



# ICFHR 2010

Introductory words – Lambert Schomaker

# International Workshop Conference on frontiers in Handwriting Recognition

2

- 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

3

- 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

4

- 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

5

- 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

6

- 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

7

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

# Is HWR solved in 2010?

8

- 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

the Queen's Cabinet

# ... and complexity

1a  
1808  
Keruny van Koninkrijk der Nederlanden  
Algemeene belastingen.

2 Jan. Koninkrijk der Nederlanden 23 Dec. 1807. Decree van de koninklijke commissie van de algemeene belastingen van 1807. tot deminutie van, en, en de introductie in de douane der welke, betrefft de algemeene belastingen welke de koninklijke commissie van 1807. heeft voorgesteld over 1808.

18 Jan. Koninkrijk der Nederlanden 19 Januarij 1808. Decree van de koninklijke commissie van de algemeene belastingen van 1807. tot deminutie van, en, en de introductie in de douane der welke, betrefft de algemeene belastingen welke de koninklijke commissie van 1807. heeft voorgesteld over 1808.

18 May. Koninkrijk der Nederlanden 12 d. Decree van de koninklijke commissie van de algemeene belastingen van 1807. tot deminutie van, en, en de introductie in de douane der welke, betrefft de algemeene belastingen welke de koninklijke commissie van 1807. heeft voorgesteld over 1808.

18 July. Koninkrijk der Nederlanden 27 July 1808. Decree van de koninklijke commissie van de algemeene belastingen van 1807. tot deminutie van, en, en de introductie in de douane der welke, betrefft de algemeene belastingen welke de koninklijke commissie van 1807. heeft voorgesteld over 1808.

2 Aug. Koninkrijk der Nederlanden 29 Aug. 1808. Decree van de koninklijke commissie van de algemeene belastingen van 1807. tot deminutie van, en, en de introductie in de douane der welke, betrefft de algemeene belastingen welke de koninklijke commissie van 1807. heeft voorgesteld over 1808.

5 Oct. Koninkrijk der Nederlanden 5 Oct. 1808. Decree van de koninklijke commissie van de algemeene belastingen van 1807. tot deminutie van, en, en de introductie in de douane der welke, betrefft de algemeene belastingen welke de koninklijke commissie van 1807. heeft voorgesteld over 1808.

1158  
1803  
Datum  
No  
Handel en scheepvaart

febr	18	1	Rapp. N.D. 1257, verkenning van de toelate a/d algemeent-inspecteur v/d Scheepvaart M. J. v. d. Bosch	b.f.
"	25	1	Rapp. N.D. 134, verkenning van de toelate a/d algemeent-inspecteur v/d Scheepvaart M. J. v. d. Bosch	b.f.
Apr	24	1	Rapp. N.D. 1302, besluit van de S. de Bats tot bevestiging van de toelate a/d algemeent-inspecteur v/d Scheepvaart in buitenspersonen dienst en tot de bevestiging van de toelate a/d algemeent-inspecteur v/d Scheepvaart tot vaststelling van de toelate a/d algemeent-inspecteur v/d Scheepvaart	b.f.
mei	13	1	Rapp. N.D. 1442, verkenning van de toelate a/d algemeent-inspecteur v/d Scheepvaart M. J. v. d. Bosch	b.f.

SAH D N

# From paper to silicon

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



# Scale up!

13

- 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: now 1.5 PB □ towards 10 PB
  - Modern file systems (gpfs)
  - Live 24/7 machine learning

# The pitfall

14

- 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?

15

- 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!

16

- 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:

17

- ... 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!