive Systems,
Vienna University of Technology
schindler@ifs.tuwien.ac.at

[4] Austrian National Library
sven.schlarb@onb.ac.at

Abstract. Digital preservation workflows for automatic acquisition of image collections are susceptible to errors and require quality assurance. This paper presents an expert system that supports decision making for page duplicate detection in document image collections. Our goal is to create a reliable inference engine and a solid knowledge base from the output of an image processing tool that detects 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. 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.

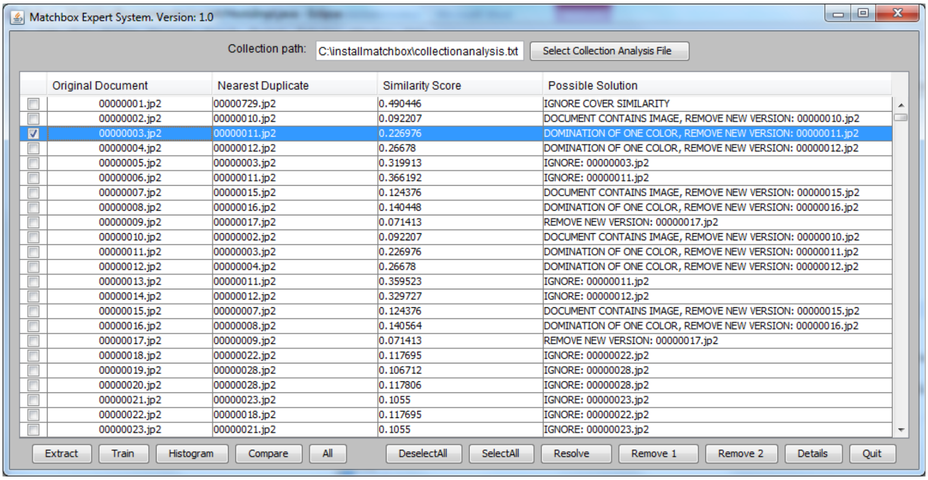
1 Introduction

Libraries and archives have carried out large-scale digitization projects. New digital collections comprising millions of books, newspapers and journals have been created. Each of the single collection items contains hundreds of document images and other information entities. Institutions are facing a paradigm shift in the manner in which preservation, maintenance, and quality assurance of these collections have to be addressed. They need automated solutions that are able to operate on the collections.

A typical task in quality assurance is an update of digitized books collections. One such project runs at the Austrian National Library in partnership with Google. This

* 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).

M. Ioannides et al. (Eds.): EuroMed 2012, LNCS 7616, pp. 251–260, 2012.


Matchbox Expert System. Version: 1.0

Collection path: C:\installmatchbox\collectionanalysis.txt

Select Collection Analysis File

	Original Document	Nearest Duplicate	Similarity Score	Possible Solution
□	00000001.jp2	00000729.jp2	0.490446	IGNORE COVER SIMILARITY
□	00000002.jp2	00000010.jp2	0.092207	DOCUMENT CONTAINS IMAGE, REMOVE NEW VERSION: 00000010.jp2
☑	00000003.jp2	00000011.jp2	0.226976	DOMINATION OF ONE COLOR, REMOVE NEW VERSION: 00000011.jp2
□	00000004.jp2	00000012.jp2	0.26678	DOMINATION OF ONE COLOR, REMOVE NEW VERSION: 00000012.jp2
□	00000005.jp2	00000003.jp2	0.319913	IGNORE: 00000003.jp2
□	00000006.jp2	00000011.jp2	0.366192	IGNORE: 00000011.jp2
□	00000007.jp2	00000015.jp2	0.124376	DOCUMENT CONTAINS IMAGE, REMOVE NEW VERSION: 00000015.jp2
□	00000008.jp2	00000016.jp2	0.140448	DOMINATION OF ONE COLOR, REMOVE NEW VERSION: 00000016.jp2
□	00000009.jp2	00000017.jp2	0.071413	REMOVE NEW VERSION: 00000017.jp2
□	00000010.jp2	00000002.jp2	0.092207	DOCUMENT CONTAINS IMAGE, REMOVE NEW VERSION: 00000010.jp2
□	00000011.jp2	00000003.jp2	0.226976	DOMINATION OF ONE COLOR, REMOVE NEW VERSION: 00000011.jp2
□	00000012.jp2	00000004.jp2	0.26678	DOMINATION OF ONE COLOR, REMOVE NEW VERSION: 00000012.jp2
□	00000013.jp2	00000011.jp2	0.359523	IGNORE: 00000011.jp2
□	00000014.jp2	00000012.jp2	0.329727	IGNORE: 00000012.jp2
□	00000015.jp2	00000007.jp2	0.124376	DOCUMENT CONTAINS IMAGE, REMOVE NEW VERSION: 00000015.jp2
□	00000016.jp2	00000008.jp2	0.140564	DOMINATION OF ONE COLOR, REMOVE NEW VERSION: 00000016.jp2
□	00000017.jp2	00000009.jp2	0.071413	REMOVE NEW VERSION: 00000017.jp2
□	00000018.jp2	00000022.jp2	0.117695	IGNORE: 00000022.jp2
□	00000019.jp2	00000028.jp2	0.106712	IGNORE: 00000028.jp2
□	00000020.jp2	00000028.jp2	0.117806	IGNORE: 00000028.jp2
□	00000021.jp2	00000023.jp2	0.1055	IGNORE: 00000023.jp2
□	00000022.jp2	00000018.jp2	0.117695	IGNORE: 00000022.jp2
□	00000023.jp2	00000021.jp2	0.1055	IGNORE: 00000023.jp2

Fig. 1. GUI of the expert system for quality assurance of document image collections

digitization project produces digital images from books through an automatic scanning process. The resulting digital collections are stored in the long term documents repository. Stored collections are maintained and constantly improved with new versions by applying image enhancement and OCR tools. Quality assurance is required in order to select between the old and the new version of the associated documents due to the high cost of storage space. A quality 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 complicates 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 proposed Expert System (see Figure 1) is based on the *matchbox* tool [6] that implements image comparison for digitized text documents.

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

The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the knowledge base 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

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.

The implementation of an expert system for color retrieval described by Yoo et al [16] 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. Image similarity is determined based on the degree of color-spatial overlap through pixel-wise comparison. In our system the similarity computation task is provided by the *matchbox* tool, which is based on SIFT [8] feature extraction.

The rule-based system presented by Bernard [3] 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.

Typically, approaches in the area of image retrieval and comparison in large collection 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 [17]. A so called bag of words (BoW) [5], derived from local descriptors, was described as an efficient approach to near-duplicate video key frame retrieval [15]. 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 [4], [11].

A state of the art with respect to technical requirements and standards in digital preservation is given by Becker et al [2]. 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.

For the assessment of evaluation results we use ground data truth, manually created by Austrian National Library experts. We designed and implemented a customized rule-based Expert System in order to find duplicates in a given collection. We compared our results to the ground truth data.

3 Knowledge Base Aggregation Process

The Knowledge Base determines the quality of the Expert System. A Knowledge Base is required in order to collect information and to perform automatic document assessment and duplicates detection. Existing information about the documents in a digital collection like file name, file size, collection name, sequence of pages is not structured

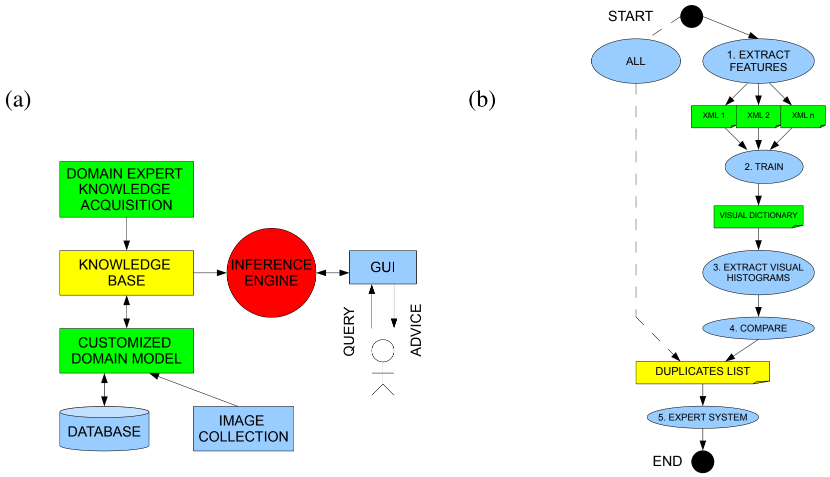


Fig. 2. Expert system: (a) overview and (b) workflow

or is only partly structured. A collection should not contain duplicates or ambiguous entries. A basis for accurate reasoning is information aggregated from digital documents and from knowledge provided by human experts.

A process of decision making for image quality assurance in digital preservation requires knowledge of image processing, file format standards and library processes. The search for such knowledge is very time consuming. A manual approach to quality assurance is often not feasible because of the large volumes of data. To facilitate the management of the Expert System we provide a user interface (see Figure 1) that supports quality assurance processes for digital preservation.

Collection analysis is conducted according the quality assurance workflow shown in Figure 2. The user triggers a complete collection analysis, the results of which are stored in a text file. In order to handle collection analysis in separate steps the user can extract document features, then 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.

3.1 Suggested Method and Rule 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. The knowledge acquisition for the Knowledge Base is performed by librarians who provide the knowledge engineer with typical application use cases, metrics and parameters that characterize the preservation processes. Information retrieved from the image collection is processed by the customized domain model. This model enables structured and maintainable handling of

Table 1. Dependency chart with interactions among the rules and associated impact factors. Rules already supported by Expert System depicted by "*".

RULES/ACTIONS	REMOVE	IGNORE	ADD	SCAN	FILTER	RENAME	RELOAD
Similarity score (*)	+	+			+		
Similarity offset max/min (*)					+		
Metadata	+	+	+		+		
File format (*)					+		
File size (*)					+		+
File name (*)					+	+	
File creation date (*)	+				+		+
Collection size (*)			+	+			+
Cover image	+	+					
Odd page			+	+			+
Fingers detected	+	+		+			+
Text/picture variation	+	+			+		
Sequence detected (*)	+	+			+	+	
Sequence distortion	+				+		
Dominating color	+	+		+	+		
Known pair (*)	+	+			+		
Old/New priority (*)	+				+		
Page number	+	+	+	+	+	+	+

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.

In order to avoid common weaknesses in rule-based systems as described by Arman [1], we generated a dependency chart showing the interactions among the rules [9]. The dependency chart helps to find potential rule problems and to keep an overview over the rules. Potential problems with rule definition and coverage are redundant rules, conflicting rules, rules that are subsumed by other rules, unreachable rules, inconsistent rules and circular rule chains. The dependency chart presented in Table 1 gives an overview about the identified rules and associated impact on the knowledge base. 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. The "Fingers detected" rule becomes significant in the case of corrupted documents which were unintentionally scanned with the scan operators fingers over the document surface and require special treatment. The issue of "Text/picture variation" means that a collection has significant variations in similarity scores caused by the mix of text and picture documents.

Another group of rules operates on image processing results. "Sequence detected" leverages the fact that duplicates mostly originate from automatically performed scan runs resulting in a sequence of documents. Typically, isolated duplicates can be omitted from duplicate analysis, as it was observed than duplicates always occur in bursts, a fact justified by the specifics of the scan machine operation mode. The "Sequence distortion" rule describes the case of the sequence in which there is a gap of a single page image, which caused by the poor quality of the version of the image duplicated at the end of the sequence. The “Odd page” rule can also be used to efficiently filter a collection. The automatic scan run assumes that always two pages of a book are scanned together. Therefore, it is not possible to have an odd duplicate of the even page. Additional information yields a rule “Page number”. With the help of OCR techniques one could attempt to find out the page number in scanned document images, which is topic of future work.

Possible actions according to the advice provided by the Expert System include removing of a document, ignoring the advice, adding a new document, performing a new scan for the particular image, filtering a collection by different filter arguments, renaming a document and reloading a document or a collection including similarity analysis.

3.2 Image Processing

In cases of geometric modifications as well as filtering, color or tone modifications, the information at the image pixel level might differ significantly, although the image content is well-preserved. Therefore, we use interest point detection and derivation of 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 keypoint extraction and description approach [8]. 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.

Learning the visual dictionary is performed using a clustering method applied to all SIFT descriptors of all images, which 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 [10], 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.

3.3 Algorithmic Details

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 detection is presented in Figure 3. Forward chaining 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 detection there is no need for any conflict-resolution procedure to resolving possible rule conflicts.

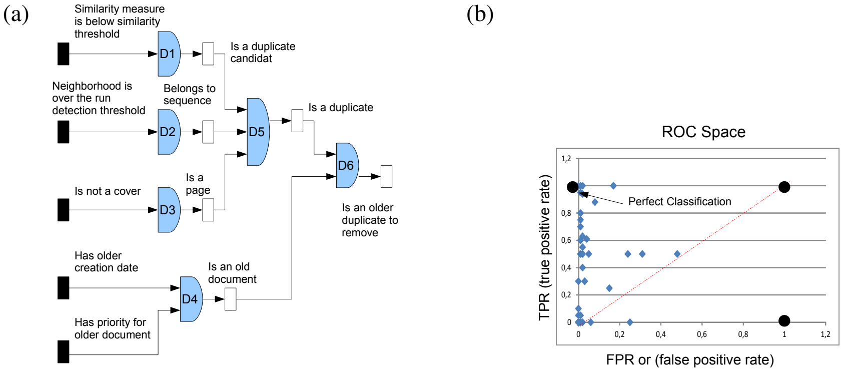


Fig. 3. Evaluation: (a) forward rule chaining for duplicate detection, (b) ROC space plot

In Figure 3 we present rules distinguishing duplicated pages from non-duplicated ones. The rule-base system starts duplicate identification with the rule D1. Suppose that similarity measure of particular document is below similarity threshold. Then if the antecedent pattern matches that assertion, the value x becomes “is a duplicate candidate” and the rule D1 fires. Because the document belongs to a sequence, rule D5 fires, establishing that the document “is a duplicate”. Similarly we go through remaining rules. The final conclusion of the rule-based system is whether there is a duplicated page observed. The inference engine performs conditional rules and similarity score analysis, infers appropriate action and formulates advice using relation of linguistic labels to specific consequences.

4 Evaluation

The suggested Expert System processes reasons on found duplicates pairs and generates advice on how to clean up the collection. Our hypothesis is that automatic approach should be able to detect duplicates with reliable quality. Then this method would be a significant improvement over a manual analysis.

4.1 Collection Analysis

The considered collection contains 730 documents corresponding to a single book. Manually created ground truth was available.

Table 2 lists all duplicates pairs (MV1; MV2) that were discovered by an expert. Duplicated images automatically inferred by the Expert System are denoted with (AV1; AV2), whereas V1 and V2 stand for the original and new version of the document. The distance column presents the number of pages between the original and the new version of the duplicated documents in the collection. We employ the rules from Table 1. In the first instance we apply “Similarity score” rule. For this rule we compute average similarity score over the all similarity scores provided as an output of the matchbox tool. In conjunction with similarity offset rule we are able to isolate most of the duplicate pairs. The “Cover image” rule emerges from the fact, that the first and the last documents in

Table 2. Manually and automatically detected duplicates

MV1	MV2	Distance	AV1	AV2	Distance
2	10	8	2	10	8
3	11	8	3	11	8
4	12	8	4	12	8
5	13	8	-	-	0
6	14	8	-	-	0
7	15	8	7	15	8
8	16	8	8	16	8
9	17	8	9	17	8
-	-	0	108	118	10
-	-	0	109	119	10
-	-	0	110	120	10
-	-	0	111	121	10
-	-	0	112	116	4
-	-	0	113	125	12
-	-	0	114	126	12
-	-	0	115	127	12
-	-	0	117	125	8
-	-	0	124	112	12

collection often are the front and back cover and are often wrongly identified as pair of duplicates. This case should be ignored. 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. Visual histograms comparison by color could provide additional help for working with these documents.

4.2 Effectiveness of the Duplicates Search

As presented in Table 2, the automatic approach of duplicate search did not find two duplicates which were identified as duplicates by manual analysis. The corresponding documents 5 and 6 obtained similarity scores 0.319 and 0.366, respectively. The reason for that is the computed average similarity score was 0.0635 with the offset 0.01. The resulting threshold is smaller than the scores of documents 5 and 6. In this specific case we have to deal with nearly empty documents with dominating white color, which makes it difficult to identify these documents as a pair of duplicates.

The documents in range from 108 to 115 and document 116 are detected as false positives by the automatic analysis. As distinguished from the case dominating color here similarity scores are in range. But manual checking of mentioned documents confirms that there are no duplicates. As we can see high image similarity not always means semantic equality. We could tackle this issue applying additional refining rules like “Sequence detected”, “Sequence distortion” or “Page number”. As depicted in Table 2 the distances between pretended duplicates vary between 4 and 12. Therefore there is no proper sequence detected for these documents. The gap in the supposed sequence between 115 and 117 boosted by mixed numbers of the new version (AV2: 121-116-125) proves that we face no duplicates in this case.

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 151694702) in two groups ”duplicates” and ”single” by similarity threshold. The inference engine 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).

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

Therefore the sensitivity of the presented approach is 0.75, the FPR is 0.012. The ROC space demonstrates that the calculated FPR and TPR values form a point (0.012, 0.75) that is located very close to the so called perfect classification point (0, 1). These results demonstrate (see Figure 3b) that an automatic approach for duplicates detection is effective and it is a significant improvement compared to manual analysis. Some of the rules are not yet implemented in the algorithm of the Expert System or matchbox but have a potential to improve the quality of evaluation. The effectiveness of the approach could be improved in the future when all rules are implemented. We evaluated the ROC space plot (Figure 3) on a set of 58 digital books containing 34.805 high-resolution scans of book pages. Each point is defined by FPR and TPR rates of associated collection. 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.

A typical text document image contains $d = 40.000$ descriptors on average. Matching two images based on the BoW representation requires a single vector comparison. For a typical book $n \times (n-1) \approx 350.000$ vector comparisons are necessary. In contrast, direct matching of SIFT descriptors requires $d^2 = 1.6 \cdot 10^9$ vector comparisons for a single pair of images. Therefore, direct SIFT matching is only suggested as a verification step. The total time for duplicate detection for a typical book from the considered collection takes 3000 seconds on an Intel Core2 Quad (4 cores at 2.66GHz) computer. The average relative computational costs are 13 percent for feature extraction, 43 percent for BoW construction and 54 percent for actual comparison.

5 Conclusion

We have presented an expert system that supports decision making for duplicate detection in document image collections. This system uses automatic information extraction from the *matchbox* tool, performs analysis and aggregates knowledge that supports decision making for preservation planning.

An important contribution of this paper is the creation of reliable inference engine and the solid knowledge base from the output of the matchbox tool that detects duplicates based on methods of computer vision. We employed AI technologies (i.e. knowledge base, expert rules) to emulate reasoning about the knowledge base like a human expert. We designed a user friendly GUI that provides functionality to enact collection analysis and to execute generated reasoning suggestions.

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 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. We also plan to improve the quality of suggestions and the usability of the user interface.

References

1. Arman, N.: Fault detection in dynamic rule bases using spanning trees and disjoint sets. Int. Arab J. Inf. Technol. 4(1), 67–72 (2007)
2. Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., Hofman, H.: Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International Journal on Digital Libraries 10(4), 133–157 (2009)
3. Bernard, J.: Use of a rule-based system for process control. IEEE Control Systems Magazine 8(5), 3–13 (1988)
4. van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.: Distance measures for layout-based document image retrieval. In: 2nd ICDIAL, DIAL 2006, pp. 231–242 (April 2006)
5. 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)
6. Huber-Mörk, R., Schindler, A.: Quality Assurance for Document Image Collections in Digital Preservation. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P., Zemčík, P. (eds.) ACIVS 2012. LNCS, vol. 7517, pp. 108–119. Springer, Heidelberg (2012)
7. 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, MULTIMEDIA 2004, pp. 869–876. ACM, New York (2004)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Comput. Vision 60(2), 91–110 (2004)
9. Nguyen, T.A., Perkins, W.A., Laffey, T.J., Pecora, D.: Checking an expert systems knowledge base for consistency and completeness. In: Proc. of the 9th IJCAI, IJCAI 1985, vol. 1, pp. 375–378. Morgan Kaufmann Publishers Inc., San Francisco (1985)
10. 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)
11. 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, DAS 2008, pp. 585–590 (September 2008)
12. Schlarb, S., Michaelar, E., Kaiser, M., Lindley, A., Aitken, B., Ross, S., Jackson, A.: A case study on performing a complex file-format migration experiment using the planets testbed. IS&T Archiving Conference 7, 58–63 (2010)
13. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. of Computer Vision 37(2), 151–172 (2000)
14. Strodl, S., Becker, C., Neumayer, R., Rauber, A.: How to choose a digital preservation strategy: evaluating a preservation planning procedure. In: JCDL 2007: Proceedings of the 2007 Conference on Digital Libraries, pp. 29–38. ACM, New York (2007)
15. 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, CIVR 2007, pp. 162–169. ACM, New York (2007)
16. Yoo, H.W., Park, H.S., Jang, D.S.: Expert system for color image retrieval. Expert Syst. Appl. 28(2), 347–357 (2005)
17. 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 (2007)