Optical Character Recognition
A classic problem still unsolved

András Kornai
Plan of the talk

- **Linguistic pattern recognition**: recovering sequence of discrete cognitive units from continuous physical realization
- **The main solutions**: traditional (S-C-I) vs. modern (HMM)
- **The key issue**: reducing the number of parameters by structural decomposition
- **Applications**

András Kornai

Optical Character Recognition: a classic problem still unsolved
Linguistic pattern recognition: from signals to symbols

Input: any linguistically organized signal
Output: string of symbols

1. Speech
2. Sign language
3. Writing

Examples of non-linguistic tasks: face recognition, fingerprint recognition, speaker identification
June 27, 1933.

P. W. HANDEL

STATISTICAL MACHINE

Filed April 27, 1931

1,915,993
Early computer OCR (1960s)

By the 1960s, scanning is viewed as a separate problem, but using a computer is still considered too expensive:

**Highleyman: 1961**

Even if the optimal decision function were known, its implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipment may, in many cases, outweigh the advantages of mechanized categorization.

**As recognized by contemporary OCR:**

Even if the Optimal decision function were known, its implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipment may, in many cases, outweigh the advantages of mechanized categorization.
Linear Decision Functions

There is another practical advantage that is realized by the second approach, namely one of economic feasibility. Even if the optimum decision function were known, its implementation would require, in general, the use of a digital computer or other complex equipment. The cost of such equipment may, in many cases, outweigh the advantages of mechanized categorization. However, if the designer can limit his search to those structures which are economically feasible, and if the optimum structure in this class works well enough for the given purpose, then a technically feasible as well as cost-effective solution is feasible.
Important advances in the 1970s

**Trainable classifiers (Highleyman, Munson)**

Truthed data

- Trainer
- Model

Plain data

- Model
- Hypothesis

**Data normalization (Casey, Nagy)**

Map char area estimated from 2nd central moments to standard area by linear transformation.

Work concentrates on isolated character recognition, running text is still considered too hard, hence the OCR-A font (standardized 1976).

OCR-A 123 ABC
Important advances in the 1980s

- Better image denoising
- Omnifont recognition
- Better page segmentation
- Automated deskewing
- Commercial desktop OCR
- Standardized databases (ETL, NIST, USPS)
- Early commercial vendors: Calera, Recognita, Caere, Kurzweil, ScanSoft
- By now (2011) these are all owned by Nuance
Dimensions of the OCR problem

- dynamic vs. image data
- machine print vs. handwriting
- one vs. many writers/fonts
- isolated letters vs. continuous words
- pre-set vs. free forms
- special vs. general vocabulary
- consistent vs. inconsistent writing
- careful vs. hurried production
- high vs. low resolution
- clean vs. noisy environment
Own work

- Steal technology from speech recognition
- Better language modeling
- Integrated noise modeling
- Feature engineering
- Desktop (WordScan at Calera)
- High volume (postal, bank at IBM Almaden)
- Noisy multilingual (Arabic, Chinese, English faxes at BBN)
The S-C-I pipeline

- Input image
  - Signal Processing
    - Clean image
      - Segmentation
        - Character boxes
      - Feature Extraction
        - Feature vectors
      - Classification
        - Character hypotheses
      - Identification
        - Word and field hypotheses

András Kornai
Optical Character Recognition: a classic problem still unsolved
Main features extracted

- Topological features
- Greyness values
- Directional counts
- Zernike moments
- Subband values
- FFT
Methods used for isolated character recognition

- Template Matching
- Linear Discriminant Analysis
- Artificial Neural Nets
- Radial Basis Networks
- Kohonen Networks (LVQ)
- Nearest Neighbor Methods
- Support Vector Machines

András Kornai  
Optical Character Recognition a classic problem still unsolved
The Hidden Markov approach

input image

Signal Processing

clean image

Feature Extraction

feature vectors

HMM

word and field hypotheses
Difficulties with S-C-I

1. Segmentation decisions are largely context free
2. Segmentation errors propagate to identification stage
3. Global characteristics of font/writing largely lost
4. Language modeling only as postprocessing
5. Components optimized separately
Difficulties with S-C-I

1. Segmentation decisions are largely context free
2. Segmentation errors propagate to identification stage
3. Global characteristics of font/writing largely lost
4. Language modeling only as postprocessing
5. Components optimized separately

It was precisely for these reasons that HMMs replaced S-I-C in speech (though (3) was not solved until the mid-1990s)).

Why didn’t HMMs spread to OCR earlier?
Conversion from 1d signal processing to 2d is not trivial.
Image operations without speech counterparts

- Binarization
- Deskewing
- Zoning
- Connected Component Analysis
- Contour Following
- Skeletonization

András Kornai

Optical Character Recognition: a classic problem still unsolved
Typical bank checks

[Handwritten bank checks with various amounts and payees]

Optical Character Recognition: a classic problem still unsolved
András Kornai  
Optical Character Recognition: a classic problem still unsolved
András Kornai

Optical Character Recognition: a classic problem still unsolved
Impact of language model

Grammar
Grammar & Bigram Probs
Bigram Probs Only
No Language Model

% Within Tolerance

András Kornai
Optical Character Recognition: a classic problem still unsolved
Future directions

- Tangent distance *joint optimization of feature extraction and classification*

- Compressed sensing – almost competitive with state of the art (Ranzato et al 2006 gets 0.6% on NIST, Mairal et al 2008 gets 1%).

- Work planned on mimeographed lecture notes e.g. Fuchs’s *Algebra*