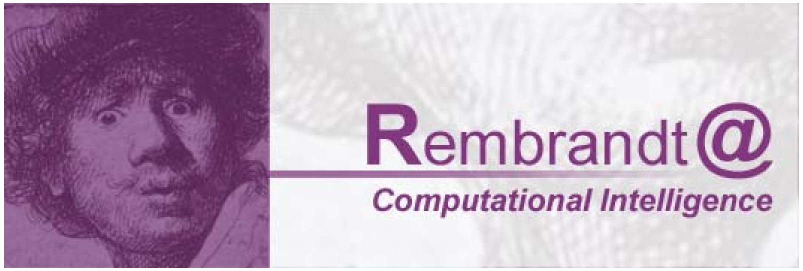
Paper Retrieval using Laid and Chain lines

Mark van Staalduinen



Fabriano, Italy
July 26, 2007

M.vanStaalduinen@tudelft.nl

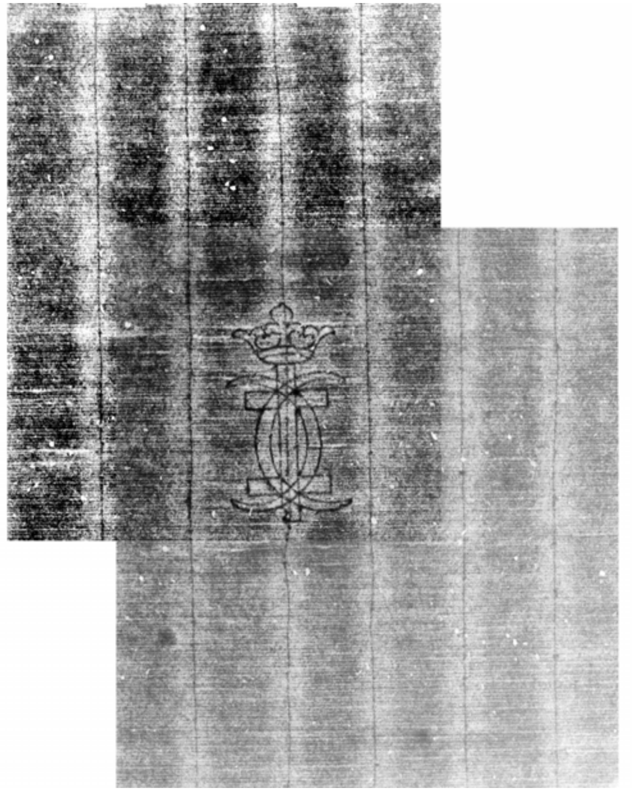
Information and Communication Theory

TUDelft
Delft University of Technology

Goal of paper retrieval - Dating

Dating is an important feature for art, documents, books, manuscripts, prints, drawings, maps etc.

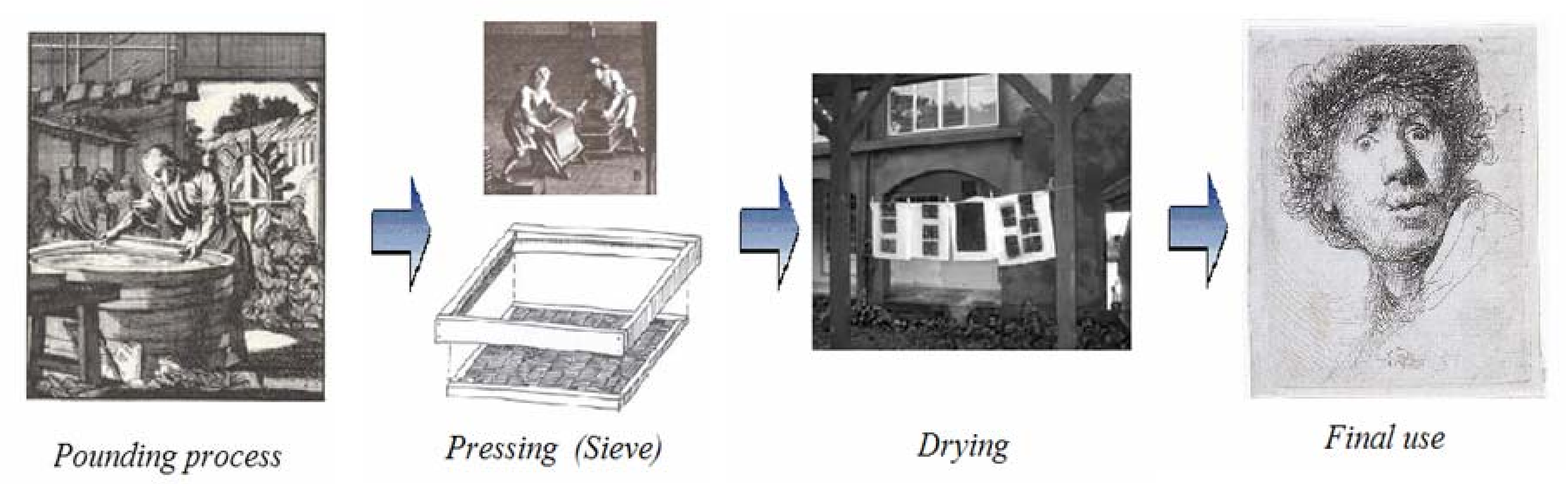
=> The answer is in the **paper**



Assumption

Paper with the **same features** is used in the **same period**

Paper features



- Pulp composition
- Water contents
- etc

- Watermarks
- Chain lines
- Laid lines
- etc

- Visual content
- Ink composition
- etc

Retrieval process

Goal: Discovery of identical pieces of paper

Mainly done manually (with automatic inspection) using features:

Visual content and Watermarks

These features are “most suitable” to manage manually

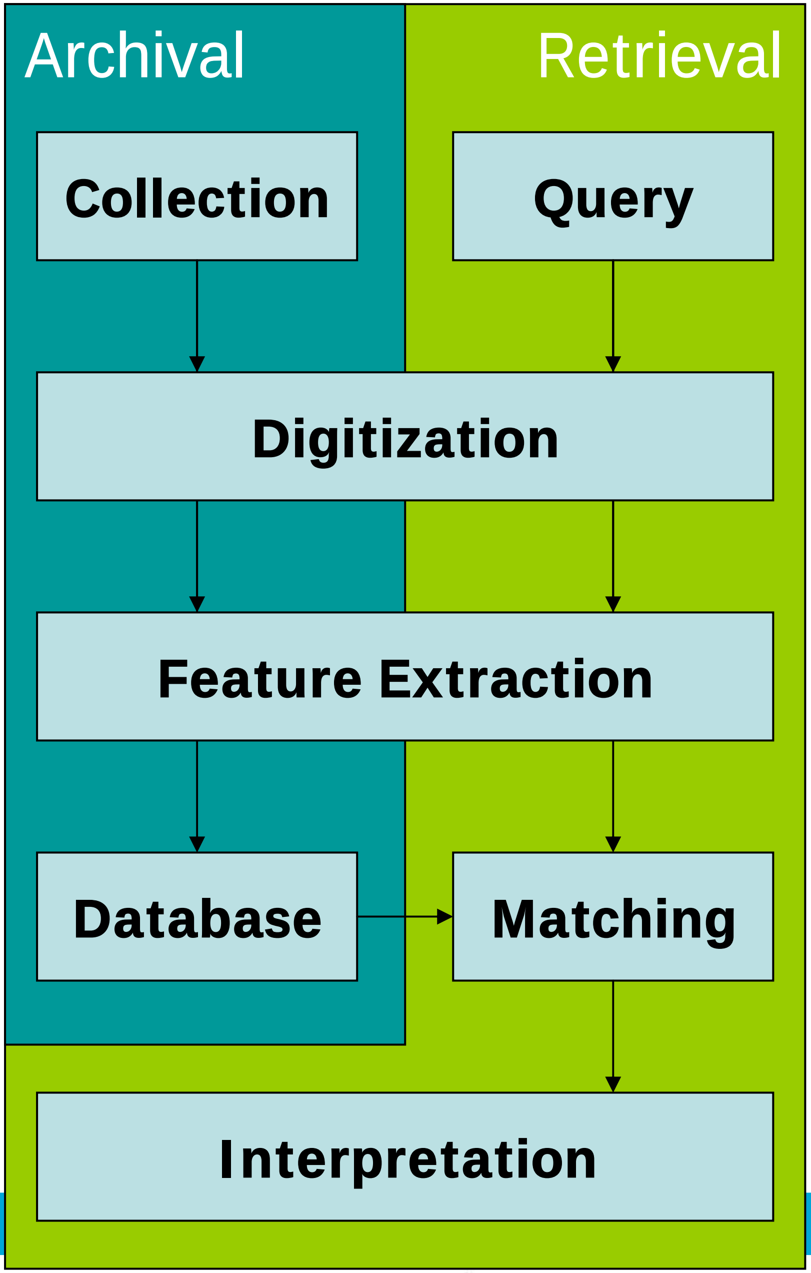
Is this also the case for computer techniques?

Our goal

Paper retrieval using Laid and Chain lines

Automatic Paper Retrieval System

- Start with query paper, which should be discovered in a collection to interpret
- Distinguish between archival and retrieval phase
- Automatic, so a matching procedure that compares database objects
- Paper features need to represented
- Digitization is needed



Digitization

Important properties

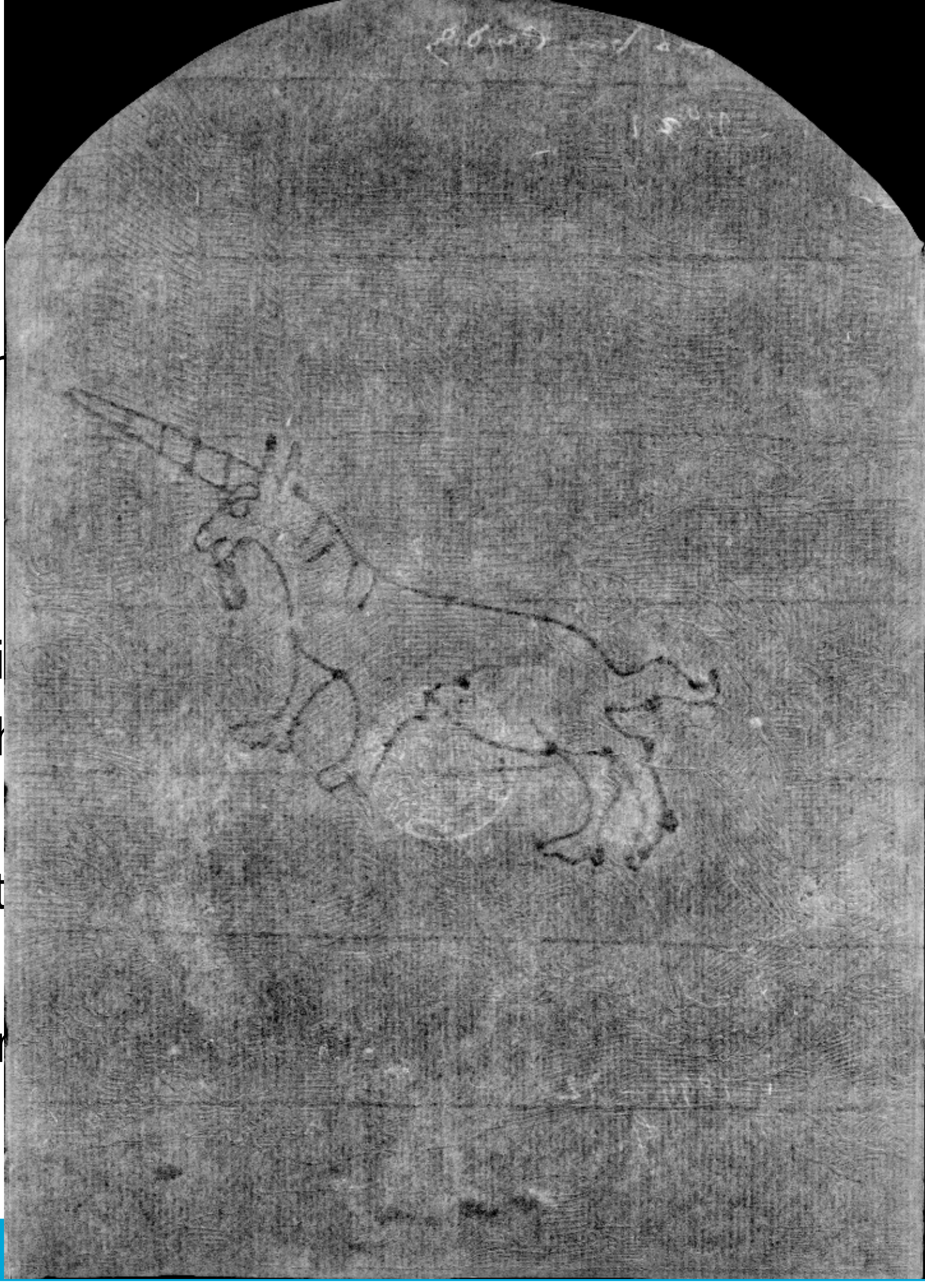
- Resolution, (minimal 75 dpi accordir
150 dpi)
- Selection of paper region

Soft x-ray imaging in cooperation wi

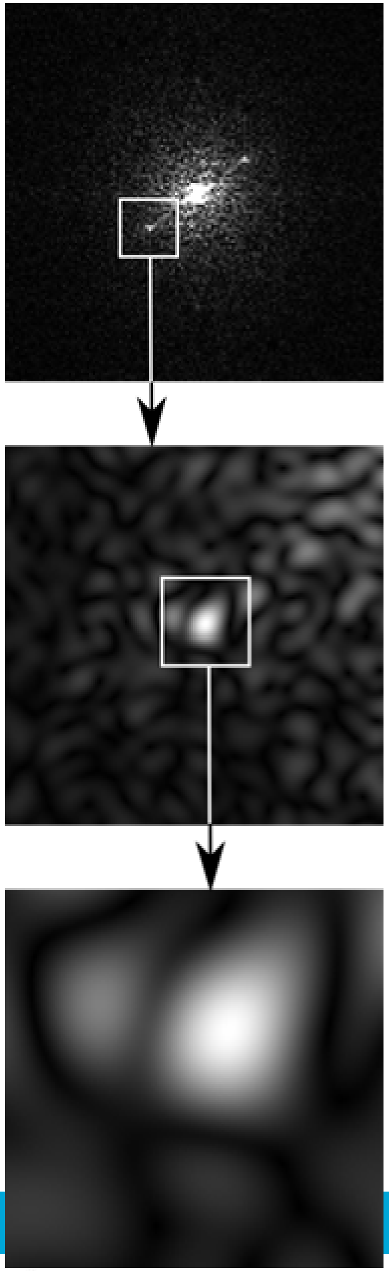
- Automatic method for selection of tl

Backlight imaging in cooperation wit

- Automatic method for subtraction
- Automatic method to estimate the ir
of the lineal



Feature Detection - Laid lines



Facts

- High-frequent regular straight line pattern
- Some variations in the frequency
- Laid line density between 5 till 15 laid lines per cm

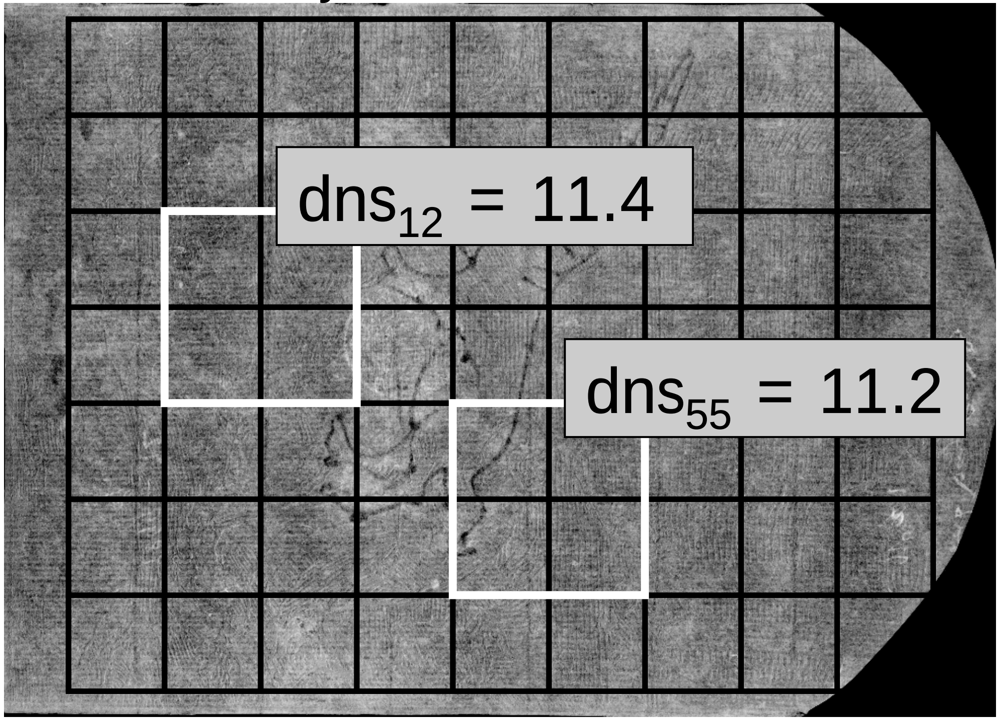
Detection of laid line density

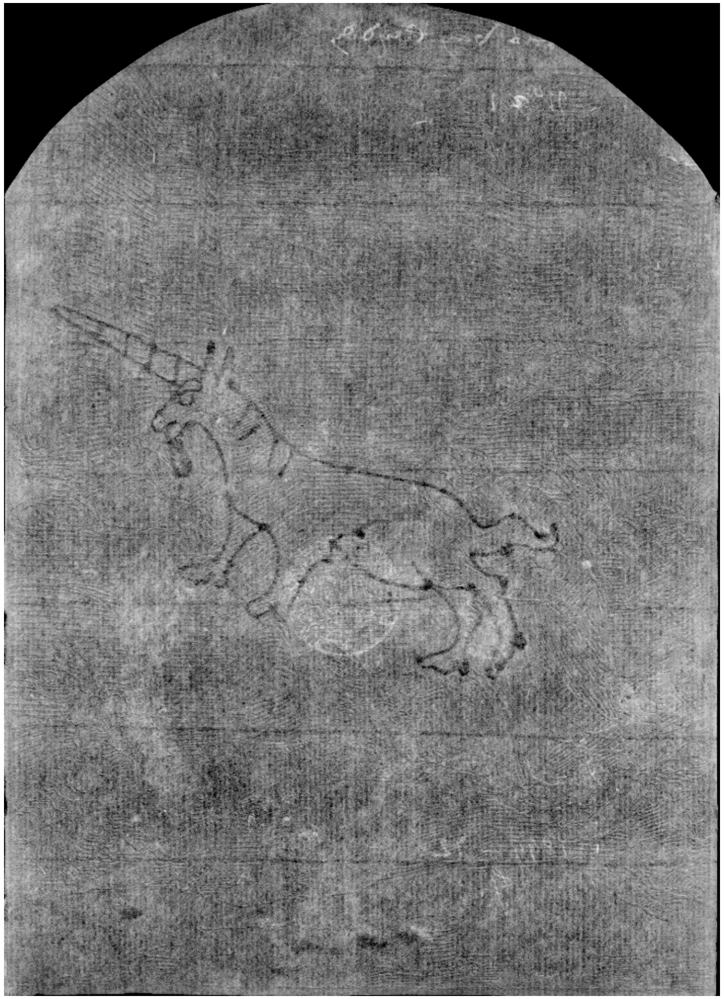
- Coarse-to-fine approach
- Coarse energy peak in bounded Fourier transform
- Due to variations, detect peak as a scattered blob
- Refine the local density estimation by the chirp Fourier transform

Feature Representation - laid lines

Concept

- Estimate the orientation and rotate the image, such that laid lines run horizontally





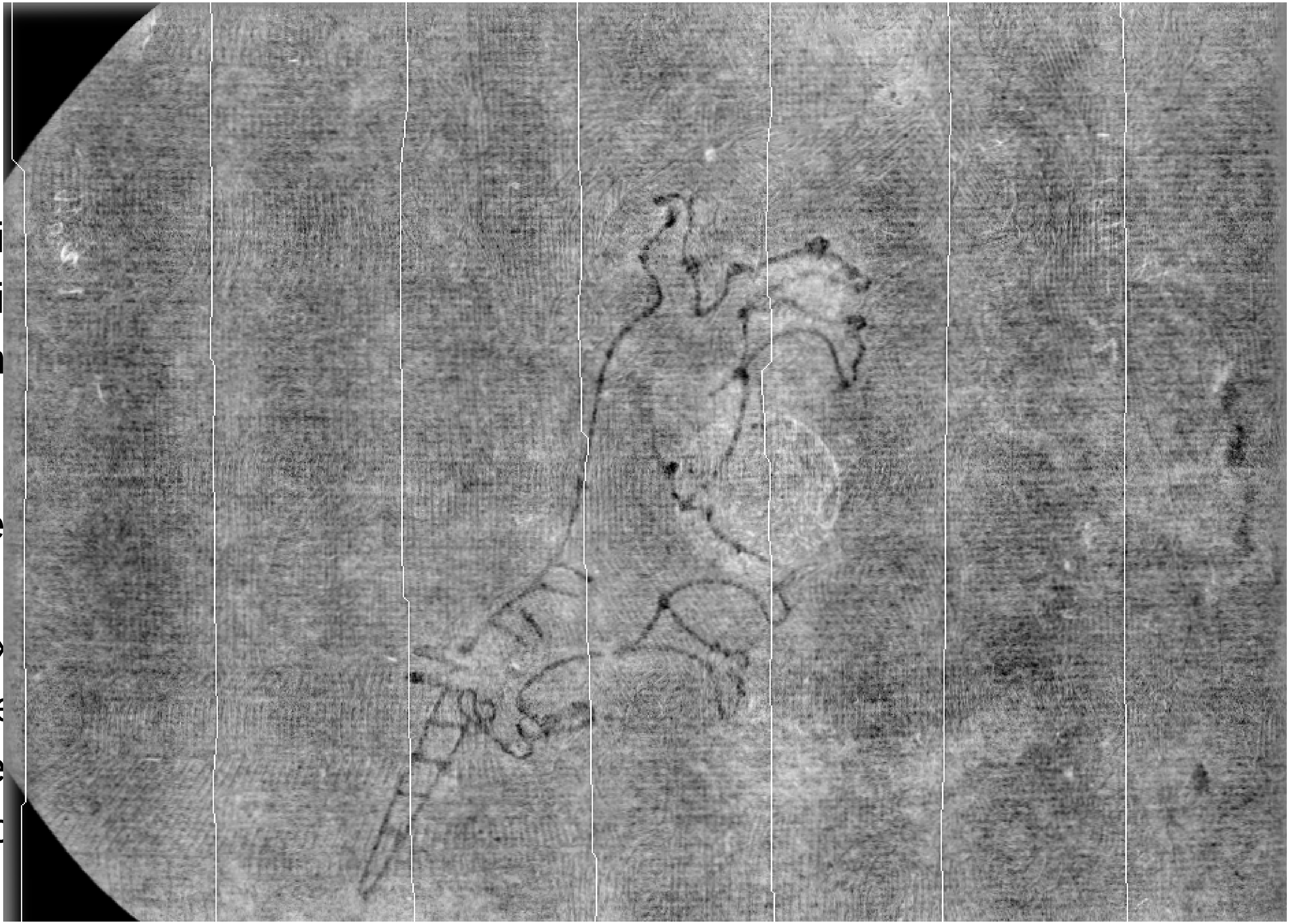
Feat

Facts

- Chai
- Chai
- Som

Chain

- Line
- Proj
  max
- Dete
- Sele
- Trac



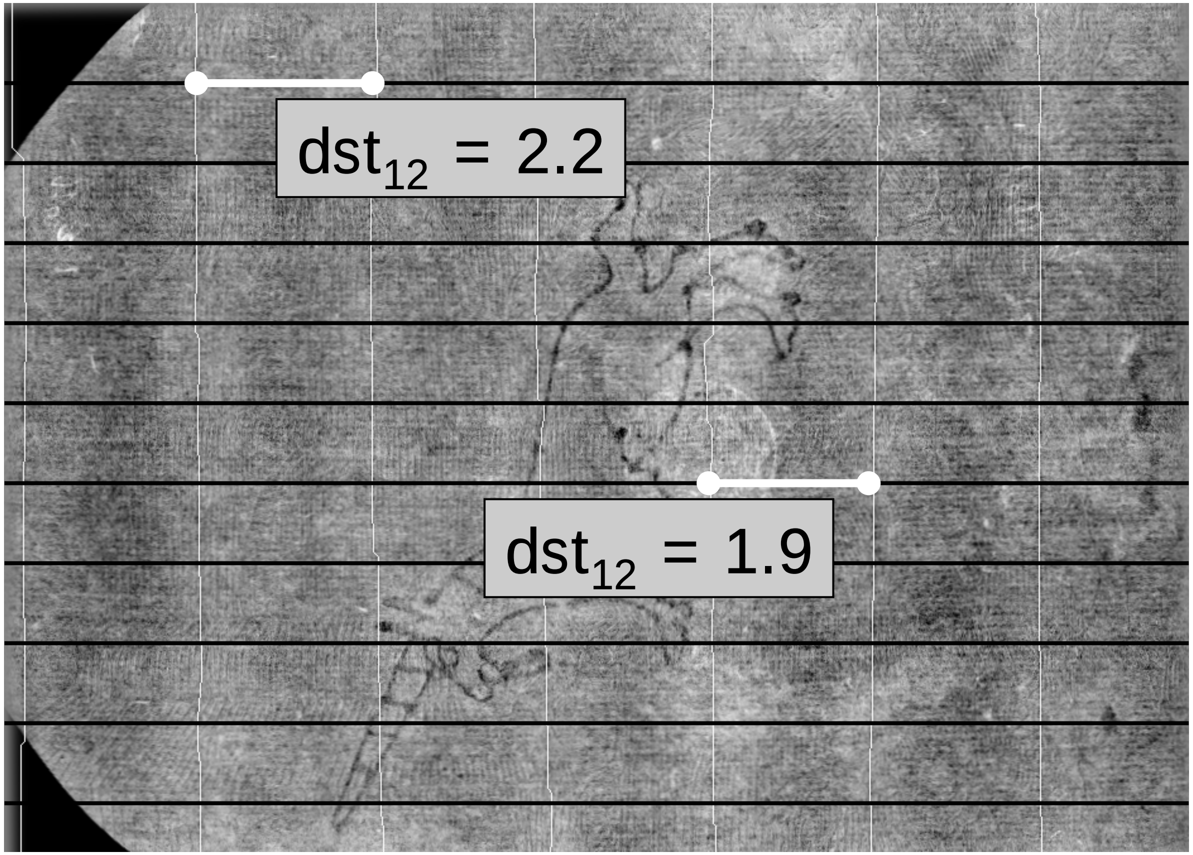
Feature Representation - Chain Lines

Concept

- Chain line distances are informative
- Chain line distances at a sampled rate

Representation

- Chain line distance matrix
- Average chain line distance
- Chain line selection mask



Matching with Laid and Chain lines

Facts

- Paper was cutted, so partial matching
- Needed balance between the amount of evidence and the error
- Four configurations: normal, flipped, rotated, flipped and rotated

Match certainty

- Independent features by decomposing density and distance matrices
- Estimation of pdfs by Gaussian distributions
- Log-likelihood ratio determines match certainty
- Best match is configuration with largest match certainty

$$\mathcal{M}(\,\delta_{ij}(t,\vec{n})\,) \;=\; \ln\left(\frac{\sigma_{\neg \mathrm{M}}}{\sigma_{\mathrm{M}}}\right)\underbrace{\boxed{|K(t,\vec{n})|}}_{\text{Amount of evidence}} - \frac{1}{2}\left(\frac{1}{\sigma^2_{\mathrm{M}}} - \frac{1}{\sigma^2_{\neg \mathrm{M}}}\right)\underbrace{\boxed{\sum_{\vec{k}\in K(t,\vec{n})} \delta_{ij}(t,\vec{n},\vec{k})^2.}}_{\text{Mean squared error}}$$

Retrieval Demo

Visual Inspection

Database objects are ranked on the basis of the match certainty of the laid and the chain lines

Interpretation Stage

Simple comparison of meta-data for identical pieces of paper

WWW

http://rembrandt.ewi.tudelft.nl

Discussion

- Automatic paper retrieval by means of laid and chain lines performs quite good. Laid and chain lines are easily represented for a computer, this is declared by the “watermark paradox”

- On the other hand, many watermark databases exist, therefore it is important to exploit the knowledge available in these databases

Watermark Retrieval

Semantics-based strategie

- Ordering by motifs (Piccard),
- Hierarchical / Rule-based ordered (Piccard Online, WZMA)
- Distance between horns (WZMA)

Context-based strategie

- WILC, laid lines (Atanasiu, van Thienen)
- Laid and Chain lines (Delft)

Feature-based strategie

- Landmarks (Ornato)
- Shape features (Pun, Eakins)

Watermark Detection

- Publications of Pun and Eakins showed rather good retrieval accuracy for tracings, the binary watermark representation

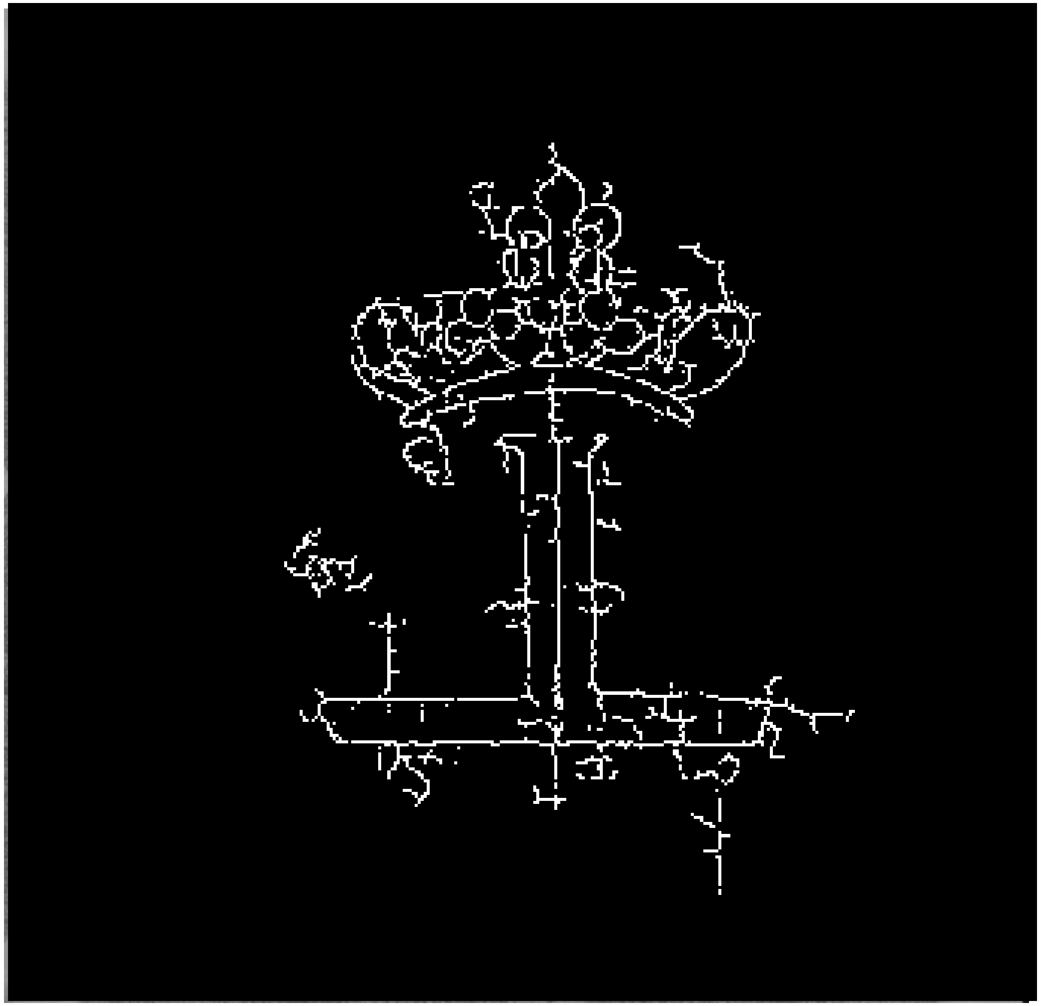
- Nowadays paper is mainly digitized as noisy images, no tracings are made anylonger

- Therefore, watermark detection is an important topic for the accessibility of the watermark databases

- Some results of cooperation with master student Hector Moreu

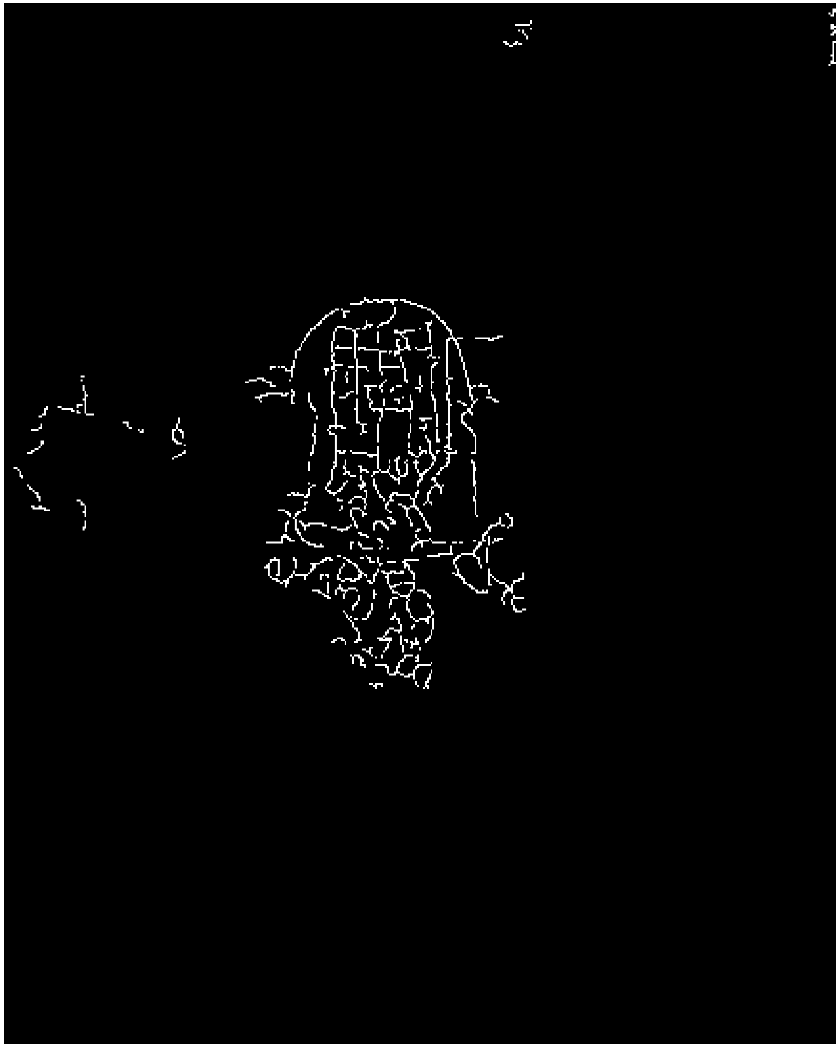
Watermark Detection

- Line profile
- Line contrast
- Line width
- Line connectivity
- Line length

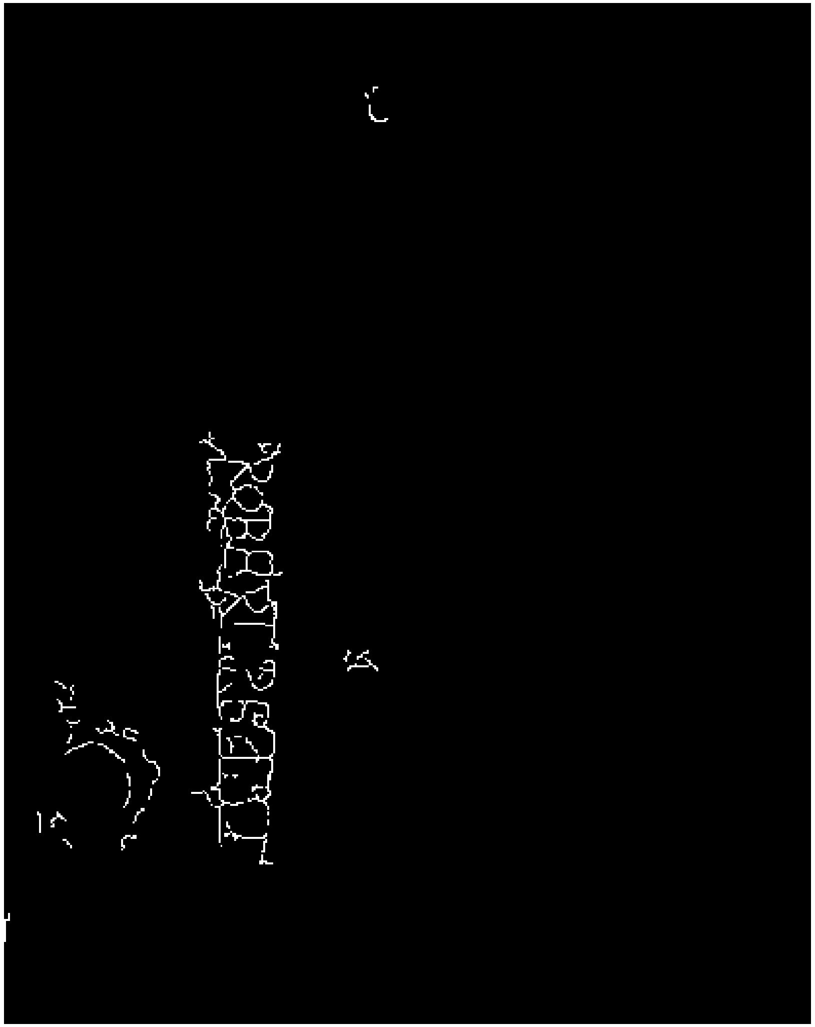


Detection is optimized by minimizing a trained error measure obtained by inspecting paper experts

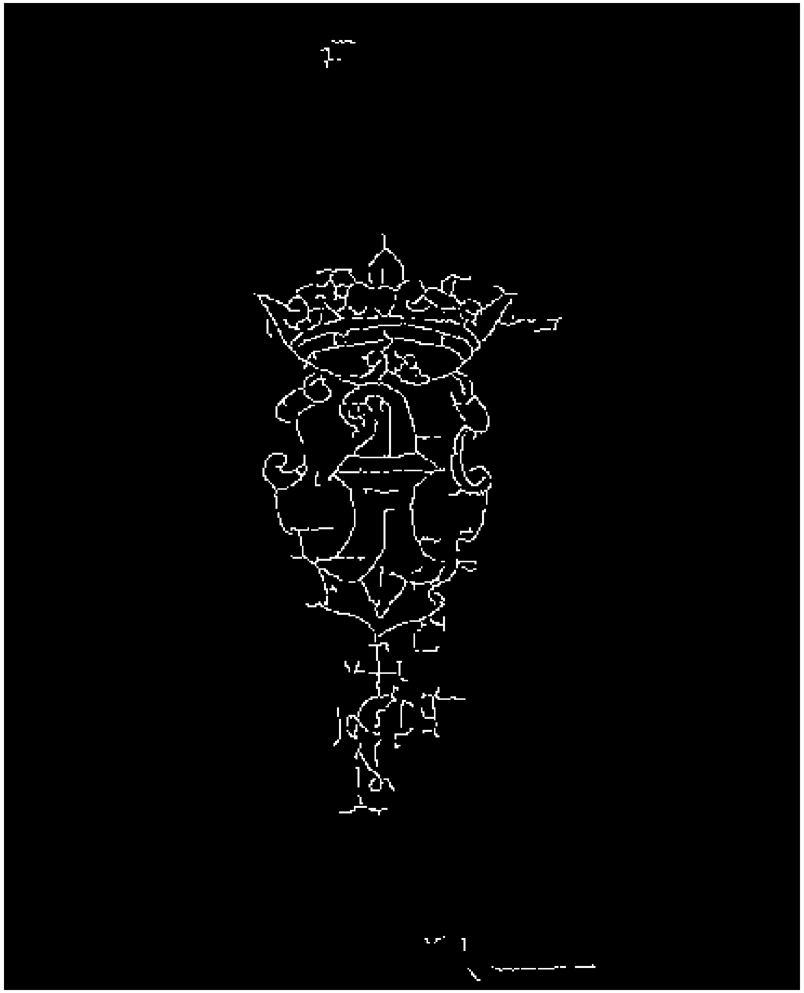
Watermark detection results



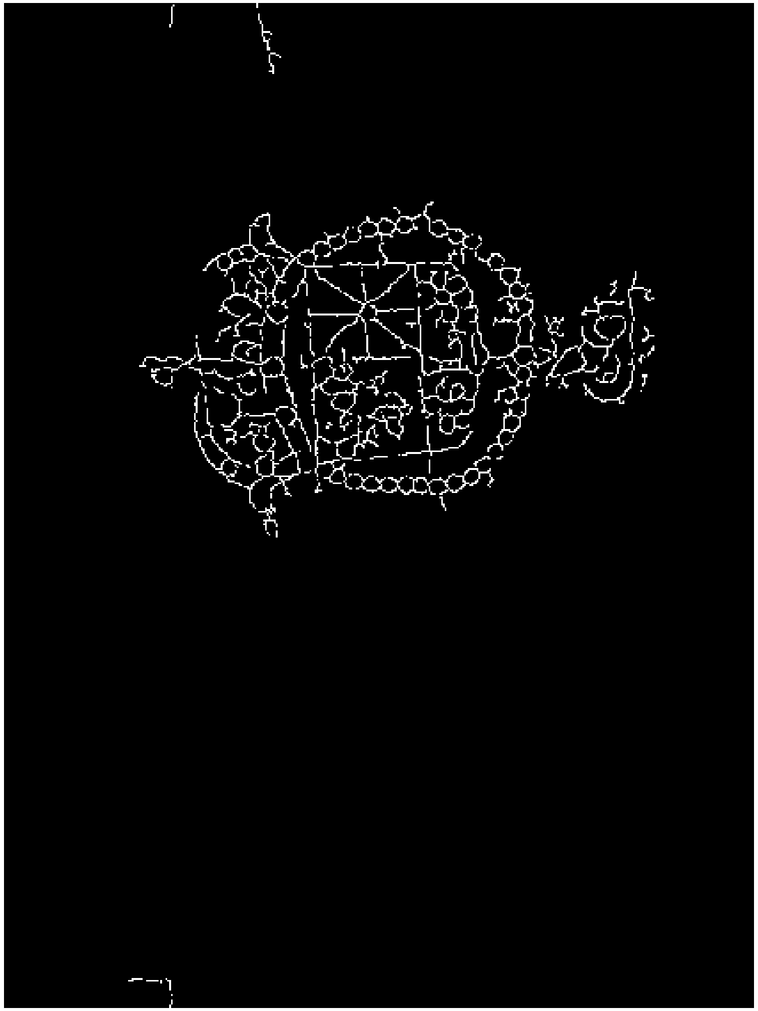
Watermark detection results



Watermark detection results



Watermark detection results



Conclusions

- Paper retrieval using laid and chain lines is a simple, but very effective approach

- Computers are better suited to represent and match with laid and chain lines, while human experts are better able to interpret watermarks

- Watermarks are very complicated shapes for which perfect detection will be difficult or even impossible, question is what is sufficient for retrieval

- TUDelft focus on content-based paper retrieval