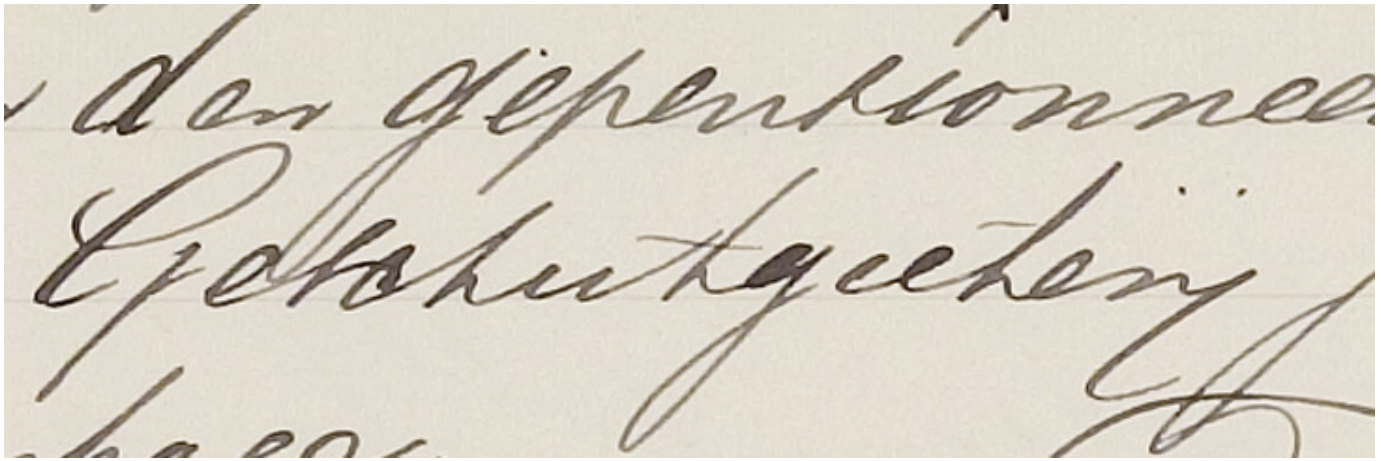


ICFHR 2010

Introductory words – Lambert Schomaker

International Workshop Conference on frontiers in Handwriting Recognition

- Handwriting recognition is such a difficult problem that:
  - We need to try out all newest methods asap;
    - And invent our own new algorithms, some of which had a solid impact on pattern recognition, machine learning and computational linguistics – *at large*

A heroic history formed at the frontiers

- Selected **feats** from ICFHR (1)
  - SVMs from the AT&T group, Boser & Guyon with their seminal paper on margin maximization
    which was the direct result of the frustrations about the overly variable results on neural-network (MLP) training in on-line **character recognition**
  - Convolutional MLPs (LeCun) as a 2D generalization from TDNNs
    IWFHR-1, CENPARMI Montreal,
    were based on **character recognition**

A heroic history formed at the frontiers

- Selected feats from ICFHR (2)
  - Raw image skeletonization is too noisy, look further than your nose and use algebra to prevent strange forkings! (Nishida, Suzuki & Mori, in Bonas, 1990)
  - MLPs and on-line character recognition, freezing the weights to the hidden layer after preliminary training, then allowing the list of output nodes to grow as new allographs come in for training (Guyon, in Bonas, 1990)

A heroic history formed at the frontiers

- Selected feats from ICFHR (3)
  - US-post funding & adress reading saga at CEDAR, end 80-ies, begin 90-ies in Buffalo (Srihari, Govindaraju)
  - Behavior Knowledge Space: Bayesian classifier combination, *avant la mode* (Huang & Suen, in Buffalo 3rd IWFHR, 1992)

A heroic history formed at the frontiers

- Selected feats from ICFHR, middle 90-ies (4)
  - HMM revolution in on-line HWR: Manke, Schenkel, Dolfing (in Colchester), Artieres
  - HMM revolution in off-line postal-address reading: Gilloux (F), AEG|Daimler|Siemens (D)

The data … the benchmarks

- (M)NIST
- Unipen
- IAM
- IrOnOff
- …

Is HWR solved in 2010?

- ICDAR 1997, Ulm (D)
  - machine-print OCR is solved!

- ICDAR 2009, Barcelona (E)
  - HWR is the buzzword

- Solved? Not at all!
  - Why so little HWR on iPad? Gestures? yes free-style cursive? Not really
  - What happened to Tablet PC?
  - How to deal with historical manuscripts?
  - etc.

Handwritten archives, a challenge …

- Example: KdK (Cabinet of the Queen) 60 shelf meters

- fan out: **one running meter** of handwritten indexes provides access to about:

- **50 running meters** of chronologic arranged Royal decrees, laws and cabinet's letters, mostly handwritten

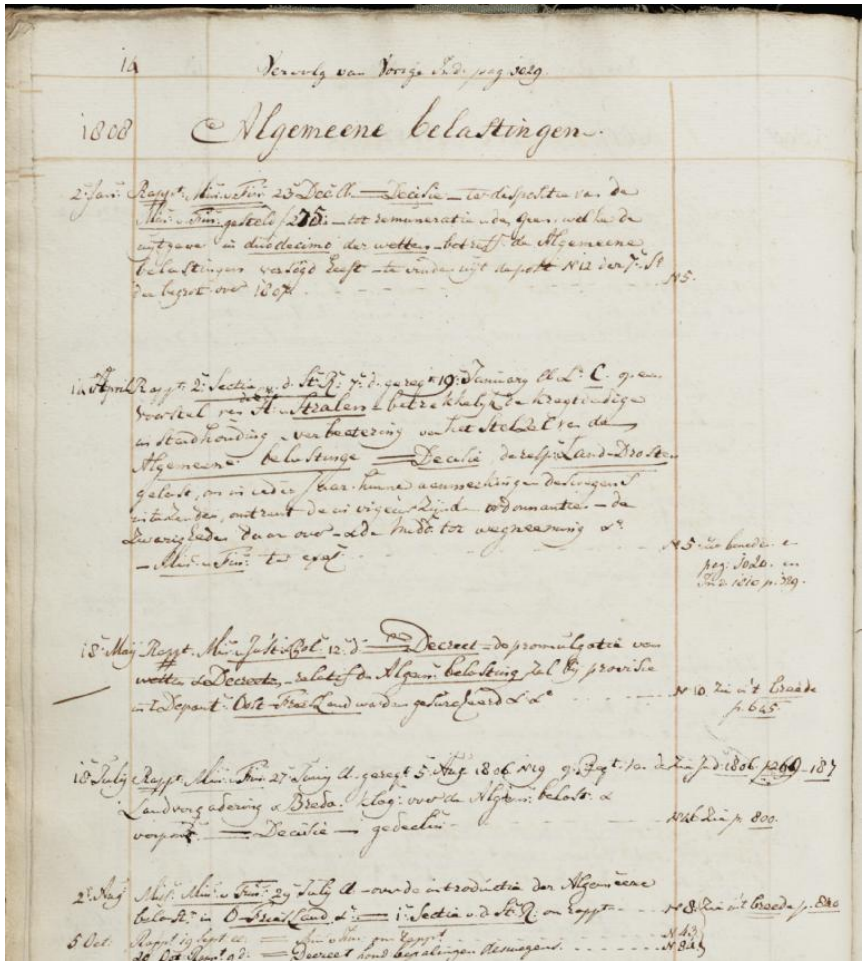
… of formidable magnitude …

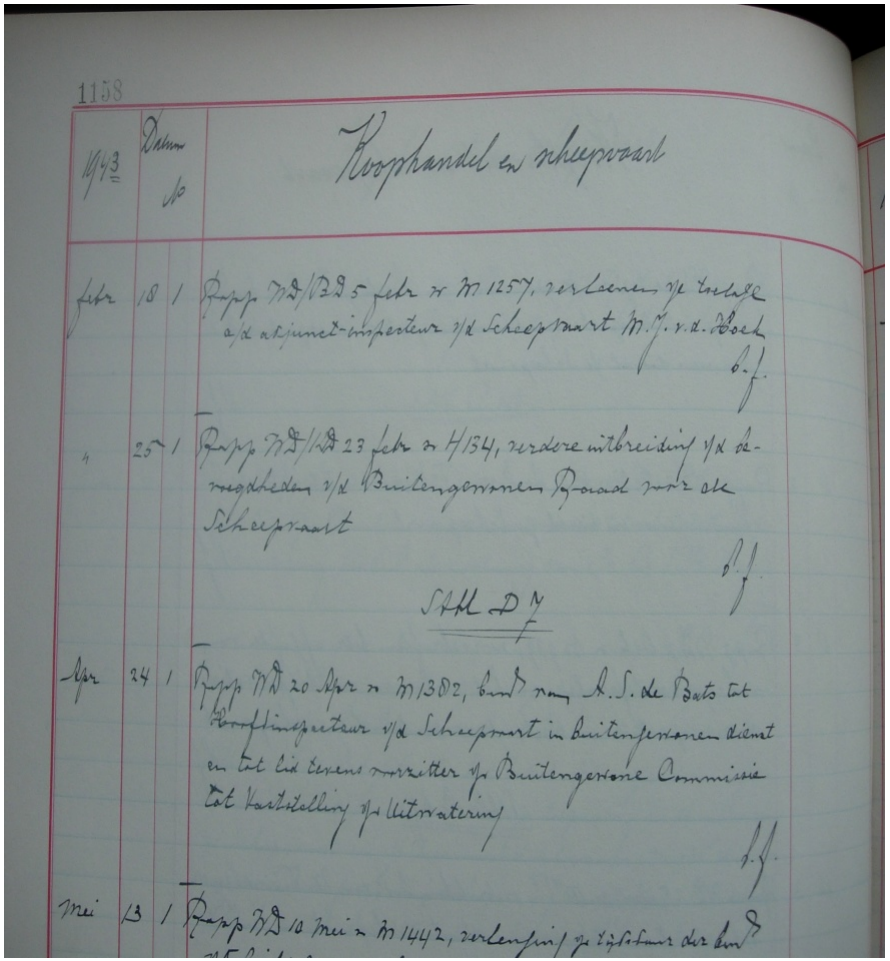
- with a total extent of (era 1798-1988):
  - 3,250 linear meter of shelves

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

    - □ 28,000,000 pages

… and complexity





From paper to silicon

- IBM Blue Gene (“Stella”)
- 14k processors
- > 28 Tflop/s
- > 6TB memory
- 150 kW

Scale up!

- Example: Monk system, Target project in Groningen

- Dutch archive Cabinet of the Queen, captain’s logs, and mediaeval manuscripts
- +60k page scans of handwriting
- disk test bed: now1.5 PB towards 10 PB
- Modern file systems (gpfs)
- Live 24/7 machine learning

The pitfall

- One algorithmic idea
- One data set
- One PhD student
- Three to four years of tinkering
- Resulting in ‘95% recognition’
- ‘our local hero has solved HWR’
- The industry yawns

How to stay away from the pitfall?

- k-fold evaluation on a closed data set is not enough: open systems need to be tested to avoid bias & overfit
- Larger, time-variant data sets are needed!
- Data diversity is cool, not scary
  ‘*an overly clean data set is nothing more than a fata morgana*’
- Code projects like Ocropus, more cooperation

Challenges galore: ICFHR is thriving!

- Scientific and engineering problems remain as tantalizing as ever:
  - character classification
  - word recognition
  - text retrieval
  - writer identification
  - layout analysis
  - image processing

ICFHR 2010 will show:

- … Script types you never knew they existed !
- … ML tricks you never thought of before !
- … Image processing algorithms that are unseen !
- … Applications presented here for the first time !

- Let’s go identify the heros of today!