AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

SCAPE
Scalable Preservation Environments

Quality Assurance Tools for Document Image Collections

Roman Graf
Research Area Future Networks and Services
Department Safety & Security, AIT Austrian Institute of Technology

SCAPE Information Day at the Austrian National Library
Vienna, 5 May 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

- Introduction
- QA challenges of document image collections
- The Matchbox tool for duplicate page detection
- The blank page detection method
- The tool for finger detection on scans
- The tool for cropping error 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).

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).

SCAPE project

1. **Matchbox tool** for accurate duplicate detection in document image collections is a modern quality analysis tool based on SIFT feature extraction.
2. **Blank page detection tool** that employs different image processing techniques and OCR.
3. **Finger detection tool** for automatic detection of fingers that mistakenly appear in scans from digitized image collections. This tool uses modern image processing techniques for edge detection, local image information extraction and its analysis for reasoning on scan quality.
4. **Cropping error detection tool** supports the analysis of digital collections (e.g. JPG, PNG files) for detecting common cropping problems such as text shifted to the edge of the image, unwanted page borders, or unwanted text from a previous page on the image.

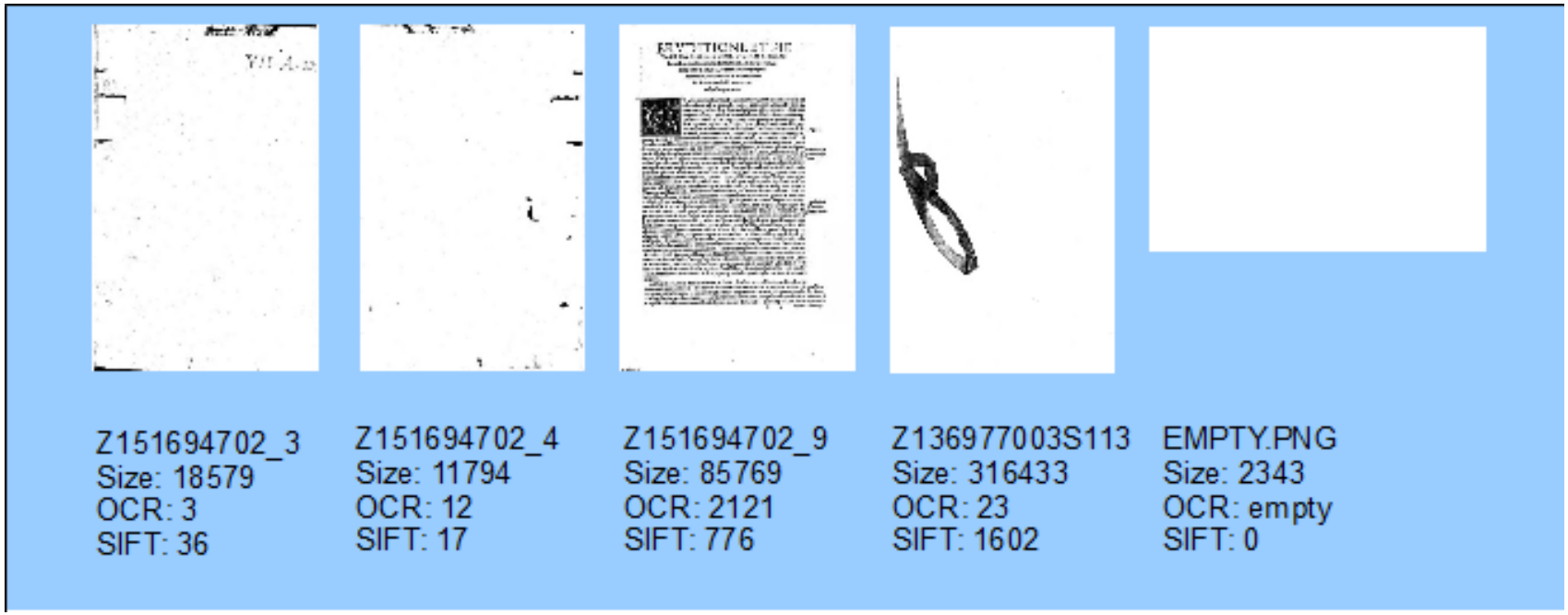
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

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 pages, cropping errors, finger on scans and duplicates.
- Definition of the expert rules with associated severity level and its automatic computation.
- Inference engine from the image processing tools based on methods of computer vision.
- 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).

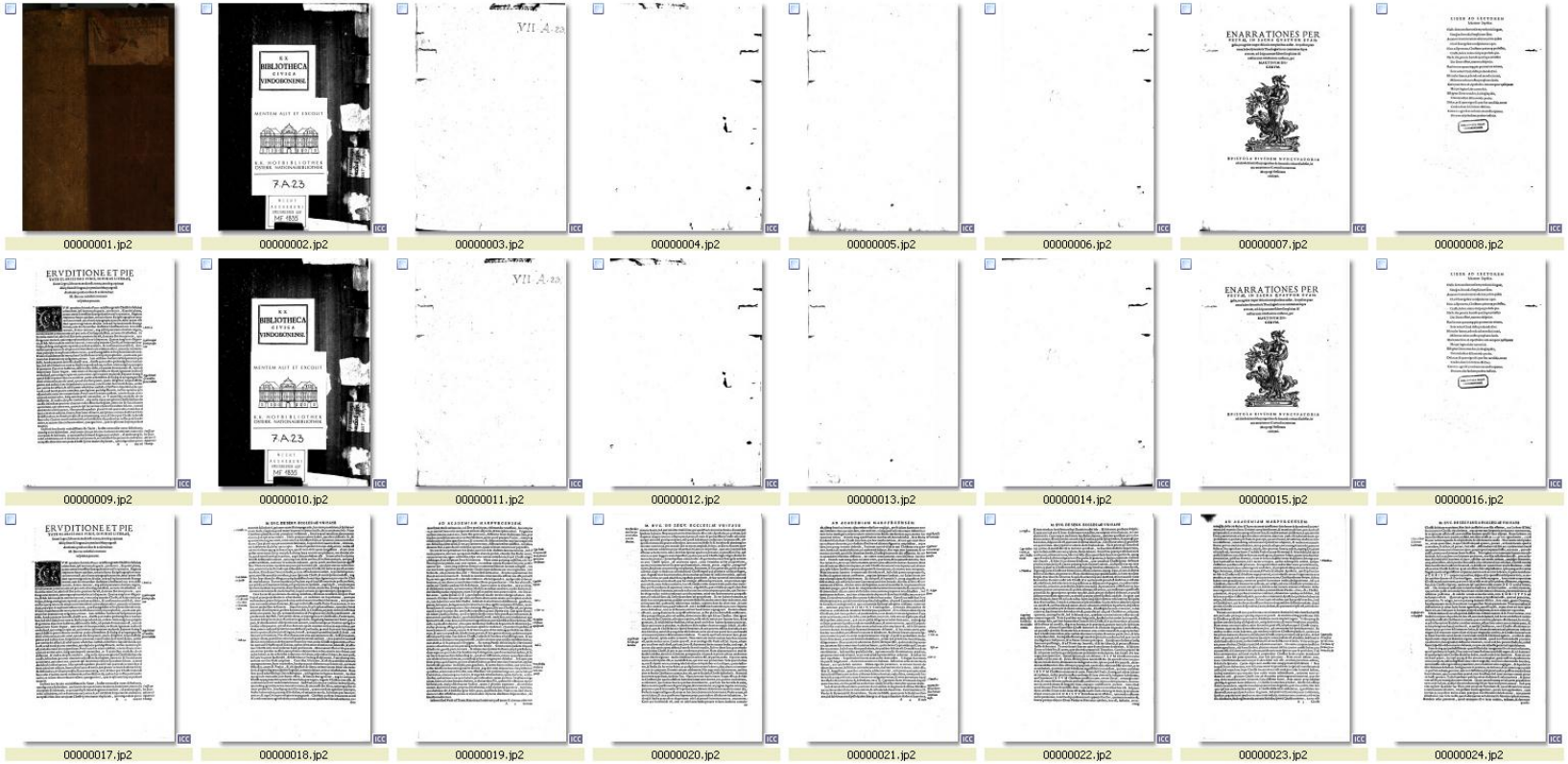
Introduction



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).

High similarity but no duplicates


00000108.jp2


00000109.jp2


00000110.jp2


00000111.jp2


00000118.jp2


00000119.jp2


00000120.jp2


00000121.jp2

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.
- Image: d = 40.000 SIFT descriptors, book: n = 700 images
- SIFT: $d^2 = 1.6 \times 10^9$ vector comparisons for a single pair of images
- BoW typical book: clustering, $n \times (n - 1) = 350.000$ vector comparisons

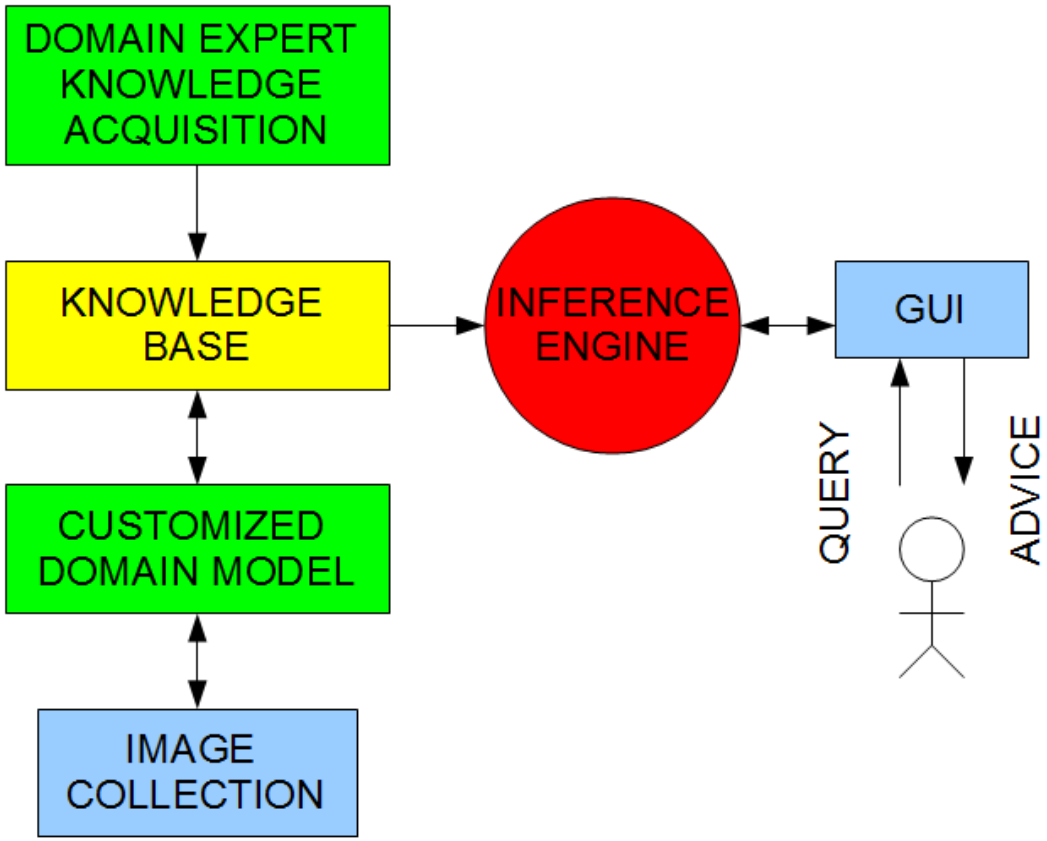
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

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).

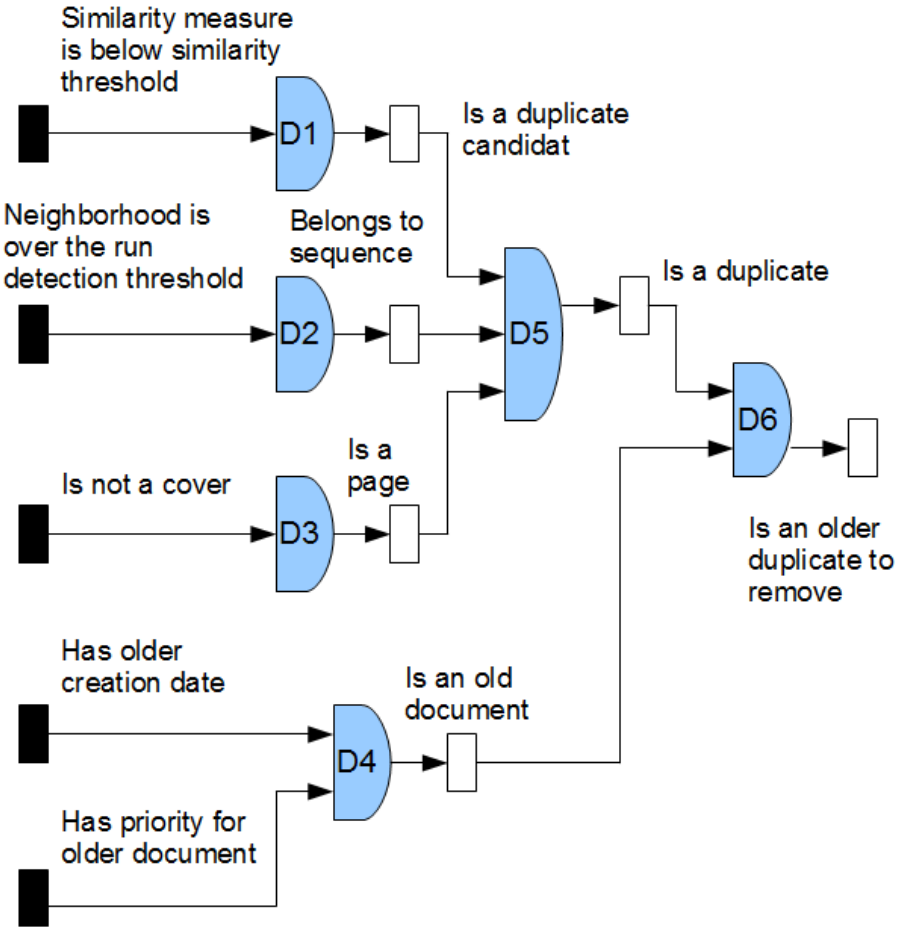
Expert rules identification



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.

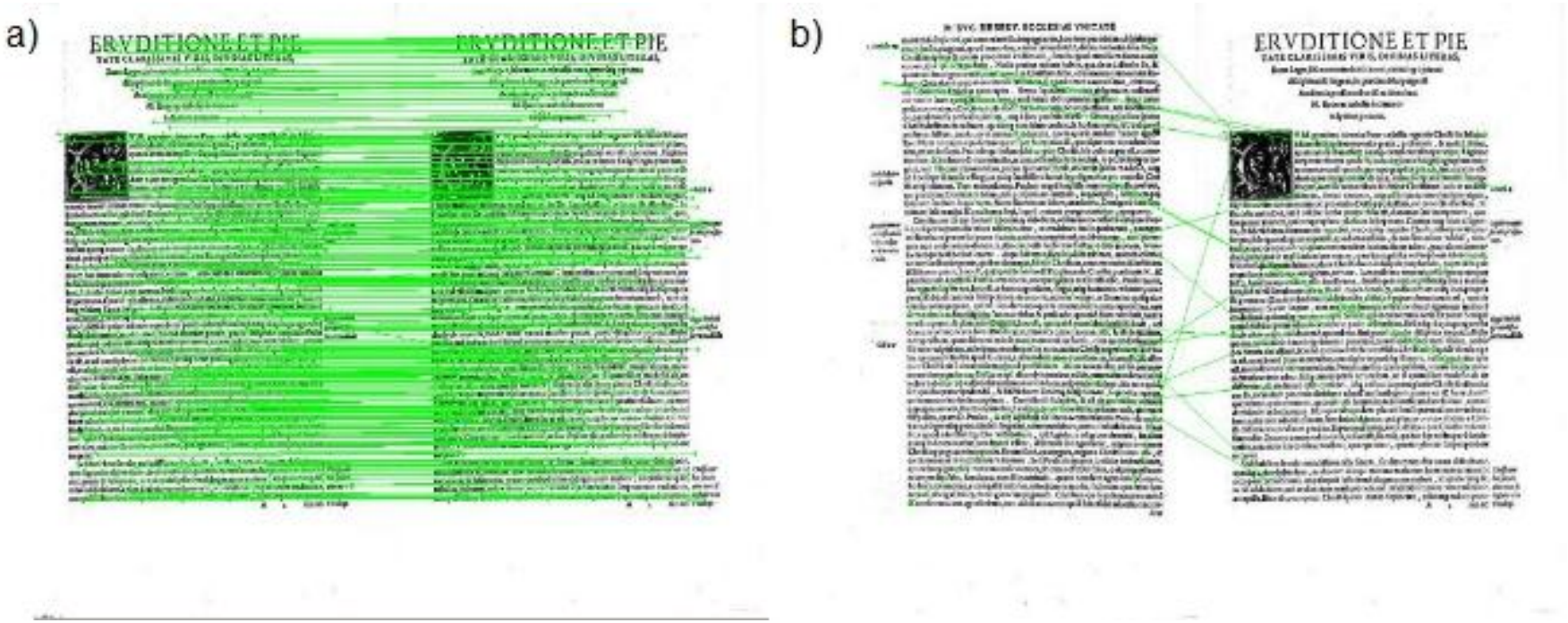
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

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).

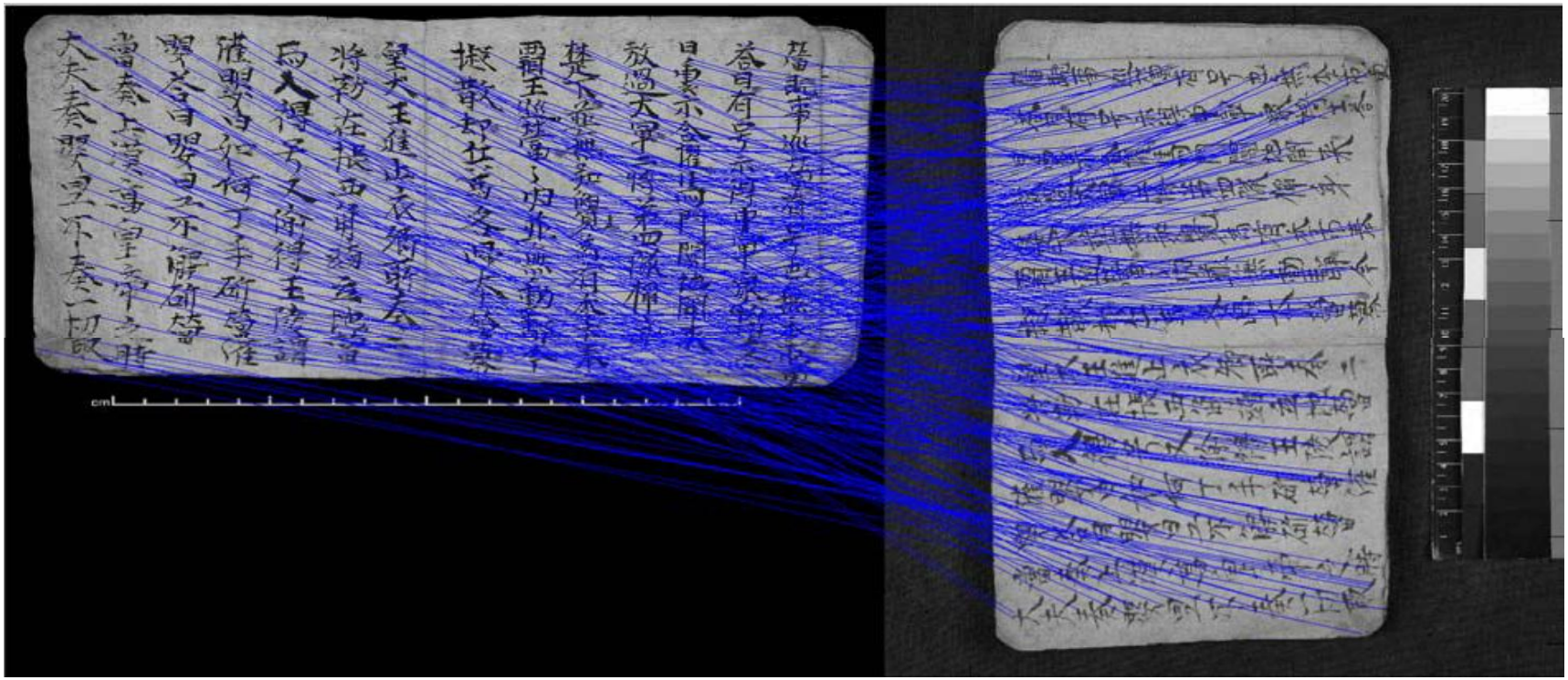
Image processing



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).

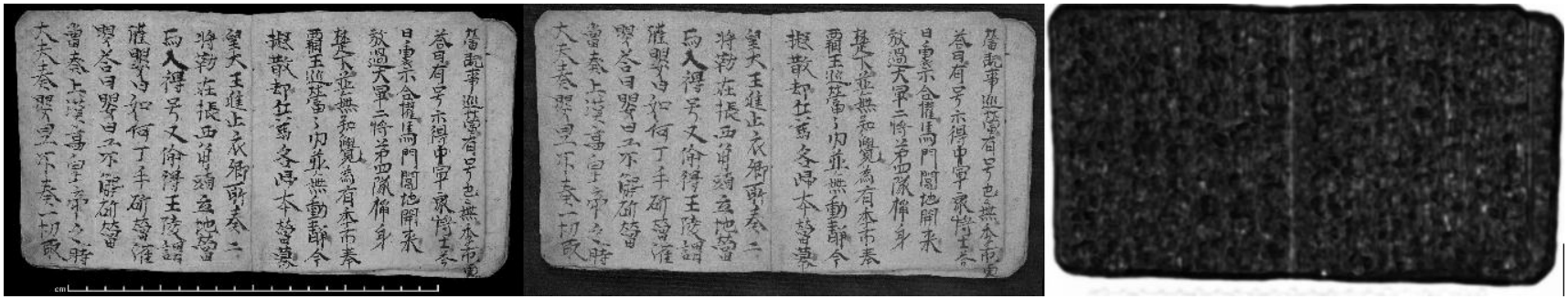
Image processing



Matching of keypoints for chinese scripts

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).

Pixel wise comparison - SSIM

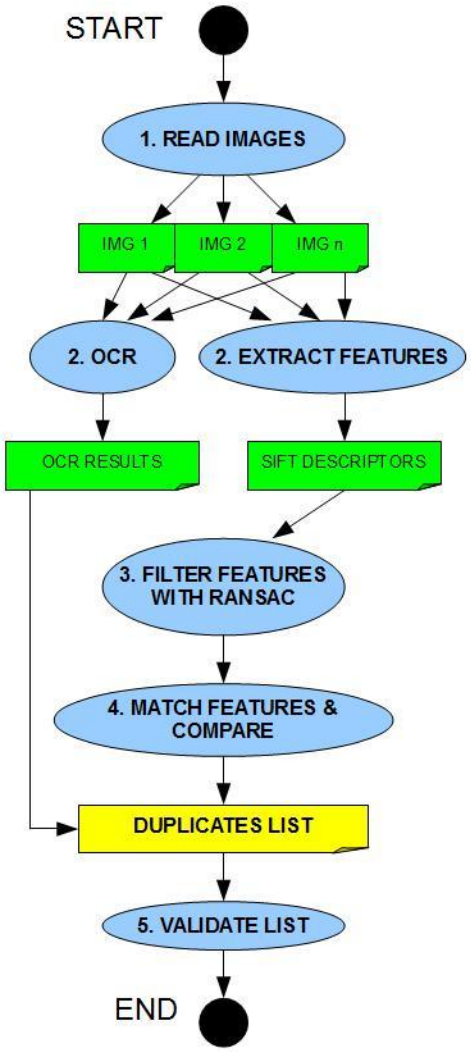


Duplicate detection process

- 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).

Image comparison tool

- Visually similar images detection
- Duplicated images
- Added images
- Missing images

Application fields

- File format
- Different color information
- Scale
- Rotations
- Different resolution
- Different cropping
- Slight differences in content

Experimental evaluation

- 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

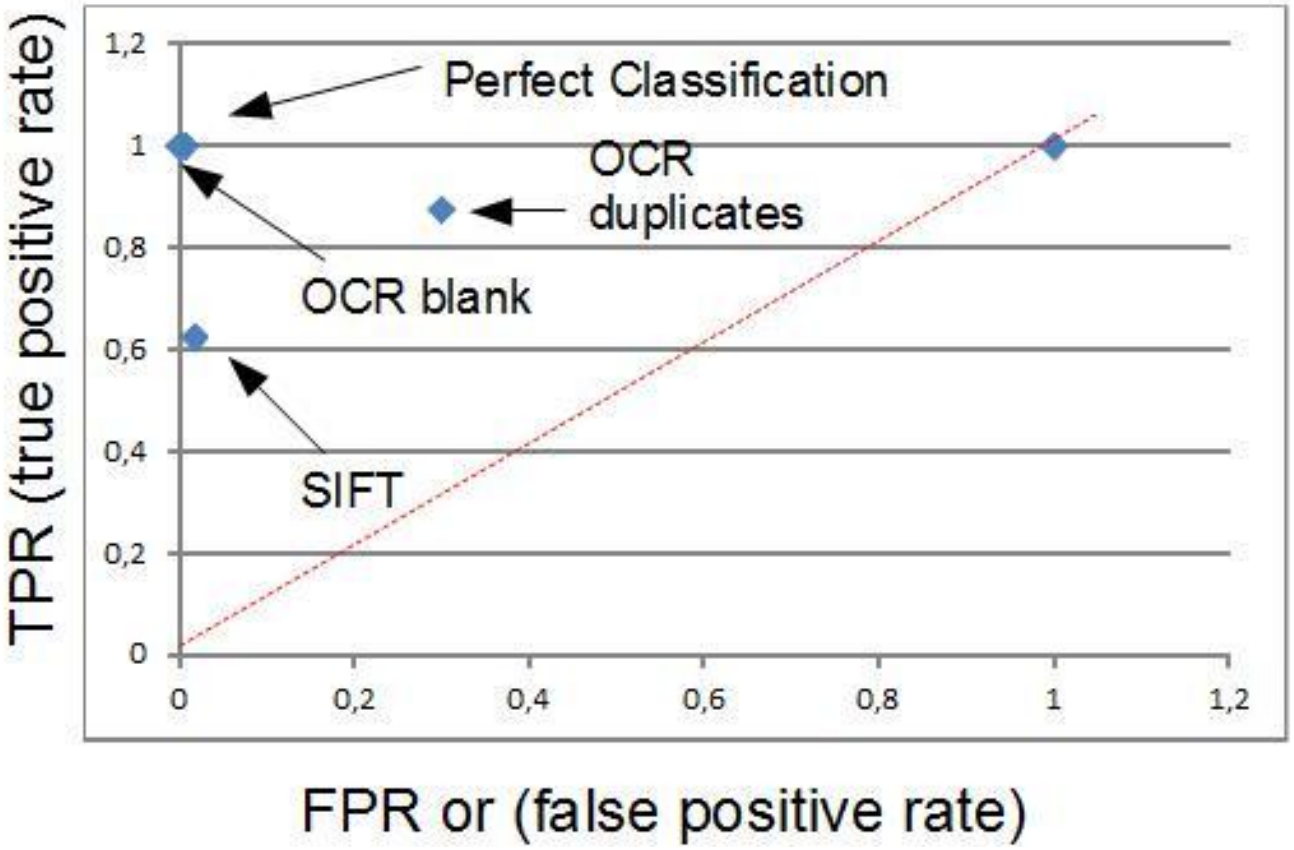
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Experimental results and its interpretation

- 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

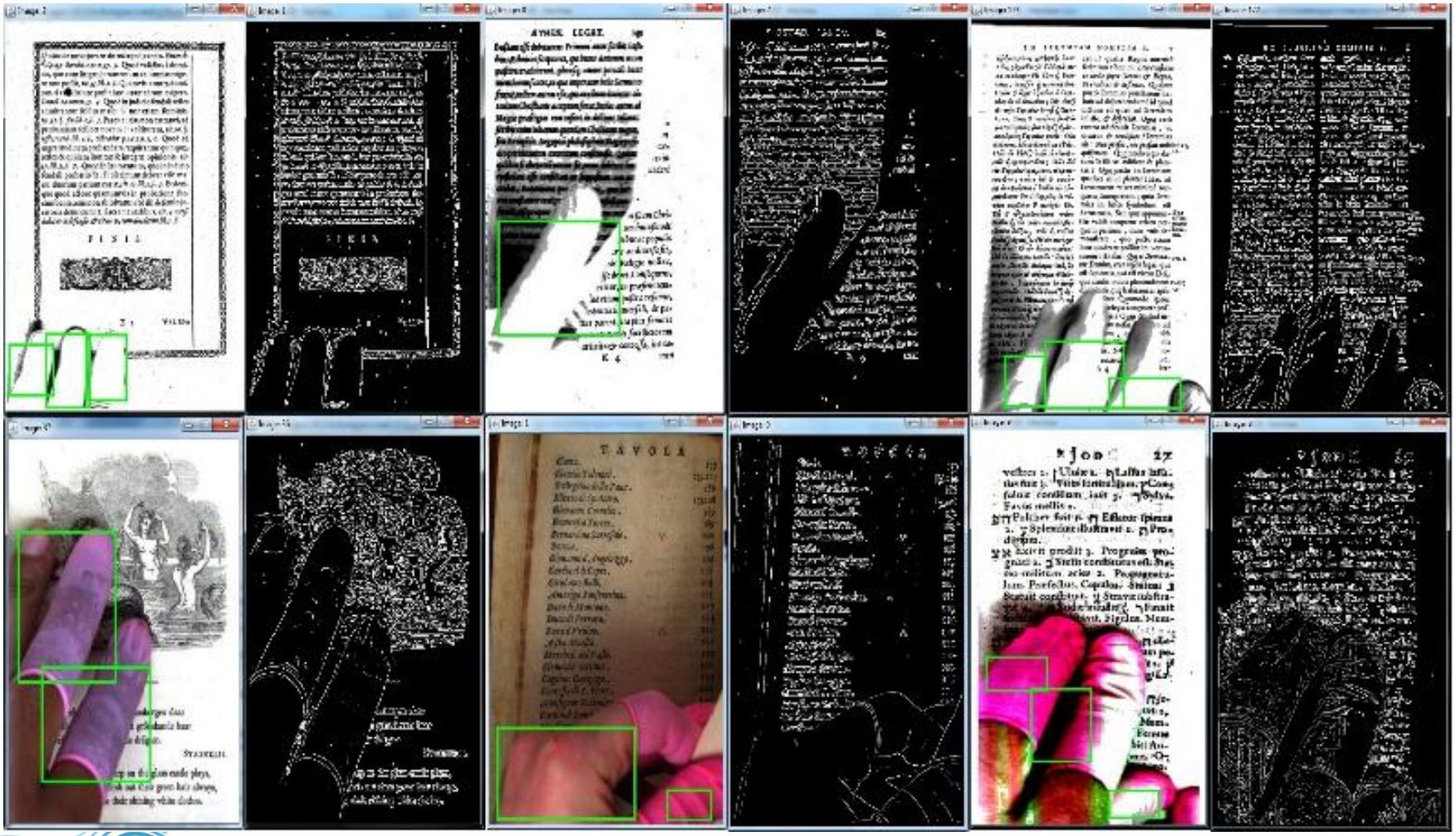
- Reduce costs
- Improves quality
- Saves time
- Automatically
- Increase efficiency of human work with particular focus
- Invariant to format, rotation, scale, translation, illumination, resolution, cropping, warping, distortions
- Application: assembling collections, missing files, duplicates, compare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

Quality assurance goals for finger detection on scans

- Automated preservation workflows are common in large digitization projects (e.g. museum collections, Google books,…).
- Automated quality assurance to ensure consistency of digital collection and to detect content modification in preservation process.
- Unintended placement of fingers over the document scan by workers performing the digitization process causes a significant quality reduction.
- The text obstructed by the fingers is lost and cannot be corrected.
- No automatic methods exist to detect fingers on scans - a human expert involvement is required.

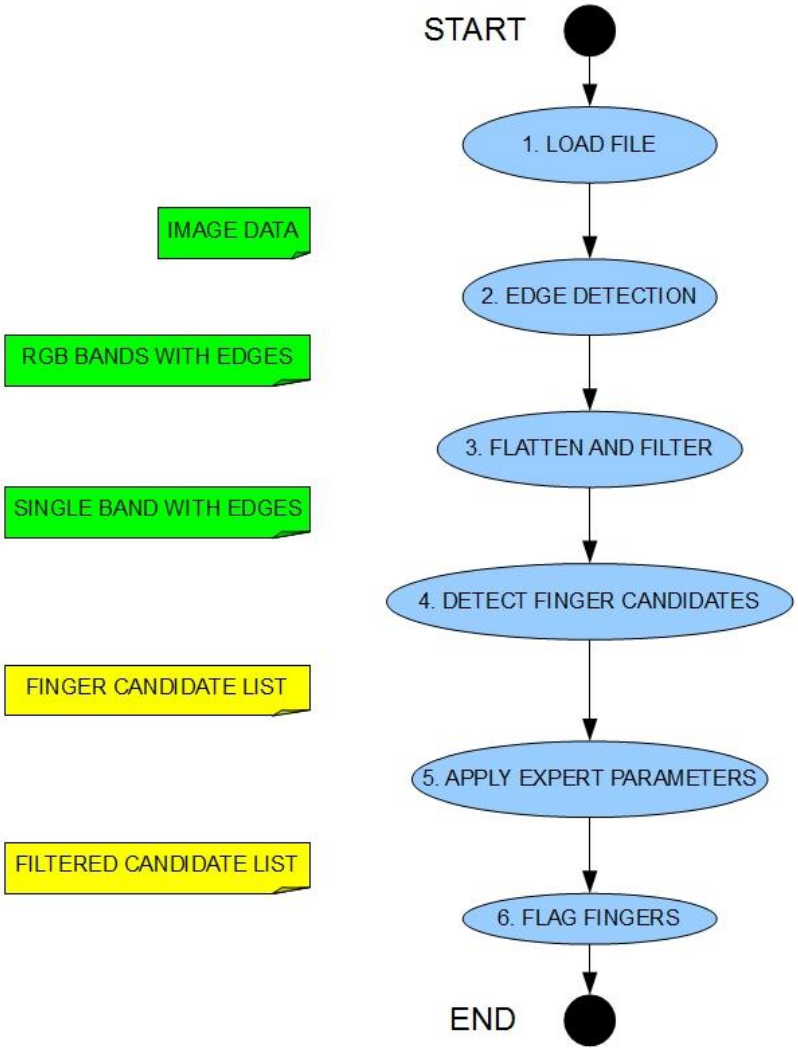
The sample positive detections



Finger detection process

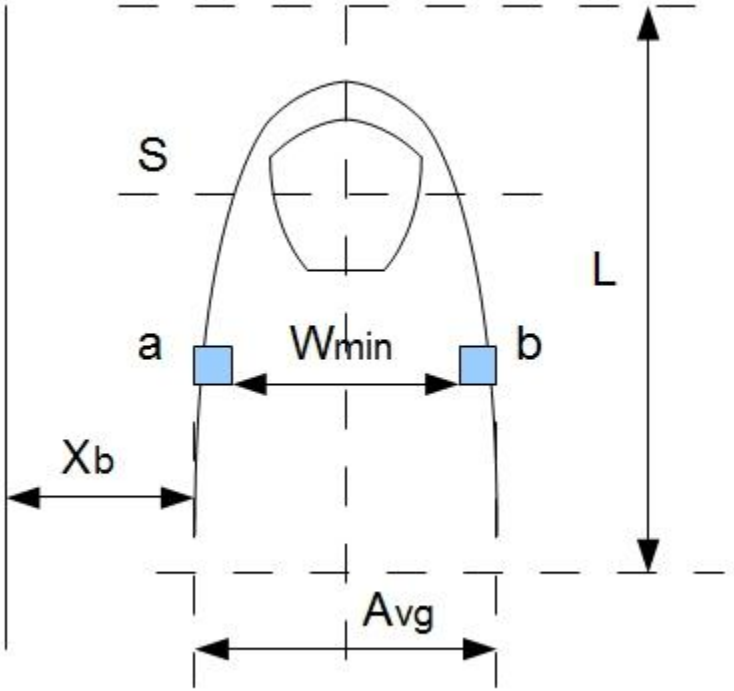
- Image processing methods
- OpenIMAJ library / OpenCV for edge detection
- QA workflow based on the expert knowledge
- Finger detection challenges
  - varying image quality,
  - different finger sizes,
  - direction,
  - shape,
  - color
  - light conditions

Finger detection workflow



Suggested method and parameter identification

- W_{min} is a minimal finger width for given Y coordinate
- X_b represents the gap between the finger edge and the page border
- ***Avg*** stands for the average distance between finger edges for all Y coordinates associated with this finger candidate
- Dashed line ***S*** represents a threshold where the **finger shape is cut** in order to facilitate calculation and to reduce the number of false positive results
- ***L*** parameter fixes the maximal accepted finger size.



The parameter used in finger detection algorithm

- Pixel value threshold ***P***. (range 0.0 to 1.0) in the flattened single band grayscale image. The pixel value threshold defines a **threshold** at which this pixel should be **taken into account** for further analysis.
- Minimal finger width $W_{n,m}$. This parameter describes the minimal pixel count that should contain a finger candidate on the X axis.
- Average width rate $A_{n,m}$. This parameter is used to set up an **acceptable threshold** for the **thinnest and the thickest distance** between finger boundaries in order to follow a relatively smooth finger shape without large volatile changes.
- Pixel variance $\mathbf{d}_{n,m}$. Due to image distortions often it is not possible to evaluate an accurate line. Therefore we have to set up **acceptable variance** for **neighboring pixels** that could build a line with a current pixel at analysis.
- Minimal finger points F_{min}. In this parameter we define the minimum number of pixel points regarded as detected pixels for a finger candidate.
- Maximal finger size $L_{n,m}$. This parameter defines the maximum size of finger candidate.
- Minimal border distances x_b, y_b. In order to involve page border factor in our calculations we define a minimal page border distances for both axes.

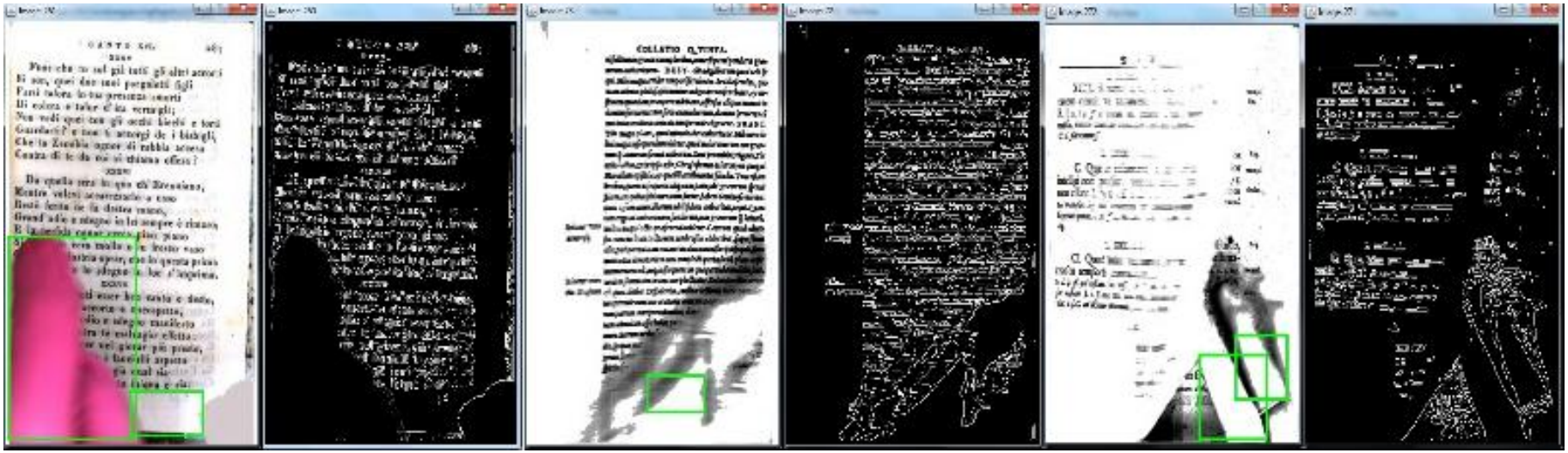
Document image processing and algorithmic details

- Canny edge detection algorithm - extract the significant shapes from the image
- Extract a gray scale image with additional filtering.
- Analyze and isolate finger candidates.
- Stroke width transform method: for each pixel we investigate pixels in the neighbouring area and follow lines according to given expert parameters.
- Finger object
- Matches all conditions defined for a finger
- Mark the evaluated region as a detected finger by green rectangle to facilitate further analysis for an expert

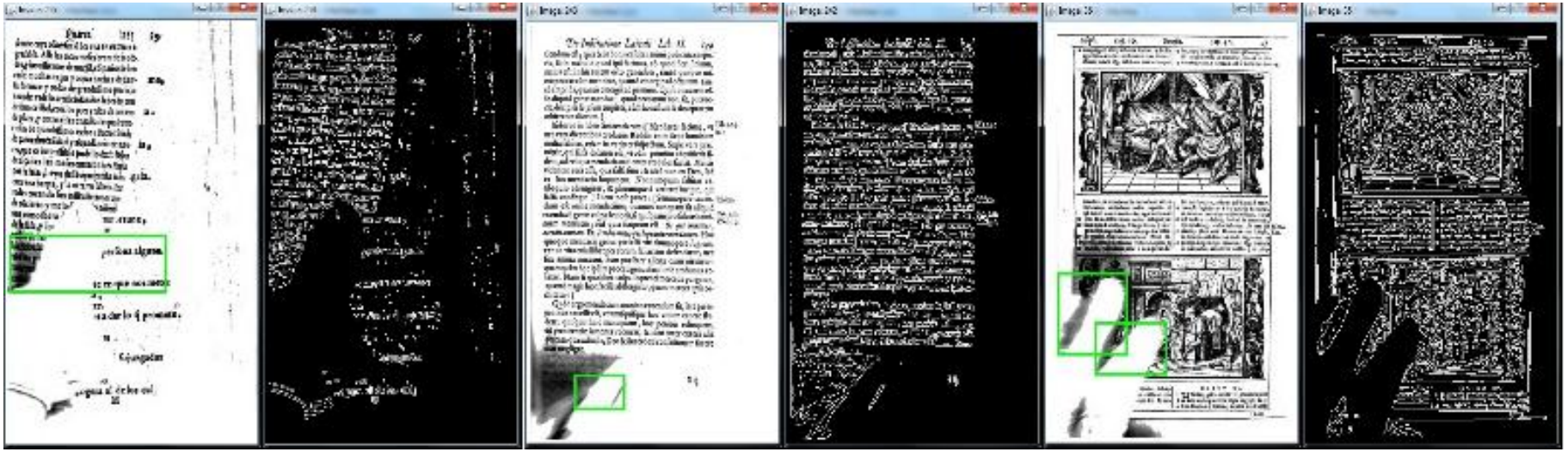
Evaluation

- Hypothesis: the reported approach should be able to detect corrupted documents with good reliability and to ignore unattended scans.
- **Scans** that are **flagged** by finger detection algorithm should be additionally analyzed by **human expert**.
- Manually created **ground truth** data for the assessment of the evaluation results.
- One collection contains 160 corrupted images that include fingers. The second collection comprises 26 images with fingers. The third collection contains 730 images and is a reference collection without corrupted documents.
- The **result** of the analysis is a **set of documents** with finger candidates marked with a green rectangle.
- Misclassifications coming along with correct results are few and happen with shapes similar to finger shape definition. Improvements in the algorithm and filtering of misclassifications are subjects of a future work.

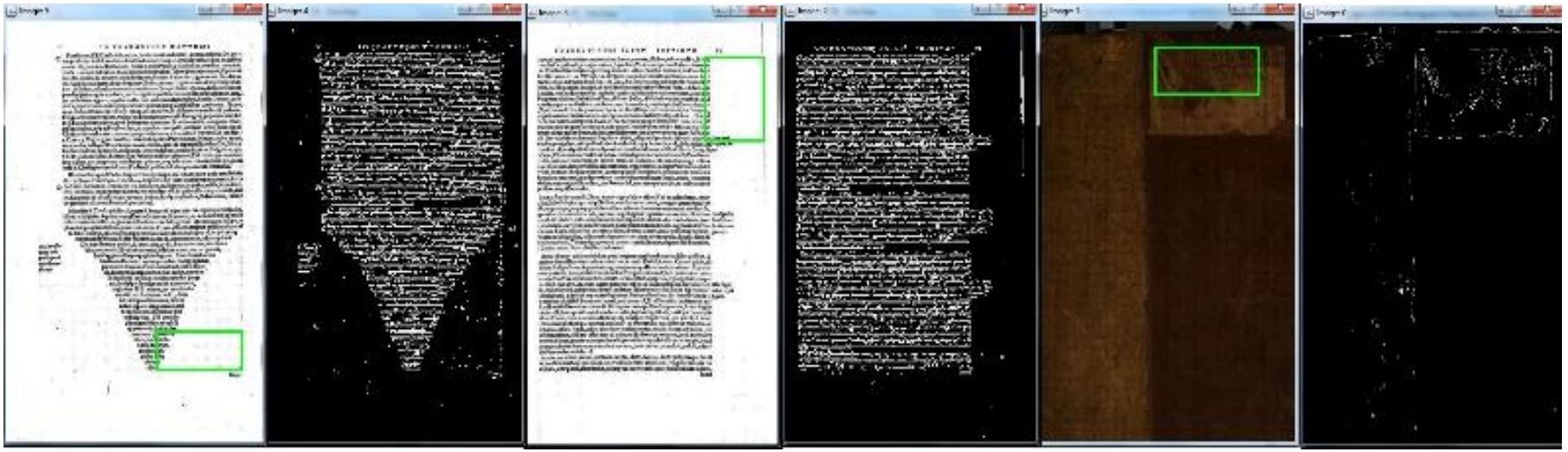
The sample for correct detection of blurred images



The sample for correct detection of fingers that are looking as a light spot on the scan surface



The sample false positive detections



The sample for scans with not detected fingers

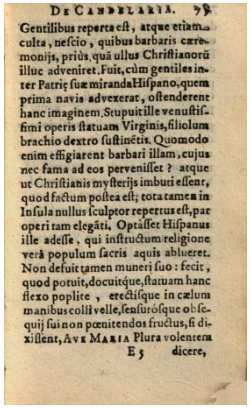
Effectiveness of the detection algorithm

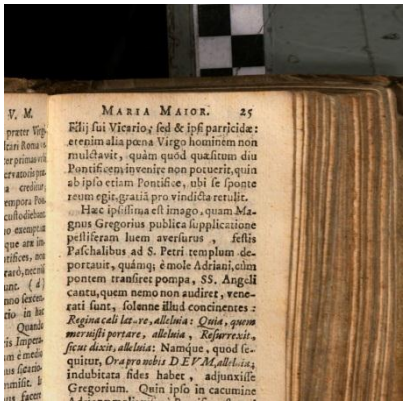
Collection name	Total scans	Correct detections	ROC point (best 0, 1)	Time (ms)
Scans with fingers (1)	160	139	0, 0.8633	185767
Scans with fingers (2)	26	20	0, 0.7692	8915
Scans without fingers	730	727	0, 0.9958	262363

- The effectiveness is measured by the distribution of collection points in Relative Operating Characteristic (ROC) space
- Results are very close to the best possible classification point (0,1) TPR/FPR
- Classification accuracy up to 86 percent
- Our approach is very promising for making the digitization process more reliable and for ensuring the quality of digital collections.

Cropping detection - a QA tool for document image collection handling

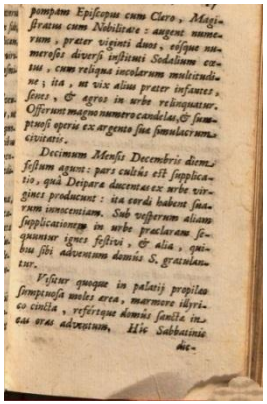
- **Border width** image width on X axis for border analysis
- **Border relation** between left and right border

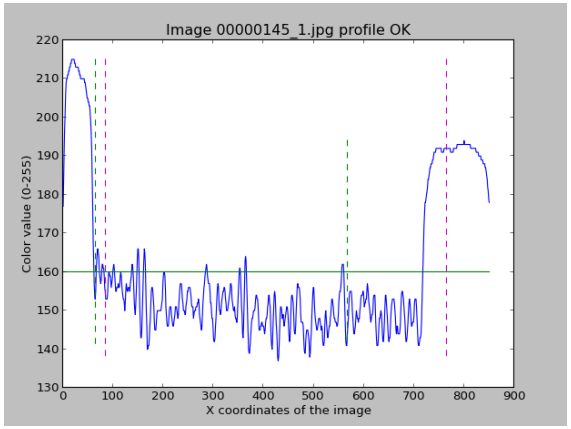


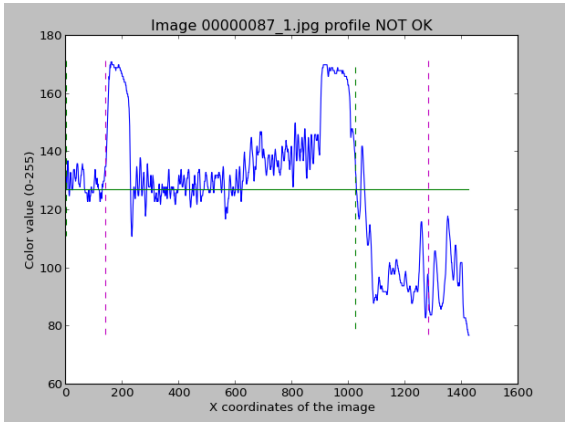


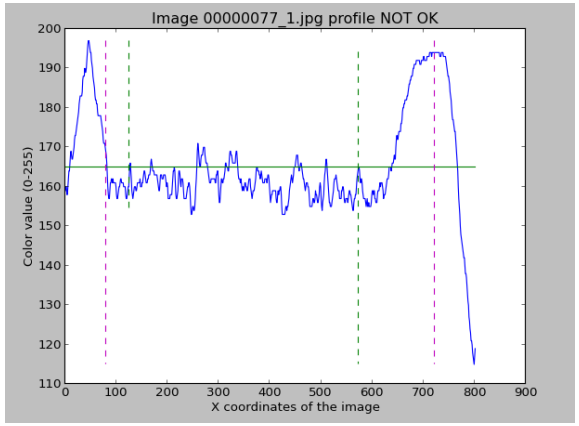
Use cases

1. text shifted to the edge of the image
2. unwanted page borders
3. unwanted text from previous page on the image









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!