The birth of a massive search engine for historical and multi-cultural handwritten collections

prof. dr. Lambert Schomaker
Continuous machine learning in Big Data

prof. dr. Lambert Schomaker
The Monk System

prof. dr. Lambert Schomaker
Monk  e-Science web service addressing these questions:

- **What?**  Word retrieval by 24/7 machine learning
  - Ambition: A European Google for handwriting
  - Server storage 1PB, filled 200 TB, with 37 collections
  - Qumran, Medieval, 17th century, 20th century, etc.
  - 400 M word images, 500 k labeled words, 30 k classes

- **When?**  Medieval manuscript dating
  - On a server (MPS), uploading of charters from 1300-1550

- **Who?**  Writer identification
  - On Monk server  over internet
  - Using GIWIS Windows tool
What?
**When?** Medieval Paleographical Scale (MPS) with prof. Jan Burgers (paleographer, Huygens Inst.)

Sub styles, geographical regions

Reference documents

Find year
Medieval Paleographic Scale (under dev.)
Writer identification (1:N) and verification (1:1)

1. Interactive, fully manual
2. Automatic, ROI based
3. OCR based

ROI based

Who?

Region Of Interest
Researchers

› Marius Bulacu, Axel Brink, Katrin Franke
› Ralph Niels, Louis Vuurpijl, Jean-Paul van Oosten, Sheng He, Jan Burgers, Petros Samara, Olarik Surinta
› “The Nijmegen Handwriting Group 1984-1993”
› Netherlands Forensic Institute: Ton Broeders, Wil Fagel, Elisa van den Heuvel
› Isabelle Guyon, Rejean Plamondon
› Users (humanities researchers):
  • Jinna Smit, Mark Aussems, Masahiro Niitsuma, Mladen Popovic, Daniel Stoekl, Jetze Touber, Grace Fong, Elaine Treharne, ...
Overview

› Digitisation of handwritten manuscripts
› Image processing challenges
› Text retrieval
  • Monk system: continuous learning
› Difficult material & the role of Knowledge
› Conclusions
Digitisation

› Int. Conf. on Document Analysis & Recognition 2009, Barcelona

  • Representative of British Library: we have computed that we would need three petabytes (three thousand disks of a terabyte, i.e., \(3 \text{ PB} = 3 \text{ million gigabytes}\))
  • ... therefore we are obliged to use ‘lossy’ compression techniques for storing images ...

› ??

› We are using \(10 \text{ PB}\) in Groningen just for some astronomers, manuscript archives and biomedical researchers!
Digitisation

- Current disk capacities allow for a full exploitation of all digital renderings of an object:
  - Gray scale
  - Color
  - Multi-spectral
  - 3D scanner (for Swedish hællristningar, Roman and Greek stone inscriptions, identification of the ‘cisel’)
- 1 PB for Monk (‘in the noise’, for astronomers)
Digitisation

› Digitisation prevents wear and tear of the books
› ... keep the books and documents in storage
› Quality considerations
  • Color mode: RGB/gray/black-and-white
  • Dots per inch? Bits per pixel?
  • Mechanical safety of scanning system
› Cost? (scanning, maintenance in digital life of object)
› Speed?
› Usage: How to make the digital access effective?
Case: Optical Character Recognition

› “The process of segmenting a text image into individual character images and classifying each as being a letter in an alphabet”

› Impressive results on contemporary printed text in machine fonts: with some linguistic postprocessing results are close to 100%

› Solved?
OCR?

› ‘OCR’ on historic documents does not work well
› On handwritten manuscripts it doesn’t work, at all!
› Problems:
  • image quality
  • unknown character shapes
  • unknown statistical language models

› However: pattern recognition and machine learning make enormous progress these days!
› Which methods? How to apply them?
Current technology: neat text/known language

› Why is ‘OCR’, i.e., letter by letter transcription on handwriting so difficult?
› Machine print: per character, per font, 8000 training examples are needed, typically
› E. Barney-Smith: 200k instances of printed c vs e
› Adress reading: reduced lexicon, zip codes etc., help
› In linguistic modeling: 20th century newspaper corpora do very little for 15th century acts
› Literary text, acts and charters each need their own knowledge models in order for OCR to work
Handwriting recognition: eat this!

- Many languages, scripts
- Over historical periods
- Contractions of letters
- ‘Suggested’ sloppy letter shapes
- Individual writer styles
- Image problems

⇒ Sliding window for character search usually problematic:

exit ‘OCR’
Monk - Design considerations

1. Don’t promise perfection
2. Don’t promise ‘transcription’
3. Don’t promise exhaustive coverage (as in databases)
4. Make use of human trainers, volunteers
   › Word spotting:
     • “a Google for handwritten documents”
Monk - Design considerations

› Word spotting:
  • “a Google for handwritten documents”
› The word is a reliable chunk of information with many shape features: redundancy

› Big Data: With sufficient data, there is always a reasonable response on a query
Monk’s world model:

- Institutes
  - Collections
    - Books (i.e., documents)
    - Pages
      - Paragraphs
    - Lines
      - Word zones and characters
    - Pixels
Example collection: Cabinet of the Queen (KdK)

› the Queen is the head of state
› the head of state is a member of the government and signs all laws and many government decisions
KdK Archival system

> all laws, decrees and correspondence are kept:
  • in a chronological order
  • and then in a numerical order
• KdK Archive
  with a total extent of
  (era 1798-1988):
  - 3,250 linear meter of shelves

• consisting of:
  - 28,000 boxes
  - average 1,000 pages per box
  • → 28,000,000 pages

• Of which Monk analyses the handwritten *Indices*, > 60k pages
Functions of Monk from user perspective

› Pages (read, annotate)
› Lines (read, annotate, search)
› Words (cut ROI, labeling, inspect hit lists)

<table>
<thead>
<tr>
<th><strong>Human</strong></th>
<th><strong>Machine</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cut out region of interest with the mouse</td>
<td>Segmentation of image into objects</td>
</tr>
<tr>
<td>Label a ROI with a text label: region</td>
<td>Classify image object</td>
</tr>
</tbody>
</table>

› General search functions, ‘google’ over several collections
KdK 1903
The Monk system / Lambert Schomaker
The Monk system / Lambert Schomaker

- RGB
- Gray
- Gray histogram equalized
- Bitonal B/W, dithered
- Gray Global contrast enhanced
- Gray Local contrast enhanced
Scanning and image preprocessing

- Quality of scanning companies varies considerably even if ‘strict’ standards (NL: Metamorphose) are applied
- OCR has very different demands from web access and human reading requirements
- ‘Simple operations’ such as ink/paper separation are not at all solved fundamentally in image-processing science due to a pervasive ‘chicken and egg’ problem:

  “In order to know the text, a system needs a good segmentation, but a perfect segmentation into lines and words is only possible knowing the text “
Coarse overview current contents Monk, 71 documents, 53000 page scans

- KdK Dutch administrative 1893-1906 20 books
- Dutch Admirality 1760-1823 5 books
- Printed, Elzevirium, 1616
- Qumran scrolls (2463 scans)
- Middelduits (example document)
- Accounts, 1425 Gelria
- Schepenbank Louvain, 1421-1559, 3 books
- Colonial diary (1932)
- Municipal year report 1855
- 20k illuminated initials
- Russian handwritten newspaper 1672
- Scholarly correspondence 1674-1682
- Chronicon Boemorum 1201
- Homiliarum Opatovicense 1150
- Resoluties Staten Generaal 1627
- Medieval charters 1300-1550 in 25yr periods
- Witch trial 1605

- ‘Beowulf’ and related, 4 books
- Ming Qing poetry, 7 books
- Arabic document, 291 scans
- Charlotte Perkins-Stetson diaries 1883- (682p.)
- Wittenbergsches (fraktur machine print)

**Cooperations with:**
- Huygens institute, Sorbonne,
- Harvard,
- Stanford,
- Czech National Library,
- Dutch National Library, McGill Univ.,
- Utrecht Univ.,
- Univ. Uppsala
- City archive Louvain,
Pattern recognition and machine learning

• Several years of experimentation
  • started in 2005
  • Monk was switched on, to autonomous mode, in 2009

• For developing and optimizing two functions:
  
  • Retrieval: return images for a given keyword
  
  • Recognition: return the most likely word given an image
Retrieval of *instances*,

Shape Feature B \[ \text{rank}_j (P(X_j|\text{Model}) \]

Recognition of *classes*,

Feature A \[ \text{argmax}_i (P(\text{Model}_i|X)) \]

Van Oosten & Schomaker (2013)
Separability & Prototypicality ...
J. Pattern Recognition
Boosting performance

› Old mine-shaft elevator principle: Fahrkunst
› After having trained with method A to its ‘max’
› An orthogonal method B can reach a higher performance
› Then method A again, etc.
› Until the real asymptote is reached
THE SOLUTION:

24/7 learning over internet

RuG HPC cluster

“Interactive Supercomputing”

Model refinement

Labeling by humans

Hit list computation

Continuous learning in Big Data
Monk architecture

Data types:
- Image files
- Text (labels, tags) files
- Key-Value index files
- Matrices (lists of labeled vectors)

Most of the files are Wordzone images

Use
Labeling
Browsing
Searching (Control)

Computing

Annotation, Labeling (.txt)

Queue:
Sor dex: “Amst er dam”,
Recog: “book-20”,
Index: “book-12”,
etc,

Word Models

Vectorial matching

Wordzones x100

Collections x20
Books x20
Pages x1000
Lines x40

Data types:
Image files
Text (labels, tags) files
Key-Value index files
Matrices (lists of labeled vectors)

Most of the files are Wordzone images
Lessons learned during Monk development

- A shape feature which is powerful for Retrieval may not be strong in Recognition!

- Feature B: hit list should provide nice, intuitive ranking in a satisfying ‘hit list’

- Feature A: target word class should survive competition with the other word classes (emerging needle from the heterogeneous hay stack)
Examples of feature patterns for two words

van der Zant, Schomaker, Haak (2008). IEEE PAMI

Amsterdam                             Groningen
Distance Measures for nearest-neighbour matching

- **Hamming:**
  \[ d = \sum_i |p_i - q_i| \]

- **Euclid**
  \[ d = \sqrt{\sum_i (p_i - q_i)^2} \]

- **Minkowski**
  \[ d = \left( \sum_i |p_i - q_i|^n \right)^{\frac{1}{n}} \]

- **Hausdorff**
  \[ d = \max_i \left( |p_i - q_i| \right) \]

- **Chi Squared \((\chi^2)\)**
  \[ d = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i} \]

- **Bhattacharyya**
  \[ d = \sqrt{1 - \sum_i p_i \cdot q_i} \]
The Monk system / Lambert Schomaker

monk: No dBASE db=* 

Monk - steekwoord zoeken

Tilburg

- KAK 1903 [Tilburg] 8 6 8 6 68 68 68 Leger Onderwijs - 1903/Novb/28 #30
- KAK 1893 [Tilburg] 8 6 6 4 5 1 8 4 [1893/]
- KAK 1893 [Tilburg] 8 6 6 4 5 1 8 4 [1893/]
- KAK 1893 [Tilburg] 8 6 6 4 5 1 8 4 [1893/]
- KAK 1893 [Tilburg] 8 6 6 4 5 1 8 4 [1893/]

Bedrijven - 1893/

Aantal [Tilburg] gevonden: 5

The Monk system / Lambert Schomaker
Tilburg:
Does it work?

› Do you need to label continuously?
  • Then transcription would have been better/easier
› How good is the pattern recognition?

› Do you need a supercomputer?
› How many people are working ‘back office’?
› Can I send my collection to Monk?
› What is the benefit, what is the output of the process?
#Labels, total as a function of ‘days on-line’, KdK 1903, 1893, 1897

![Graph showing the number of #Labels as a function of days on-line for KdK 1903, 1893, and 1897. The graph includes three lines, one for each year, with the number of #Labels increasing over time. The x-axis represents days, and the y-axis represents the number of #Labels.]
#Labels, total as a function of ‘days on-line’, KdK 1903,1893,1897

start $t=0$
synchronized
Monk

Jumps in the number of harvested word labels (y-axis) coincide with reaching a critical training dataset size (time axis).

Note the log() scale: these are ‘snowball effects’
Performance

- **Retrieval precision** increases monotonously with number of word examples

<table>
<thead>
<tr>
<th>#training examples</th>
<th>What to expect in Top 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Correct instance is good luck</td>
</tr>
<tr>
<td>5</td>
<td>A few correct hits</td>
</tr>
<tr>
<td>20</td>
<td>Many useful hits</td>
</tr>
</tbody>
</table>

- Lexicon of 11k words: > 81% **recognition rate**, KdK
- Lexwords with 50+ examples: > 95% correct
- Idem, Cliwoc captain’s logs: 93%
Good results: *bramzeijls* (top gallant’s) hit list after training on just 2 examples
ToDo / Baron / (no comment)

Items in hit list of Baron numbered: 0-2999. (now showing: 0-48)
<table>
<thead>
<tr>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
<th>Hoofdtekst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
<tr>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
<td>Exploitatie</td>
</tr>
</tbody>
</table>

The Monk system / Lambert Schomaker
At rank 343 in the list: pink newlings
Bulk labeling of all instances of [Octb] above.
Monk usage example: May 2013 Qumran scrolls:
Daniel Stoekl (Sorbonne), Mladen Popovics (Groningen)
Qumran scrolls: 2400 photographs

› Using Monk for character labeling
› With Daniel Stoekl and Mladen Popovics
› Using its 24/7 machine-learning cycle:
  • Label ➔ Train
     ➔ Label some More ➔ Train
     ➔ Easily label Many
› Thousands of characters ‘mined’ out of the Qumran collection of photographs in just two weeks, with very little effort in human labeling

Continuous learning in Big Data
<table>
<thead>
<tr>
<th>Recognizer hypotheses:</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alef</td>
<td>0.515</td>
</tr>
<tr>
<td>Gimel</td>
<td>0.555</td>
</tr>
<tr>
<td>Paleo_Tsadi</td>
<td>0.582</td>
</tr>
<tr>
<td>GreekAlpha</td>
<td>0.597</td>
</tr>
<tr>
<td>@PAPYRUS_EDGE</td>
<td>0.607</td>
</tr>
<tr>
<td>Yod</td>
<td>0.614</td>
</tr>
<tr>
<td>@Top_Of_Waw-triangular-closed</td>
<td>0.618</td>
</tr>
<tr>
<td>Paleo_Shin</td>
<td>0.657</td>
</tr>
<tr>
<td>@FRAGMENT_EDGE</td>
<td>0.662</td>
</tr>
<tr>
<td>Delta</td>
<td>0.675</td>
</tr>
</tbody>
</table>

Continuous learning in Big Data
Continuous learning in Big Data
Allograph harvest in Qumran collection
Allograph harvest 33k labels in NWO/MPS project with Huygens Institute

Items in hit list of @a_charter_1300 numbered: 0-3999. (now showing: 0-99)

... harvested during March 2014

The Monk System – Lambert Schomaker
Yes, also Chinese characters (handwritten & woodblock, Grace Fong & Harvard Yenchin collection)
Necessary metaphor for the Monk process

• Not: a stream-lined mechanical factory

• Yes: crystal-like growth, arborization,

• Randomly over a large surface

• Energy comes from:
  • human labels
  • machine re-compute actions
Time-lapse of Vitamin C Crystallization
by microbehuntercom
The Monk system / Lambert Schomaker
Monk collection on 02-04-14
52 ‘books’ and growing fast

On 28-10-14
72 ‘books’
Static index example: Chronicon Boemorum (1201 AD, Cosmas)
artificial intelligence
HWR group

Note: some @Monk-codes for word shapes have not been codified here, the user can click on them to do that
[bovenrand-boek]
[marge-boven]

Item *Renerus deveri* Lemmens de Stenhuffle filius Petri quondam Lemmens recognovit se debere Henrico et Johanni *deveri* Lemmens suis *fratibus* Willeimo de Ophem filius dominus Johanni quondam de Ophem *militer* et Henrico Lemmens filius *Reveri quandam* Lemmens *fratris predicti?* Petri centum libras grossorum Turonensium *veterum* bonas *legales* ad monicionem et *preterea famulum conductum* Albus Graven

junii xxvii

Item Johannes dictus Vos filius Nycolai dicti Vos de Steenhuffle recognovit se debere domino Johanni dicto Jacops de Steenhuffle *presbytero Renero* Henrico et Johanni dictis Lemmens de Steenhuffle *fratibus filis* Petri quondam Lemmens et Henrico de Horenbeke de Steenhuffle centum libras *grossorum* Turonensium *veterum* bonas et *legales* ad monicionem et *preterea famulum conductum* coram cisdem

Item Johannes Vos pistor in presentia promisit Quintinum de Valkenborch pistorem *ac* domum et curtem cum suis pertinentibus eiusdem Quintini site in Dorpstrata inter bona Willelmi Nagel et bona Johannis de Troembeke *exuantes* cum una parva curte in Bakeleystrata inter
Item Reinerus deveri Lemmens de van Stenhufel zoon van Peter voorheen Lemmens heeft erkend zich, dat hij, dat zij moeten Hendrik en (aan) Johannes deveri Lemmens zijn (bez.) fratibus Wim de van Ophem zoon heer van Johannes voorheen de van Ophem militer en Hendrik Lemmens zoon Reveri quandam Lemmens fratris predicti? van Peter

100 ponden grote van Tours veterum goede|goederen legales bij|tot|naar|op waarschuwing en preterea famulum conductum Albus Graven

juni 27

Item Johannes genoemd Vos zoon Nycolai genoemd Vos de van Steenhufel heeft erkend zich, dat hij, dat zij moeten aan de heer (aan) Johannes genoemd Jacops de van Steenhufel presbytero Renero Hendrik en Johannes genoemd(pl.) Lemmens de van Steenhufel fratibus filius van Peter voorheen Lemmens en Hendrik de van Horenbek de van Steenhufe 100 ponden grossorum van Tours oude goede|goederen en legales bij|tot|naar|op waarschuwing en preterea famulum conductum hoofdelijk aan dezelfden

Item Johannes Vos bakker in aanwezigheid heeft beloof Quintinus de van Valkenborch bakker ec huis en hof|tuin met zijn (bez.) aanhorigheden van dezelfde Quintinus gelegen in Dorpstraat tussen goed|goederen Willem Nagel en goed|goederen van Johannes de van Troembeke exeques met een klein|kleine hof|tuin in Bakeleynstraat tussen goed|goederen van dezelfde Willem Nagel en goed|goederen aan meester Hubertus voorheen de van aquis presbyteri inempnes penitus bewaren en relevare de van lilio modo siliginis mensuris Leuvense annue pensionis quam heer

Johannes Sampain presbyter quo bij|tot|naar|op leven suam heeft bij|tot|naar|op goed|goederen voornoemde en bij|tot|naar|op maiorem securitatem genoemd Quintinus de van premissa facundia prefatus Johannes Vos 3 gouden
Word-shape codification: attributes of a literal or @Monk-code

Monk - Word-shape codification () *

Latest=[..... hebben (MMonk.version);...] ....

Labels for words, characters, syllables or special symbols need to be described in detail. You entered this page because you have typed an hitherto unknown word label or because you wanted to modify the current definition.

User: Ischomaker

Unique Symbolic Monk Code (label): petri

Taboo this label?  
Take this ZonID as example?

ASCII characters, intended spelling: Petri
Characters, variant spelling: Petri
LaTeX encoding: Petri
Html diacritics code: Petri
Translation: van Peter
Visual type: Handwriting
Category: Proper (given) name (e.g., John)

Comment:

Guys, the second stroke of the e and the dot on the letter i may have been removed by the despeckler but the word is still quite unique so I don't expect many problems.

Web site (URL) with info

Test this URL

Rules:
1. Please do not try to be hypercorrect by entering diacritics in the main label: it is 7-bit USASCII only.
2. Ligatures and special pattern codes start with '@'
3. Do not enter ASCII special characters, write them out in @CODE notation:
   e.g., '@\ should be @SLASH
provisional transcriptions. The text shown comes either from human line transcription or (human) confirmed recognized words (see ...)

Cuper Braun 0014

6 [R] IOANNI BRAVNIO (continued) ut patet ex Liv. ...

... et ... 3 ...
tintinnabula illa summi sacerdotis ex ore fuisse, quae tū fuerunt ex auro

... B ... quibus ... u ...
vocant et quia oris in saevis Brachii, in quibus et nutrices @GR_chalko-
... quia ... qui ... a ...
@GR_drutai sacro sine dubio noie vocantur; plurimus usus, strepiturque
... g ... ipse_doces ... p ... fuisse ...
undem Deū colis et ipse doces p 96. Solenne Ethicus fuisse Bacchi
sacra celebrate, oneis crepidis et cymbalis ejusmodic calcimenta
habuisse Empedoclem, teste Tertulliano c.q. de pallio; i Andque probas
... @Epsilon ... s ...
auctoritate Plutarchi; apud quem tū l. 6 Sym. quē laudas, nihil tale
auctoritate Plutarchi, apud quem tū l. 6 Sym. que laudas, nihil tale
repperi. Sane tintinnabulorum in vestibus, ut apud Judaeos ...., usum
apud veteres fuisse nullibi reperio. Lego quidem apud M. Paulum
venetum 1.2.c.23 cursores Tartarū Regium habere, cinetoria
... nihil ... ad ...
sonoris tintinnabulis plena gestantes, sed hoc nihil ad rem
... n ... a ...
unde o abs re est, quod putem comparatione H m
... @o_- ... a ... quod ... ut ... u ... m ...
strepituum, quos .... tintinnabulis, Bacchantes oris cymbalis edunt
... quo ... s ... n ...
... b ...
instituere, nec illus dixisse Sacerdotē Bacchi et tintinnabula habuisse,
illa autem @GR_ho deiknumenos en tois enantiois tou meteooron thursos
... a ... bene ... n ...
Conclusions

› Great progress is made in handwritten text retrieval in recent years
› The concept of iterative refinement by human labeling & computing really works
  ➔ interactive machine learning over Internet
› Image preprocesesion and layout analysis are important, still crude and not fundamentally solved
› Image-quality standards (NL norm: *Metamorfoze*) do not guarantee high-performance handwritten text recognition
› Approach works for several historical periods (1200, 1400, 1600, 1800): but each needed some minor tweaking of the image preprocessing
› Think big!
How to improve on human and machine transcriptions?

The legal transactions in the acts of the Leuven Alderman’s rolls, 1421: Towards a formal semantic model of acts (Marianne Ritsema van Eck, august 2011). Presented at Digital Humanities 2012

- experts fight over the transcription
- students need to know more than paleography
- the machine needs a model for top-down reasoning
Semantic modeling and layout of acts: example: ‘IOU’

Fig. 4.9 (MS page V64.1 / scan 128) The relevant regions of interest, that correspond to the structural elements the act is composed of, are indicated on the MS page with bright colours.
Spatial modeling: Layout vs Semantics in the acts of the Schepenbank

The Monk system / Lambert Schomaker
Digging into Data / Global currents -- US/Canada/NL project with Mohamed Cheriet of Ecole de technologie supérieure (Montréal) 
Andrew Piper (McGill University) (principal investigator) 
Elaine Treharne (Stanford University)

The Monk system / Lambert Schomaker
Digging into Data / Global currents -- US/Canada/NL project with Mohamed Cheriet of Ecole de technologie supérieure (Montréal) Andrew Piper (McGill University) (principal investigator) Elaine Treharne (Stanford University)
Conclusions

› Great progress is made in handwritten text retrieval in recent years
› The concept of iterative refinement by human labeling & computing really works
  ➔ interactive machine learning over Internet
› Image preprocessing and layout analysis are important, still crude and not fundamentally solved
› Image-quality standards (NL norm: *Metamorfoze*) do not guarantee high-performance handwritten text recognition
› Approach works for four historical periods (1200, 1400, 1600, 1800): but each needed some minor tweaking of the image preprocessing
› Think big!