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

Roman Graf, Ross King

Research Area Future Networks and Services

Department Safety & Security, AIT Austrian Institute of Technology

Sven Schlarb

Austrian National Library, Vienna, Austria

APA / C-DAC Conference, 2014
New Delhi, India, 4-6 February 2014

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).

Overview

- SCAPE project
- Introduction
- QA challenges of document image collections
- The duplicate and blank page detection process
  - Expert rules identification
  - Image processing
  - Duplicate detection workflow
- Experimental evaluation
  - Hypothesis and evaluation methods
  - Experimental results and its interpretation
- Conclusion

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).

SCAPE project

- Scalable Preservation Environments
- Scalable services for planning and enactment of institutional preservation strategies
- Semi-automated workflows using large-scale collections of complex digital objects
- Services help to
  - Identify the need to perform preservation actions
  - Define preservation plans
  - Automated and scalable processing
  - Monitor the quality of preservation process

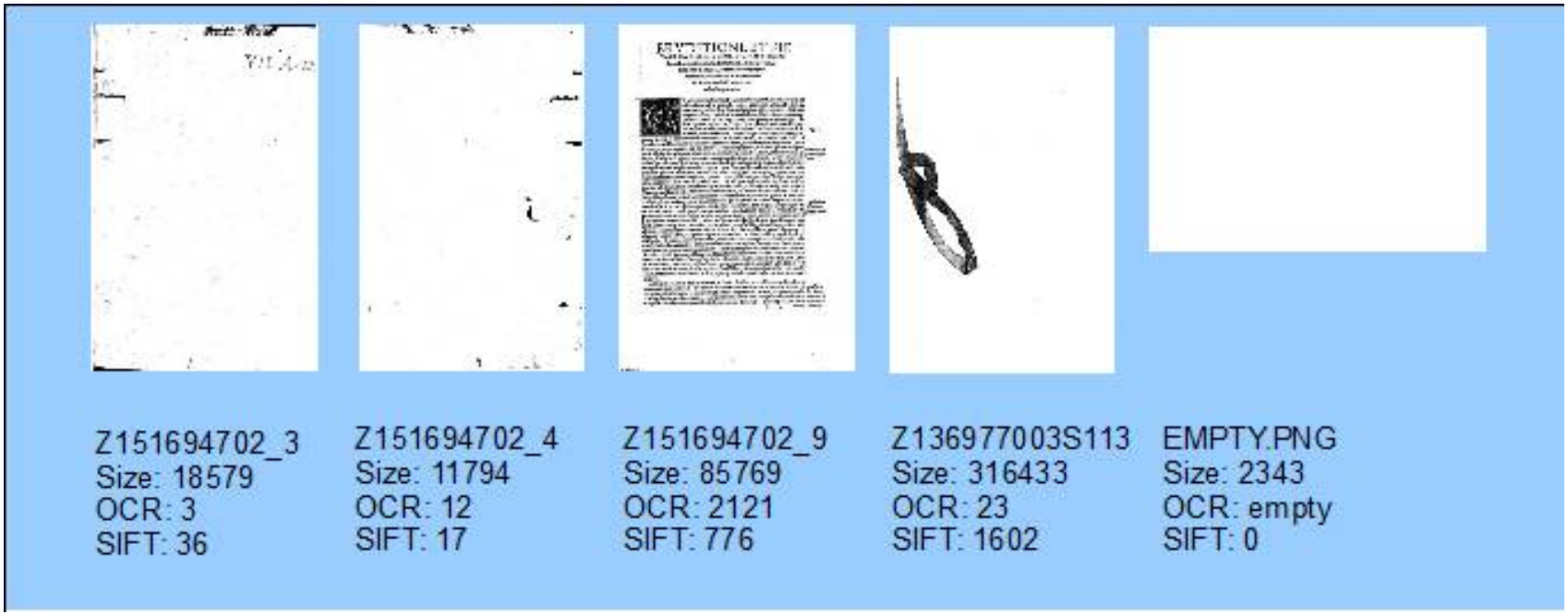
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Introduction

- Digitization workflows for automatic acquisition of image collections are susceptible to errors and require quality assurance.
- Automatic expert system for long term preservation - decision making for blank page and duplicate.
- Definition of the expert rules with associated severity level and its automatic computation.
- Inference engine from the image processing tool based on methods of computer vision.
- To improve analysis accuracy we use OCR tool for blank page and duplicate detection.
- Statistical analysis of the aggregated information.

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).

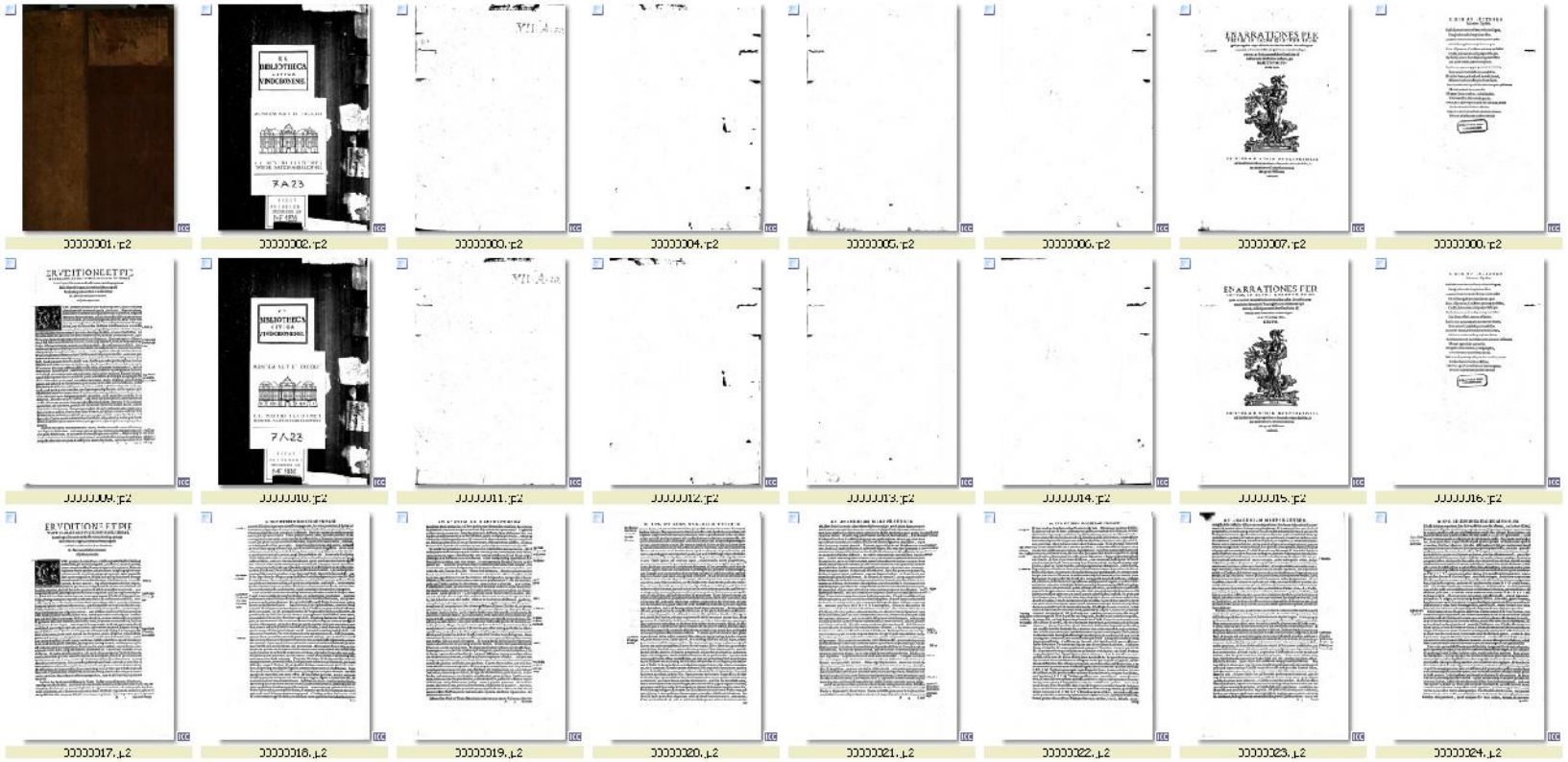
Introduction



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.

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).

Introduction



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)

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).

QA challenges of document image collections

- Large-scale digitization projects
- QA of document image collections increasingly important
- Manual maintenance and QA - time consuming, high personal and storage costs
- Need for automated solutions
- Typical QA task - update of digitized books collections
- ÖNB - automatic scanning process
- Stored collections are maintained and constantly merged with new versions (OCR): old/new version
- Stored data is not structured
- Decision support system is required (lack expertise, huge set)

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).

Duplicate and blank page detection workflow goals

- Manual search not possible.
- A consistent collection should not contain duplicates or blank pages.
- Support decision making regarding the collection cleaning
- Challenge - Information not structured.
- Collect information and to perform automatic document assessment and duplicates detection.
- Basis is information aggregated from digital documents and from knowledge provided by human experts.

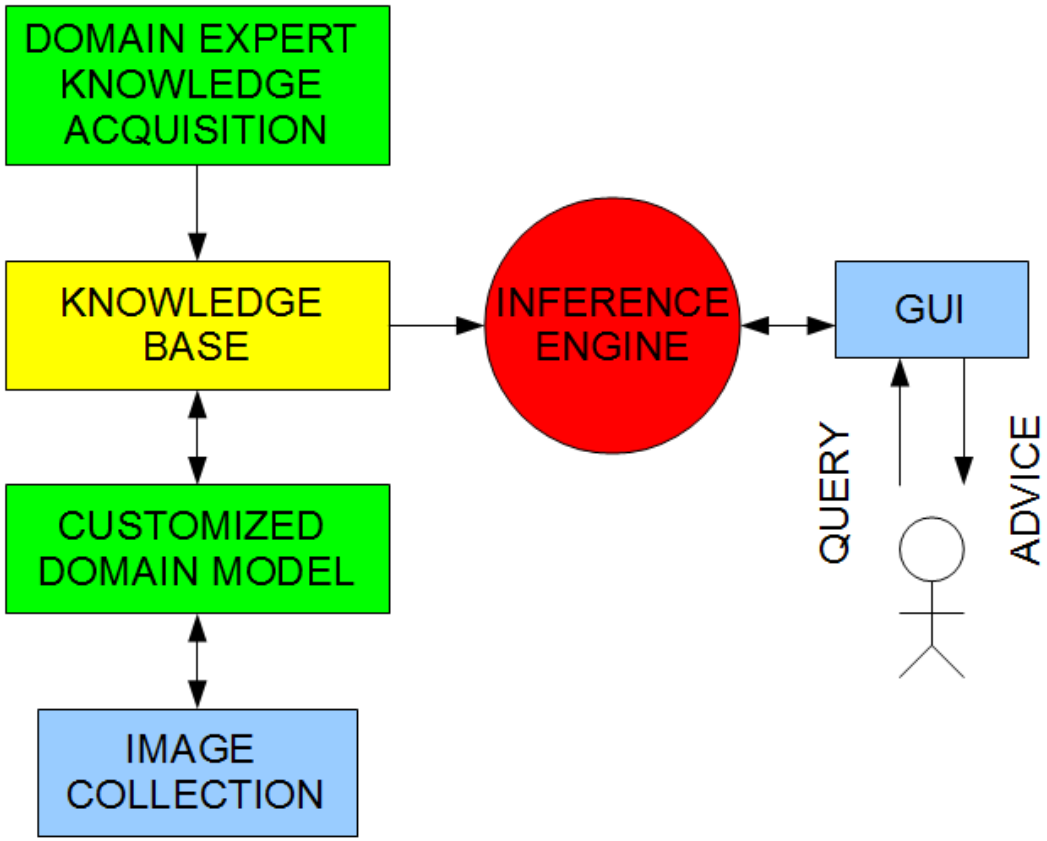
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Expert rules identification

- Structure the information from the domain experts of digital preservation and from conducted experiments.
- Define typical scenarios and identify the parameters used by library experts for collection handling.
- Define the linguistic labels to classify measured values.
- 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 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).

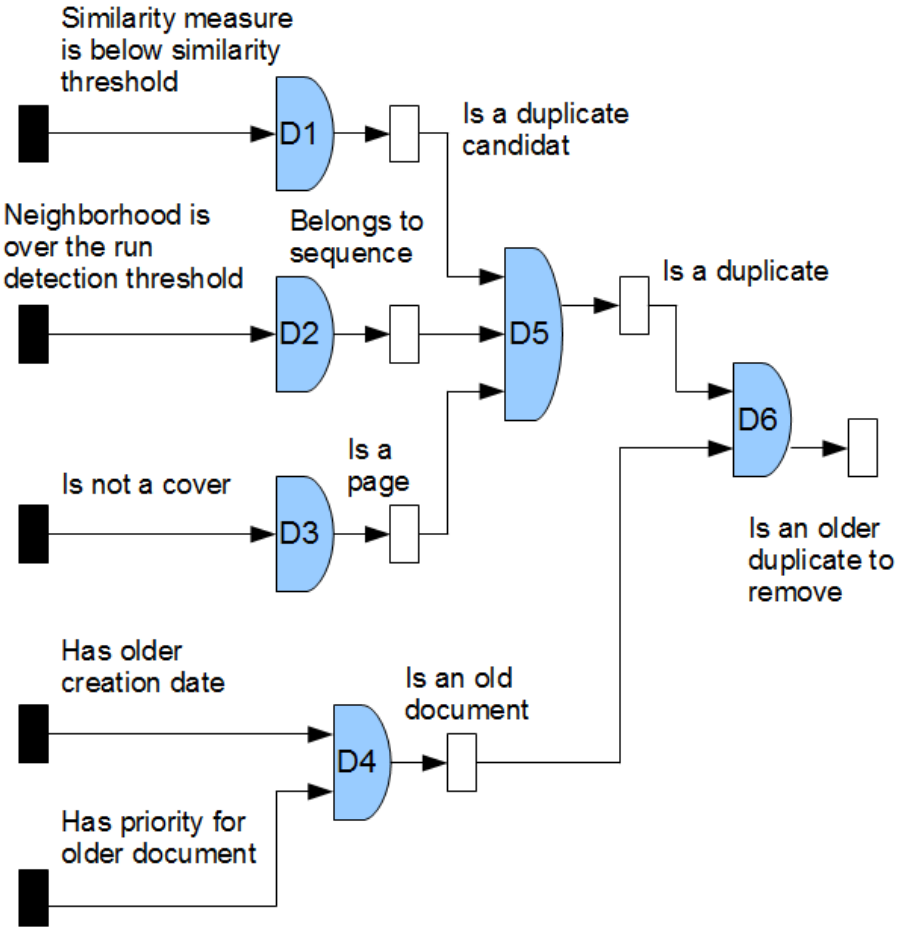
Expert rules identification



Expert system overview.

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).

Expert rules identification



Forward rule chaining of expert system for duplicate detection.

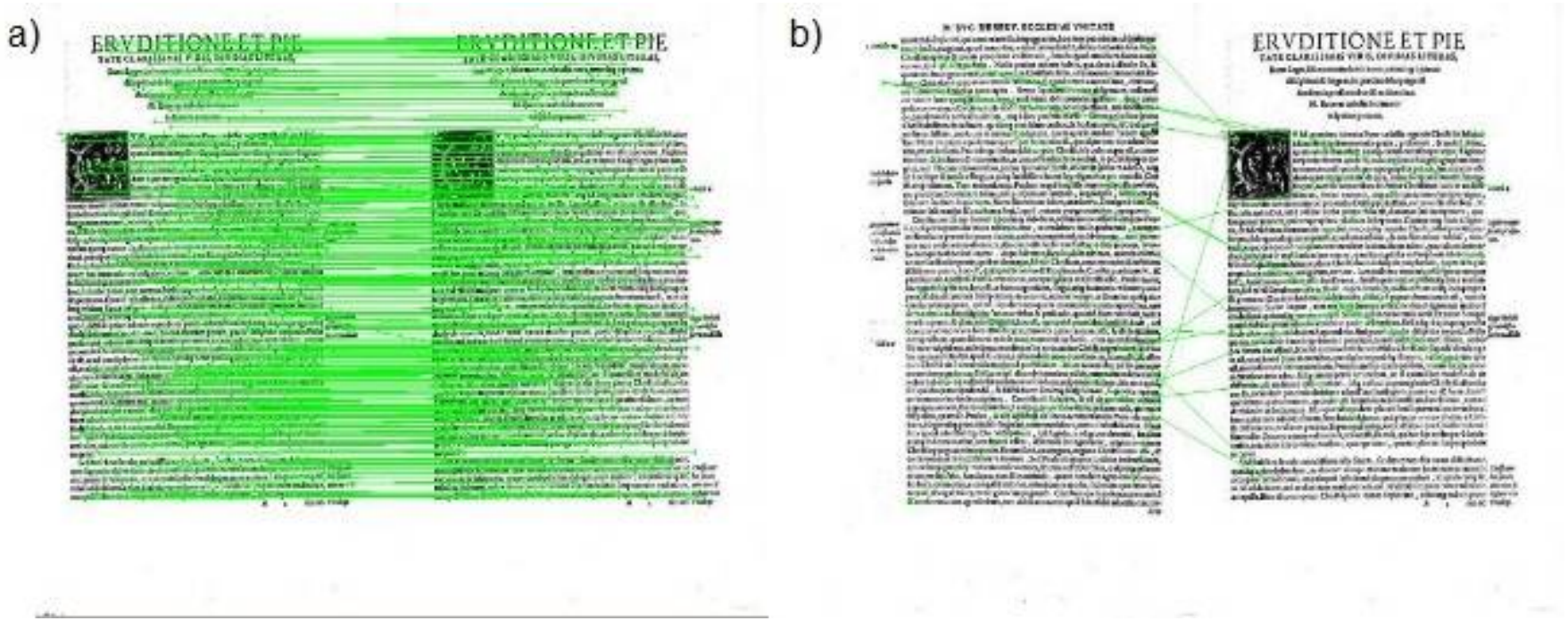
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Image processing

- Matchbox tool (new, innovative) implements image comparison for digitized text documents
- Matchbox tool (SIFT feature extraction, BoW)
  - interest point detection
  - local feature descriptors
  - invariant to geometrical and radiometrical distortions
  - preclustering of descriptors
- SIFT descriptor matching
- OCR limited with respect to accuracy and flexibility
- More descriptors – more accurate – better quality

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).

Image processing



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.

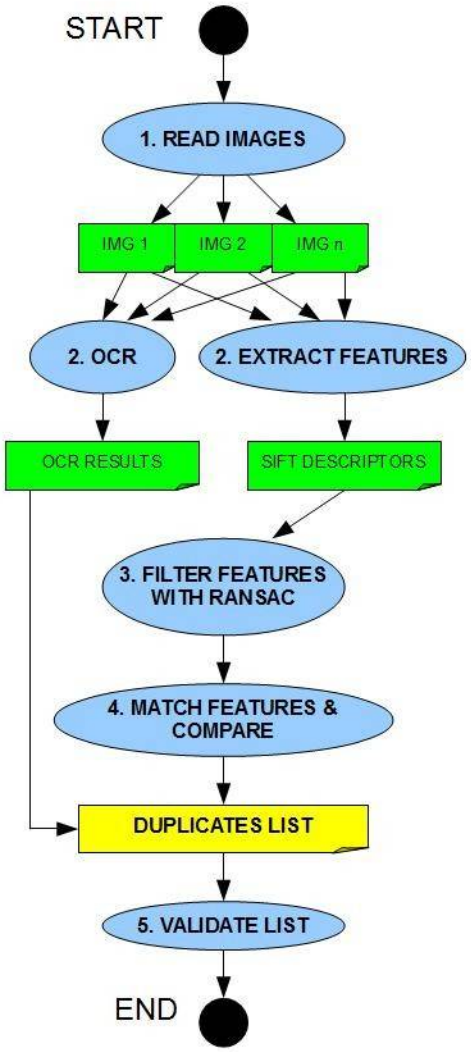
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Duplicate detection process

- Document feature extraction
  - Interest keypoints - Scale Invariant Feature Transform (SIFT)
  - Local feature descriptors (invariant to geometrical distortions)
  - Robust descriptor matching employs the RANSAC algorithm
- Learning visual dictionary
  - Clustering method applied to all SIFT descriptors of all images using k-means algorithm
  - Collect local descriptors in a visual dictionary using Bag-Of-Words (BoW) algorithm
- Create visual histogram for each image document
- Detect similar images based on visual histogram and local descriptors. Structural SIMilarity (SSIM) approach
  - Rotate
  - Scale
  - Mask
  - Overlaying
- Analysis results stored in text file
- Human expert validates the list of duplicate candidates

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).

Duplicate detection process



Duplicate detection workflow of expert system.

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).

Experimental evaluation

- Goal – two use cases for collection analysis for duplicates and blank page detection
- Expert System processes reasons on found blank pages and duplicates and generates advice on how to clean up the collection.
- Hypothesis: automatic approach should be able to detect blank pages and duplicates with reliable quality.
  1. Duplicate detection. The OCR analysis should prove the results of image processing methods and OCR scores for similar files should have similar OCR scores.
  2. Blank page detection. OCR scores should be null or near to null as well as SIFT descriptors score should be very low.
- Evaluate whether file size of blank page could be a reliable parameter for blank page analysis.
- Evaluation data set:
  - Austrian National Library collection with identifier Z151694702 (730 documents, associated ground truth)
- Significant improvement over a manual analysis
- We evaluate duplicate candidate pairs, calculation time and calculation accuracy for each evaluation method.
- Intel Core i7-3520M 2.66GHz computer using Java 6.0, Python 2.7 and C++ languages on Linux OS, for OCR analysis we use Tesseract 3.02 tool

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).

Experimental results and its interpretation

- The threshold value 0.9 was determined using statistical approach and robust estimators.
- 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 (computed average similarity score was higher than the scores)
- 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. 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.
- SIFT 5 true positives (95940 seconds), 13 false positives
- OCR 3 false positives - demonstrates sufficient accuracy with 7correct detections among 8 possible and 3 false positive results (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.
- Manually 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 3. Manually we also detected 13 pages with large empty areas that takes approximately the half of the document. 7 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.

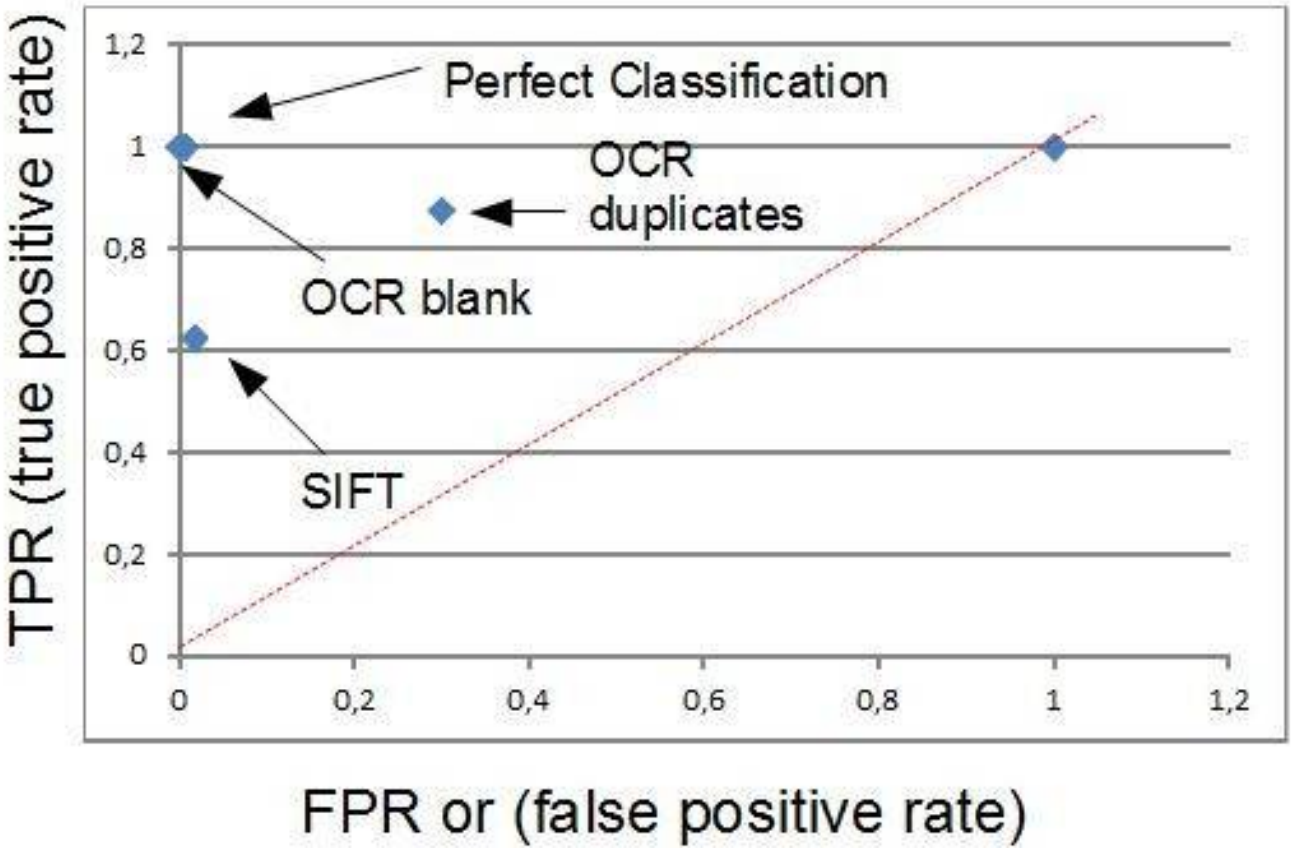
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Experimental results and its interpretation

- Typical text document image in matchbox workflow contains up to d=40.000 descriptors.
- 2000 descriptors on average for SIFT method
- 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 OCR vector comparisons are necessary.
- In contrast, direct matching of feature descriptors requires $d^2 = 4 \times 10^6$ vector comparisons for a single pair of images.
- $O(n^2)$ instead of $O(n^2 * d^2)$ in original space (SIFT matching), where n – number of pages, d - descriptors
- Direct feature matching is much more computationally intensive but its workflow is simpler then matchbox implementation.
- The average relative computational costs for matchbox workflow are 53 percent for feature extraction, 28 percent for BoW construction and 19 percent for actual comparison.
- Matchbox tool demonstrates the best detection accuracy combined with relative good performance
- 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**. **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. **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. Our initial hypothesis is true.

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).

Experimental results and its interpretation



Relative Operating Characteristic (ROC) space plot

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).

Matchbox Tool Features

- Reduce costs
- Improves quality
- Saves time
- Automatically
- Increase efficiency of human work with particular focus
- Invariant to format, rotation, scale, translation, illumination, resolution, cropping, warping, distortions
- Application: assembling collections, missing files, duplicates, compare two images independent from format (profile, pixel)

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).

Conclusion

- Automatic expert system that supports decision making for blank page and accurate duplicate detection in document image collections.
- Use automatic information extraction from the image processing tools, performs analysis that supports quality assurance process for preservation planning.
- Definition of expert rules and creation of reliable inference engine with the knowledge base from the output of the image processing tools that detects blank pages and duplicates based on methods of computer vision and OCR.
- 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.

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).

Future work

- 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.
- Further research is required to improve performance and accuracy metrics of mentioned methods.

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).

Thank you for your attention!