

G. Nagy, Back to the Future (invited, abstract only), Family History Technology Workshop, Salt Lake City UT, February 2012.

BACK TO THE FUTURE

George Nagy
Professor Emeritus, RPI

BACK TO THE FUTURE
George Nagy
Professor Emeritus, DocLab†, RPI

Nearly one hundred years after the invention of OCR, why is it still so difficult to convert some documents to a symbolic, electronic format? I discuss some of the factors that affect the conversion of printed, hand-printed, and handwritten documents. I present three problems that, if solved, would advance significantly the conversion of near-contemporary and cultural-heritage documents: 1. automated, domain-specific feature design, 2. integrated segmentation and classification, and 3. green interaction.

OCR in context

Three open problems:

Features

Segmentation

Interaction

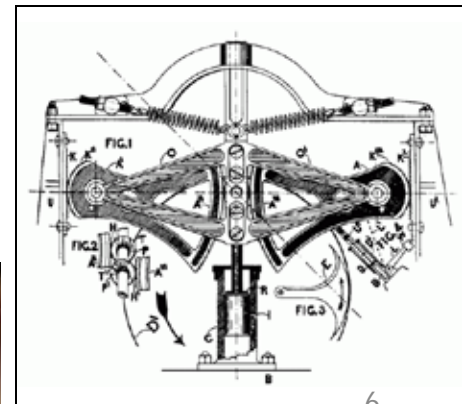
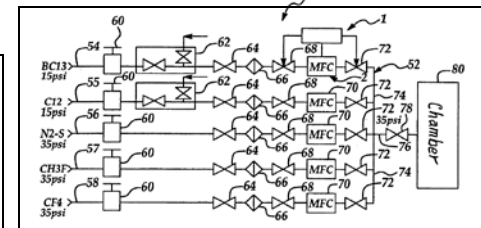
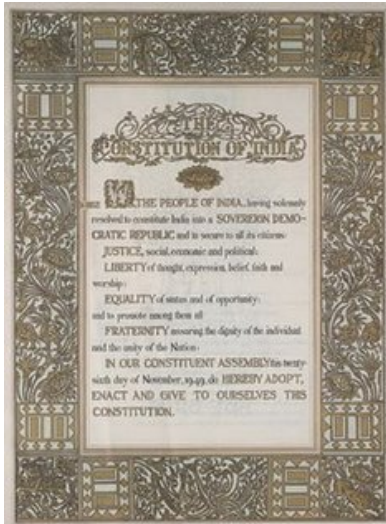
OCR vs. Pattern Recognition & Machine Learning

- Character recognition is an excellent illustration of **pattern classification / machine learning techniques**.
- Many published studies of PR/ML (including mine) use character recognition to illustrate and test algorithms, **which is quite different from building an OCR engine**.
- Very few papers are published by folks who actually build **OCR software**.
- **Feature extraction and classification** are only part of a complete OCR system, but that is all I can talk about.

C

- 1912 patent for telegraph input device
- 1914 patent for **reading-aid** for the blind
- 1931 patent for template matching
- 1938 IBM patents optical scanning
- 1950 Shepard demonstrates GISMO, patents feature extraction, reads typewritten letter, predicts 99.9% accuracy
- 1955 **First commercial OCR installation at Readers Digest**
- 1959 Readers Digest OCR reads its billionth character
- 1963 Nagy joins IBM Research Division
- 1965 First postal address reader
- 1966 OCR-A introduced and IBM 1975 delivered to the SSA
- 1970 IBM 1288 multifont page reader
- 1977 Machines cost **300K\$ – 2M\$**, 2K forms/minute, 2K chars/sec
- 1980 70,000 REI Wands for stylized characters
- 2012 **OCR software for < \$400, >4000 published papers**
Some Open Source development: GOCR, Tesseract , IMPACT







There are *billions* of documents



3/2/2012

<-- Future

Some current OCR applications

- Postal address reading 
- Check reading  
- Forms processing 
- Back issues of technical journals 
- Books (> 10% titles digitized) ?
- Legal material
- Historical records: transcription, annotation, indexing 
- Reading aids for the blind
- Personal and mobile OCR
- Scene OCR

Alphabets and Scripts

- Bank checks, tax returns 14
- Forms, letters, books ~100
- European Patent Office ~600
- Han, Kanji ~4K – 10K

Most scripts have no upper/lower case,
but many have allographs

Devanagari & Arabic have about 100 glyphs.



OCR, MICR, OMR, ICR, SCR

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Y J H \$ % | & * { } - + ■ : " , . ?
1 2 3 4 5 6 7 8 9 0 - = £ ; ' , . /

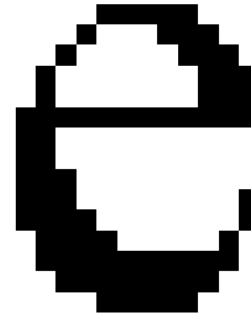


MICR E13 B Font



Bilevel patterns

Here is a 16 x 12 bitmap:



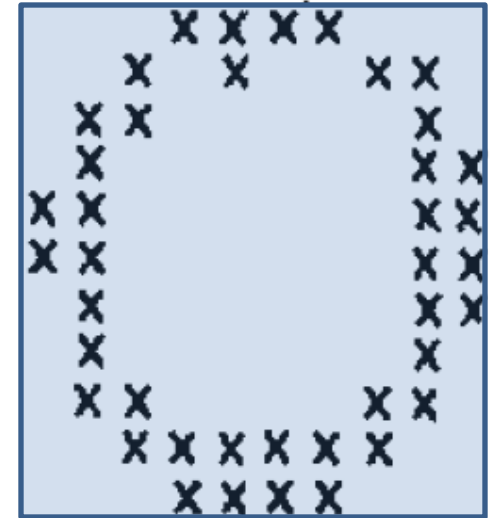
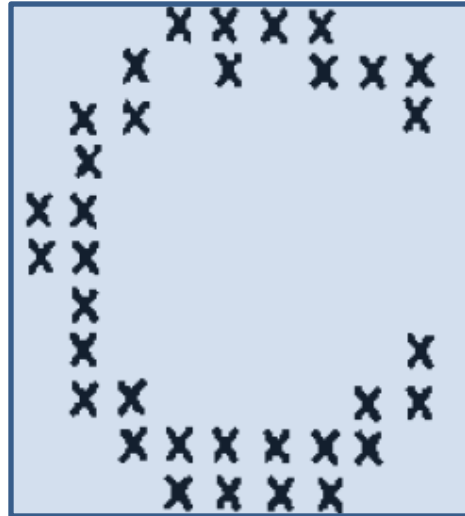
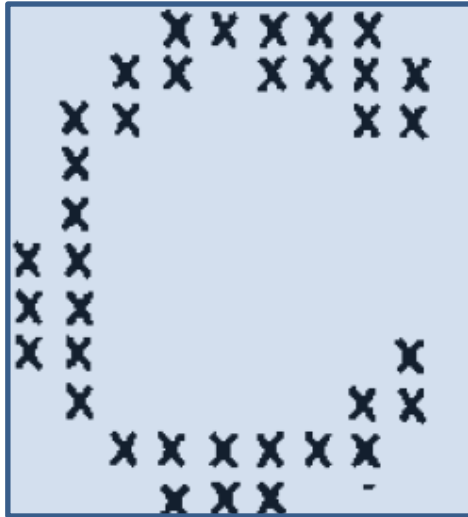
The number of possible bitmaps of this size is about

627 710 000 000 000 000 000 000 000 000

000 000 000 000 000 000 000 000 000 000

A classifier ought to know how to classify any of them!

A nine-pixel difference



Same difference, with or without a distinction

A distinction with a very small difference: **1/1**

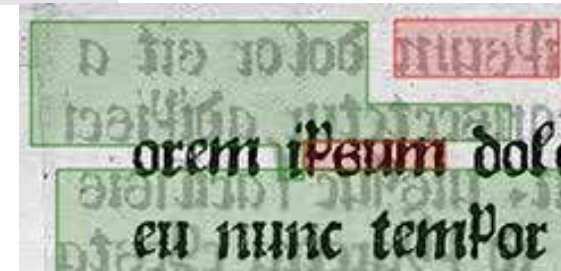
Factors that affect accuracy of OCR on **printed** matter

- Typeface (**serif**, sans-serif, display) and type size
- Style (regular, tight, **bold**, ligatures, *italic*, *subscript*)
- Quality of original
(age, paper, printing and copying technology)

Beyond
our
control

Bleed through

<http://njournet.com/images/docueval.png>



- Sampling rate (pixels per character height)
- Bit depth (grayscale or bilevel) **Quality of digitization**
- **Preprocessing**
- Features
- Classifier
- **Post-processing**

What is hard about printed character recognition?

GQg GQg GQg GQg GQg
GQg GQg GQg GQg GQg

GQg GQg GQg GQg GQg
GQg GQg GQg GQg GQg

GQg GQg GQg GQg GQg
GQg GQg GQg GQg GQg

Even perfect characters have significant variations in shape, stroke thickness, and topology

COMMEND
sites
1/4-lb.
Fig. 1b.
Congressional
Reporters

Common OCR errors

Similar characters:

1 | l | i | / e c o a s b 6 g q 9 r n h

Splits: m à rn

Mergers: rn à m

Super/subscripts:

Punctuation: . , : ; *e*⁻² *d_{m1}* " "

mis-segmentation

Most OCR errors are due to mis-segmentation

Hand-printing and *handwriting*

- Scribes (Latin, Chinese, Arabic) are consistent.
- In the US, hand-printing now rare (mainly forms, diagrams).
- It can be read by machine when there is plenty of context.
- In the last hundred years, most handwritten text self-read,
 - so low premium on legibility.
- No commercial OCR **for unconstrained** handwritten, omni-writer *text*.
- Single-writer script much easier. Train both machine & writer!.
- *Spontaneous* printing/writing is very different from *copying*.
- Most public test data sets consist of copies.
- *Word spotting*: **Spitz, Srihari, Hull, Manmatha, Barrett, ...**
- *Alignment* with transcripts.

Constraints that help

- On-line rather than off-line (with immediate feedback)
 - Dual entry (Anoto)

- Block letters rather than *cursive*

- SINGLE CA

Structured	and	semi-structured	forms
------------	-----	-----------------	-------

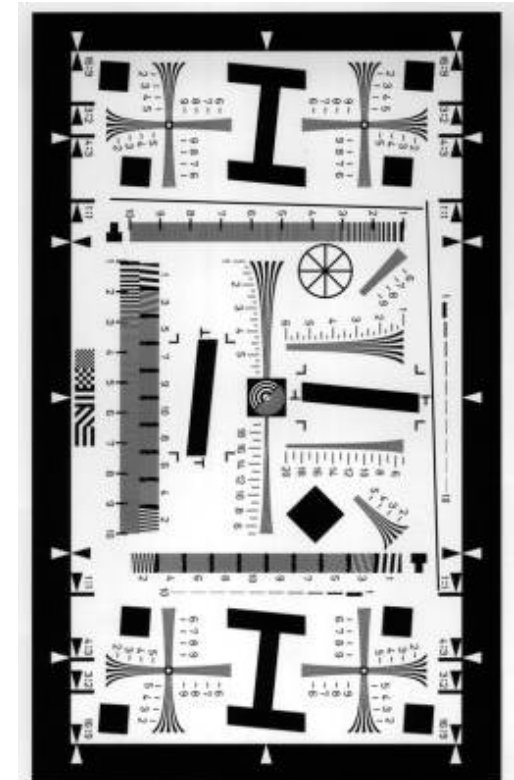
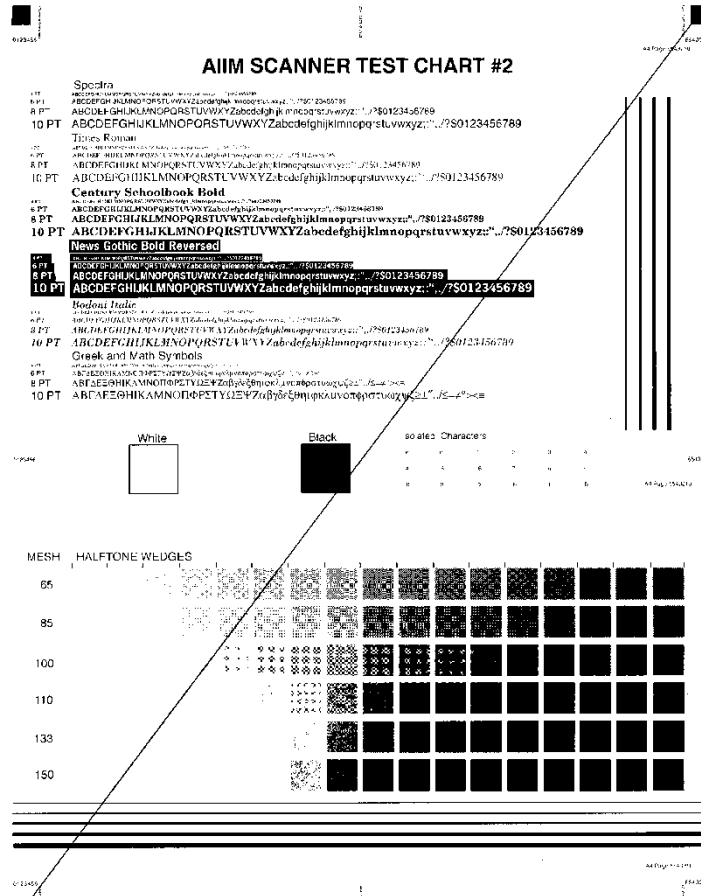
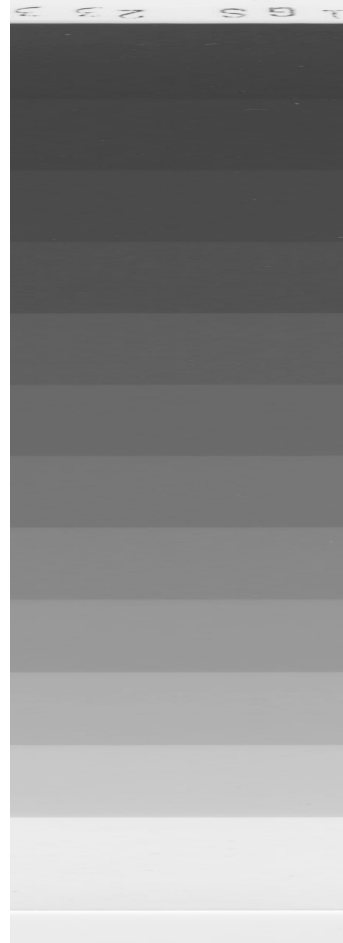
- Drop-out ink

- **Registration** (skew), index (line and column), tick (comb) marks

- Domain-specific text
- Restricted vocabulary
- Redundancy (repetition, totals, city/state/zip, ...)
- Careful scanning with quality control



Scan test charts for before every batch!



Formulate and heed specifications!

Family & Church History Department Digital Image Specification

Version 1.9a

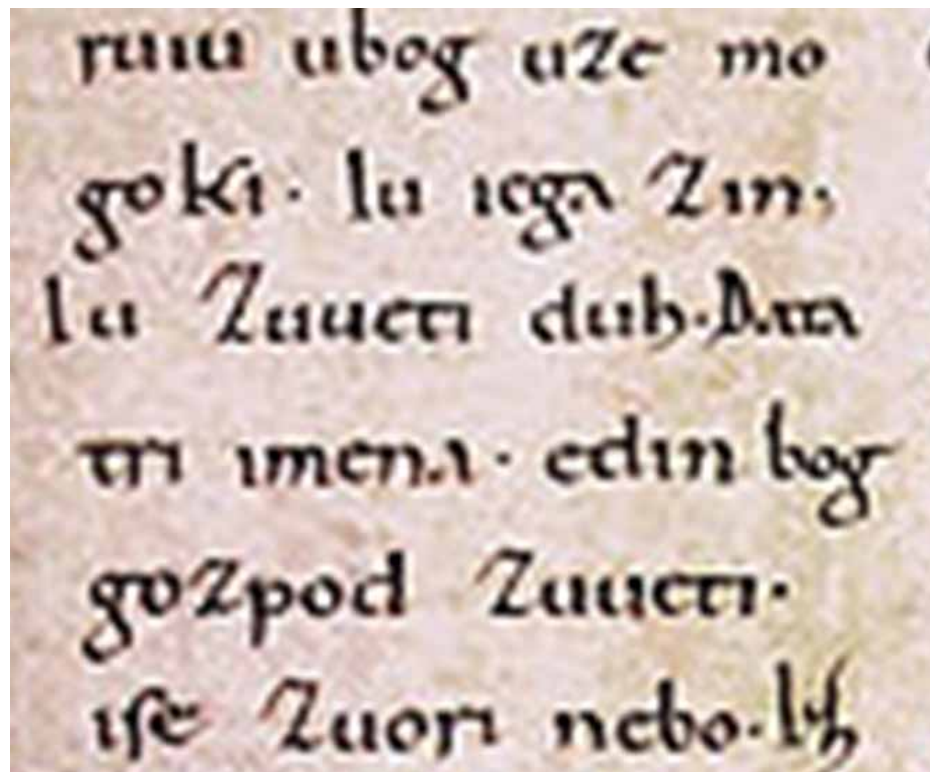
Revised 30 Apr 2007

(Last version is 1.8b, content edits to sections 16 through 21.)

(45 pages)

Cf. also E. Barney Smith, H. Baird, W. Barrett, F. Le Bourgeois, X. Lin, G. Nagy, and S. Simske, "DIAL 2004 Working Group Report on Acquisition Quality Control", Procs. 2nd IEEE International Conference on Document Image Analysis for Libraries, Lyon 2006.

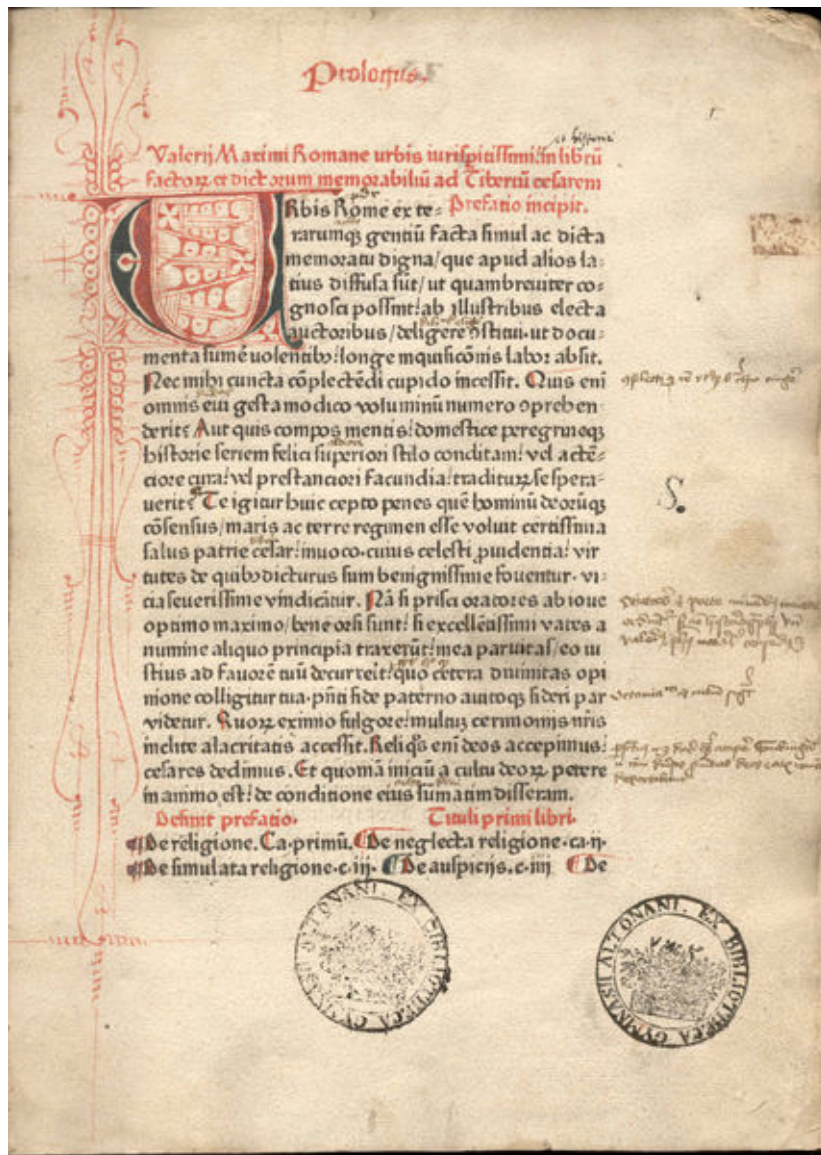
Old print is not that different
from old manuscripts



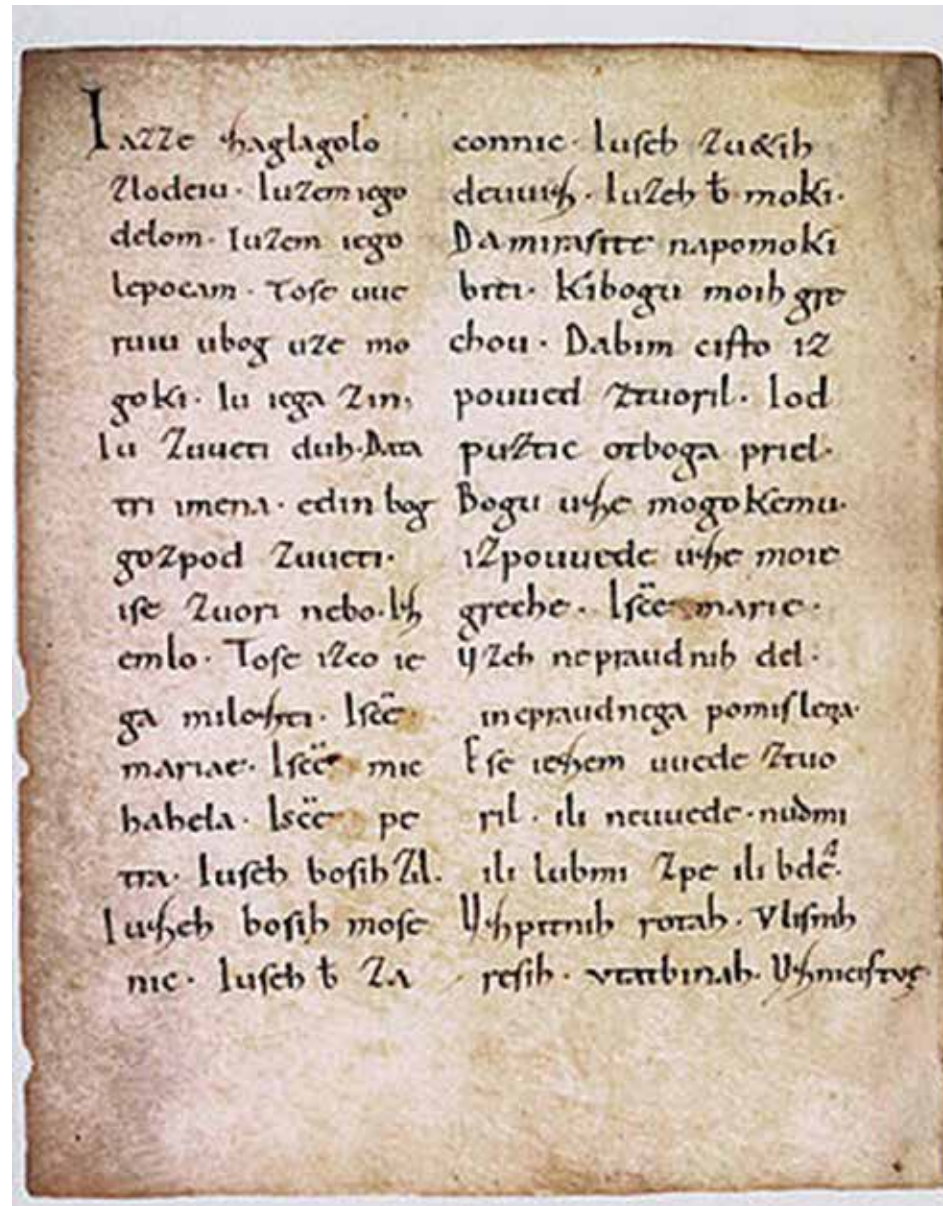
The **Freising Manuscripts** (also *Freising Folia*, *Freising Fragments*, or *Freising Monuments*; Slovene *Bržiński spomeniki*, Latin *Monumenta Frisingensia*) are the first Latin-script continuous text in a Slavic language and the oldest document in Slovene
http://en.wikipedia.org/wiki/Freising_Manuscripts.



http://en.wikipedia.org/wiki/Valerius_Maximus



Incunabulum (<1500)



manuscript

19th Century census form (from DWE)
Many forms written by same census-taker!

U.S. DEPT. OF COMMERCE
TWELFTH CENSUS OF THE UNITED STATES.

108 A

SCHEDULE No. 1.—POPULATION.

State Illinois County Cook Township or other division of county West Town Name of incorporated city, town, or village, within the above-named division Chicago Name of institution John Marshall Supervisor's District No. 1 School No. 6 Enumeration District No. 536 Ward of city, 17th
Enumerated by me on the 5th day of June, 1900. Examiner

LOCATION	NAME	RELATION	PERSONAL DESCRIPTION	SEXES			CITIZENSHIP	EDUCATION, TRADE OR BUSINESS		EDUCATION	PROFESSION OR INDUSTRY
				Male	Female	Total		At each year from age 15 to 24	At each year from age 25 to 34		
1	2	3	4	5	6	7	8	9	10	11	12
1	Harry	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
2	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
3	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
4	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
5	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
6	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
7	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
8	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
9	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
10	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
11	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
12	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
13	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
14	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
15	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
16	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
17	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
18	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
19	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
20	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
21	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
22	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
23	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
24	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
25	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
26	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
27	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
28	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
29	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
30	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
31	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
32	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
33	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
34	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
35	John	Boys	2 Nov. 1897 7	White	White (Cns)	White (Cns)	1897 7	At school			
36	John	Boys	2 Nov. 1897 7	White							

[illegible]

1941 Death Certificate (from DWE)

We know what to expect in each box!

**U. S. DEPARTMENT OF COMMERCE
BUREAU OF CENSUS**

**STATE OF OHIO
DEPARTMENT OF HEALTH
CERTIFICATE OF DEATH**

1 PLACE OF DEATH
County Wyandot Registration District No. 1416 File No. 58302
Township Paris Primary Registration District No. 3507 Registered No. 36
or Village Paris Ohio No. _____ St. _____ Ward _____
(If death occurred in a hospital or institution, give its NAME instead of street and number)

2 FULL NAME Alvioleta Wahlgrenmuth Did Deceased Serve in U. S. Navy or Army _____
(a) Residence. No. RFD #2 Paris Ohio Ward DCT 1941
(Usual place of abode) (If nonresident give city or town and State)

PERSONAL AND STATISTICAL PARTICULARS

3. SEX Female 4. COLOR White 5. SINGLE, MARRIED, Write the word Married
or Widowed or Divorced
6. DATE OF BIRTH (month, day, and year) April 21 1876
7. AGE (years) Months Days 65 6 1
8. Trade, profession, or particular kind of work done, as spinner, weaver, bookkeeper, etc. Housewife
9. Industry or business in which work was done, as mill, mine, saw mill, bank, etc. _____
10. Date deceased last worked at this occupation (month and year) _____ 11. Total time (years) spent in this occupation _____

12. BIRTHPLACE (city or town) Wyandot Co Ohio
13. NAME Milton T. Wahlgrenmuth
14. BIRTHPLACE (city or town) Wyandot Co Ohio
15. MAIDEN NAME Jarah Beak
16. BIRTHPLACE (city or town) Hancock Co Ohio
17. The Signature of Informant Jugan Wahlgrenmuth
and (Address) Paris Ohio
18. BURIAL, CREMATION, OR REMOVAL Place St Paul Date Oct 24 1941
19. FUNERAL FIRM W. H. T. Funeral Home
20. BURIED BY P. E. Dringer Lic. No. 761
21. EMBALMER Paul C. Hoff Lic. No. 3307
22. FILED 11/1/1941 Jarah Beak

MEDICAL CERTIFICATE OF DEATH

21. DATE OF DEATH (month, day, and year) 9-22 1941
22. I HEREBY CERTIFY, That I attended deceased from 5-15 1941 to 9-22 1941
I last saw him alive on 9-11 1941
to have occurred on the date stated above at 9-11 P.
The PRINCIPAL CAUSE OF DEATH and related causes of death must state as follows:
Libro poisoning - above
skin metastasis
55B
CONTRIBUTORY CAUSES of importance not related to principal cause:
Name of operation _____ Date of _____
What test confirmed diagnosis? _____ Was there an autopsy? _____
23. If death was due to external causes (violence) fill in following:
Accident, suicide, or homicide? _____ Date of injury _____
Where did injury occur? _____ (Specify city or town, county)
Specify whether injury occurred in industry, in home, or in _____
Manner of Injury _____
Nature of Injury _____
24. Was disease or injury in any way related to occupation? No
If so, specify _____ (Signed) P. E. Dringer
Date 10-24 1941 Address Forest St

DEATH RECORD

25. BIRTHPLACE (city or town) Paris Ohio
26. STATE OR COUNTRY Ohio
27. IDENTIFYING NAME Jarah Beak
28. BIRTHPLACE (city or town) Hancock Co Ohio
29. STATE OR COUNTRY Ohio
30. MAIDEN NAME Jugan Wahlgrenmuth
31. ADDRESS Paris Ohio
32. CREMATION, OR REMOVAL Place St Paul Date Oct 24 1941
33. FUNERAL FIRM W. H. T. Funeral Home
34. BURIED BY P. E. Dringer Lic. No. 761
35. EMBALMER Paul C. Hoff Lic. No. 3307
36. FILED 11/1/1941 Jarah Beak

Even clear handwriting is difficult to read without context

rendezésen ~~sz~~ munka (bár az idők premier aug 28,
 és lehetünk) deget együtt. Kérlek, gondold Te is
 a deget és így mind hamarabb követhető címe:
 Maria Karin, 2 Hamburg - Schenefeld, Dietrichsenweg 8
 Juli 20-ig kb. áll lenni, talán még egy hely.
 A telefon: 0411 (az Hamburg száma) 8304324.
 Augustus első felében még folyamatosan mives, hol lenni,
 szeretnék pár magyar Salzburgban menni, de anyagilag elég
 rosszul állunk és ezt hisszük. De is lehet, hogy

Teknik penun 21-12-19

D-64380 ROSSDORF (BEI DARMSTADT)

Felion an eddigi telefonain, de may new nyth
bich, hoy er manad twibly is:

06154 / 8556



George Nagy and **William Barrett** at 2nd IEEE International Conference on Document Image Analysis for Libraries in 2006.

Given all these givens,

Professor Bill. Barrett asked me to talk about three (3) problems that, if solved, would get us to where we want to be. Here goes:

1. Automated application-specific feature design
2. Integration of segmentation and classification
3. Green interaction

Professor Cheng-Lin Liu,
Director of the Pattern Recognition Laboratory of the
Institute of Automation of the Chinese Academy of Science
and Executive Chair of ICDAR 2011, suggested:

1. Unconstrained handwriting recognition
2. Scene image (and video) recognition
3. Historical documents

Dr. Liu lists *applications*, while I list *methods*.

1. FEATURES FOR OCR

Preprocessing or invariant features?

Resampling inevitably introduces **noise**.

It is better to devise
size, skew, and stroke-width **invariant features**.

Desirable feature invariants

Scale **G** **G** **G** **G**

Reflectance: **G** **G** **G** **G**

Geometry: **G** **G** **G** **G** **G**



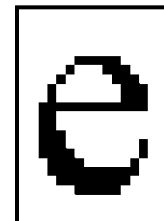
Topology: *G* **G** **G** **G** **G**

But, p and d, 6 and 9.

Feature extraction

Feature extraction maps

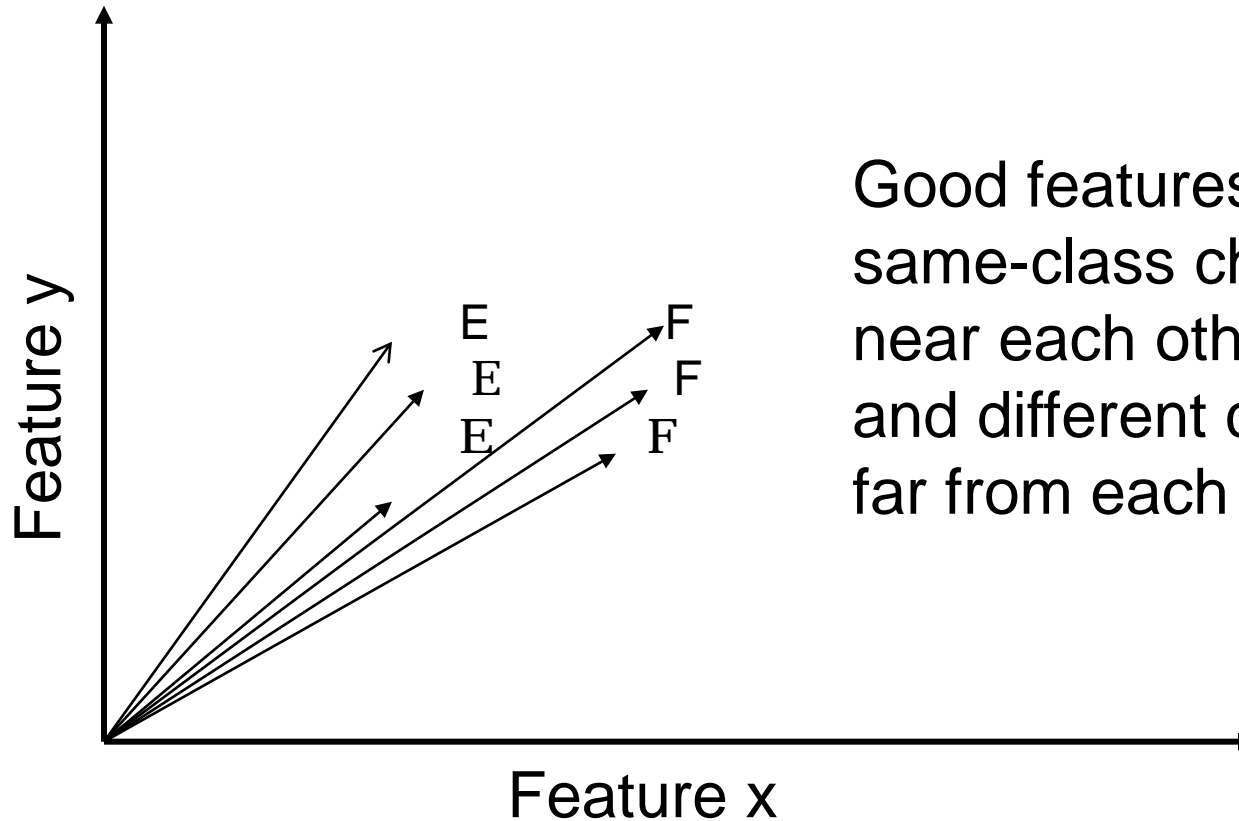
a **variable** size pixel map or bitmap
to a vector with a **fixed** number of elements.


$$\begin{pmatrix} 2 \\ 3 \\ 0 \\ 8 \\ 5 \\ 2 \\ 4 \\ 9 \\ 0 \\ 4 \\ 8 \end{pmatrix}$$

Features should preserve only **significant differences** between classes

F E 1 1

A 2-D Feature Space



Good features map
same-class characters
near each other,
and different classes
far from each other.

Instead of 2 features, 50 – 500 features may be used.

Some OCR features

- Templates
- Moments and moment invariants
- Directional gradients
- Local shapes
- Connectivity: line-adjacency graph (LAG)
- Stroke lists
- Projections and profiles
- N-tuples
- Complex boolean logics
- Autocorrelation functions
- Fourier and other orthogonal transform coefficients
- Etc., etc., etc. ...

Folks love to invent new features. We don't know what the best features are, but some are definitely worse than others.

Hand-crafted features are often specific to a given alphabet, typeface, and size.

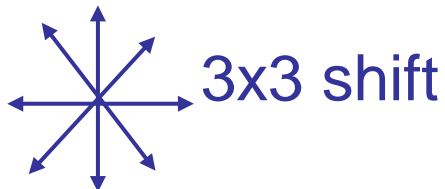
Templates

[illegible]

Ternary templates

由山而面

Confusion pairs



3/2/2012

<-- Future

[illegible]

```

N
N
N
NO      O
N        O
N          X
N    O      O
N          O
N    X      X
N    OX     O
N    O      O
N          O
N          G
N          CO
N          X
N    X      C
N    X      O
NCO      X
N    X      X
N    XX     X
N    X      C
N          O
XXXXXXXXXXXXXXXXXXXXXXXXXXXX

```

```
N
N
N
N
N
N   CO      XX          XX
N
N           x         X
N
N               x       xx
N     O        X             x
N       X
N     O   CO    X      O
N   x       O
N       GO
N     X  X      .      X
N   O
N     O   O      O      O
N
N       X      O   C  O
NXX      CO  C  O
N     O      X      x
N           X
```

Peephole templates

Templates as preclassifiers for large alphabets

降 陷 限 院 阿
隱 隨 陳 隔 陣

Fig. 5. Characters having the same radical. The radical appears on the left-hand side of the character.

THE MASK FOR RADICAL 3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
B1T 34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1T 23	2	3	0	0	2	4	2	0	0	0	1	4	6	6	2	0	0	1	3	3	0	0	0	0	0
B1T 22	9	17	9	5	6	10	13	7	0	1	5	15	23	23	18	14	4	6	10	13	10	1	0	0	0
B1T 21	45	70	54	42	40	60	53	36	22	19	29	38	61	59	56	51	31	34	33	35	31	9	1	0	0
B1T 20	96	126	115	91	110	124	121	78	52	52	60	69	88	91	89	80	74	77	75	62	49	32	5	0	0
B1T 19	111	138	121	81	106	133	131	101	54	61	70	87	99	99	88	83	93	93	100	71	48	32	6	1	0
B1T 18	103	137	110	37	107	131	159	49	48	76	93	91	82	80	66	71	88	96	79	60	41	25	5	0	0
B1T 17	89	136	92	30	120	96	31	15	42	77	74	66	71	70	58	51	77	90	77	50	39	17	10	0	0
B1T 16	78	135	84	35	109	49	7	16	55	91	77	61	65	70	56	59	71	97	83	58	49	26	12	0	0
B1T 15	67	134	78	30	93	46	15	21	57	84	80	61	63	68	69	78	81	98	84	73	68	40	11	0	0
B1T 14	67	134	80	23	83	28	36	40	62	83	97	88	82	88	100	100	100	94	86	82	69	37	4	0	0
B1T 13	64	135	80	18	58	113	78	36	47	79	94	86	78	95	100	99	82	87	93	77	66	21	0	0	0
B1T 12	61	136	84	11	27	111	115	55	56	81	89	63	65	85	93	84	80	89	96	80	56	14	2	0	0
B1T 11	60	135	98	24	37	110	130	93	75	97	97	62	65	99	102	78	77	88	100	97	74	19	1	0	0
B1T 10	66	136	100	41	63	120	132	95	86	99	83	64	74	100	100	87	79	95	100	95	74	33	5	2	0
B1T 9	67	136	100	64	93	125	117	69	62	92	81	72	68	100	101	88	66	67	73	73	49	21	9	2	0
B1T 8	61	136	100	58	100	119	79	35	41	83	93	57	50	84	86	83	62	86	63	58	38	16	4	0	0
B1T 7	60	130	109	33	72	80	38	21	41	94	84	54	48	71	75	62	70	82	81	63	36	21	5	1	0
B1T 6	64	130	100	18	19	31	25	16	68	103	88	33	48	79	72	52	57	79	92	79	60	37	10	1	0
B1T 5	68	131	104	14	7	19	15	31	72	93	56	19	42	87	81	39	30	59	69	84	74	47	14	0	0
B1T 4	73	128	109	25	9	12	35	42	63	77	56	24	53	91	86	47	46	61	75	96	87	59	19	1	0
B1T 3	95	134	120	36	15	30	36	43	60	63	58	45	58	89	91	70	71	78	85	94	89	56	14	1	0
B1T 2	101	135	118	32	23	30	30	36	46	53	42	33	54	70	78	73	67	72	72	74	61	32	5	0	0
B1T 1	71	106	73	17	14	12	13	13	35	38	18	15	31	49	45	33	40	39	36	41	31	13	1	0	0

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Preclassifiers separate **groups** of classes

Geometric (Hu) moments

$$M^{pq} = \sum_i \sum_j i^p j^q X_{ij}$$

i is the row index, j is the column index

M^{00} is the number of foreground pixels

If $M^{01} = M^{10} = 0$, then

M^{02} is large for **fat** patterns, M^{20} for **tall** patterns
(moments of inertia are always positive)

M^{11} can be positive (/) or negative (\)

There are 4 third moments: M^{03} , M^{12} , M^{21} , M^{30} (symmetries)

Central Moments

(normalization for position, size, and order)

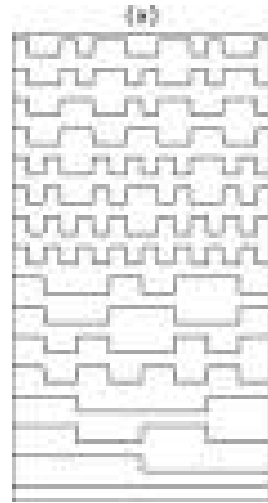
$$m^{pq} = \frac{1}{(M^{00})} \sum_i \sum_j (i - i_0)^p (j - j_0)^q X_{ij}$$

where $i_0 = \frac{M^{10}}{M^{00}}$, $j_0 = \frac{M^{01}}{M^{00}}$, i.e., the centroid of the pattern

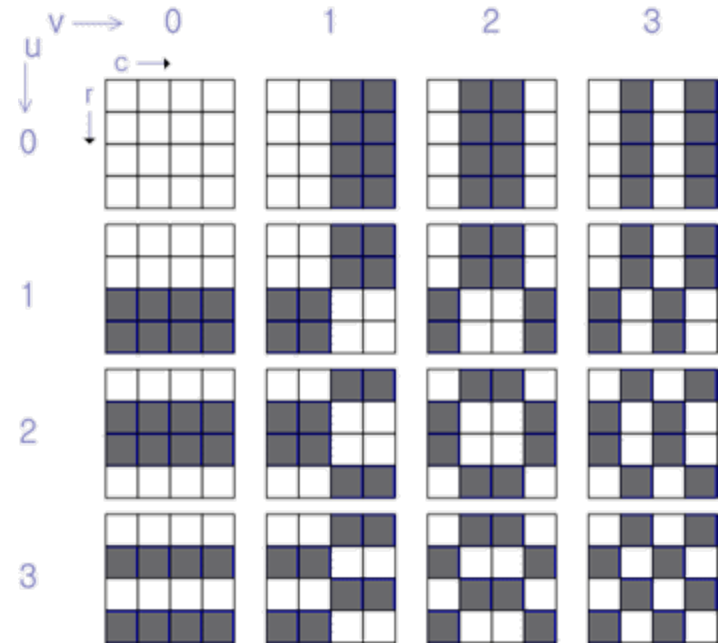
Central moments are **invariant** to size and **translation**.
Hu ratios are also invariant to **rotation**.

Other transforms (of area or contour)

- complex Zernike moments (reconstruction easier)
- Discrete Fourier (x-y or r-q)
- Haar (wavelets)
- Hadamard
- Walsh
- Rademacher



1-D

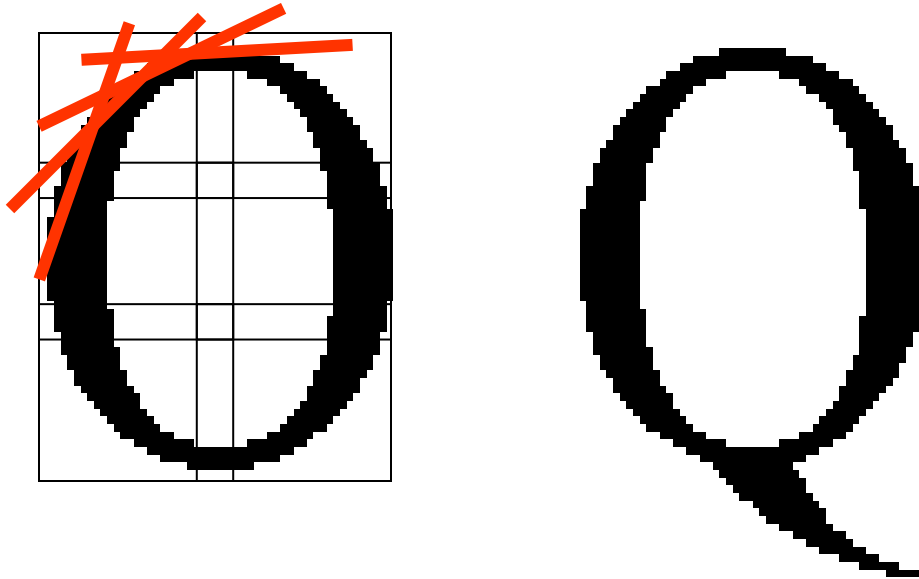


2-D

All discrete orthogonal
transforms with different
basis functions

for classification, only *fixed* low-order coefficients kept;
for compression, only *significant* coefficients kept.

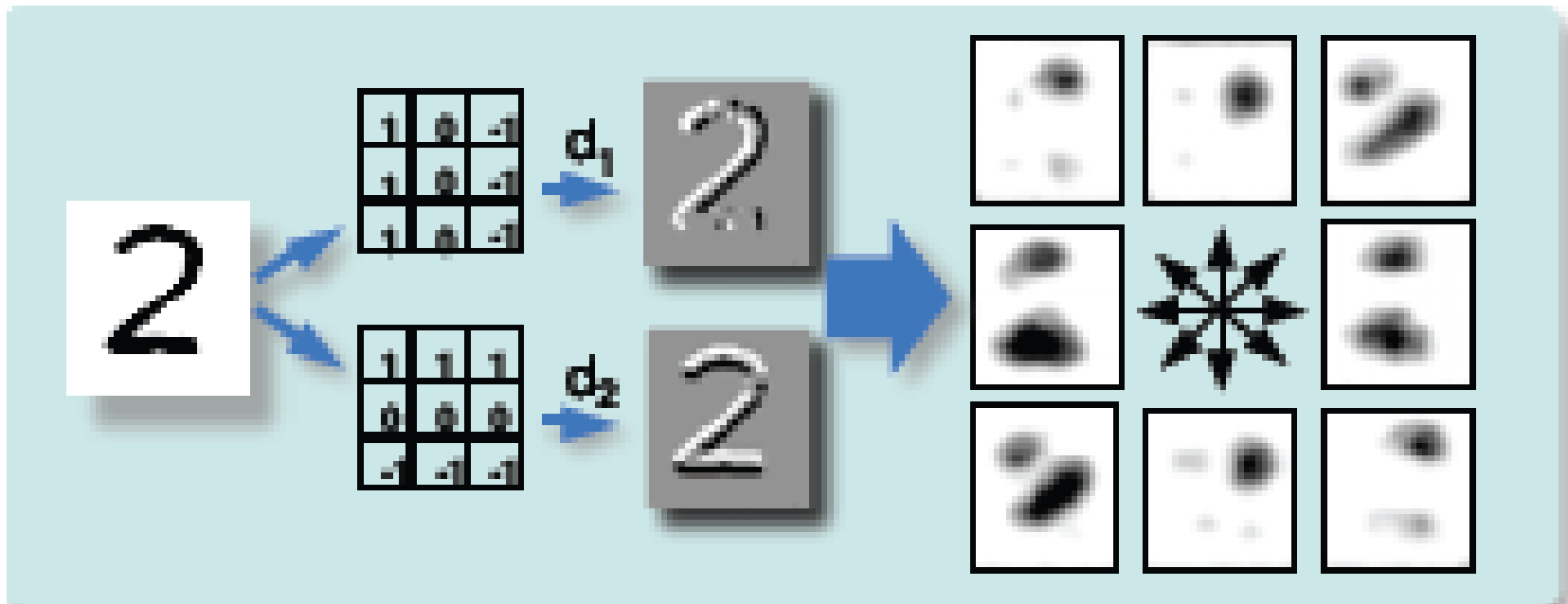
Zoned Directional Features



Extracted from contours (chain code), or center line (*skeleton*), or grayscale **gradient** histograms.

4-16 **directions**, 2 **orientations**, 6-20 **zones**, à $D = 50-200$.

Directional Gradient Features (shown without zones)



H. Fujisawa, ICDAR 07

Stroke-based features

(1) Segment strokes

(2) Extract

- length or endpoints

- location or scan order

- curvature

- orientation/direction

- stroke crossings

More useful for handprint than typeset characters,
because strokes are not well defined in most typefaces.

Local shape features

lakes

o p b

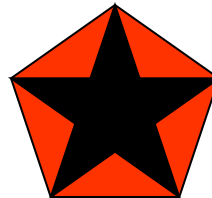
bays

c s u

lids

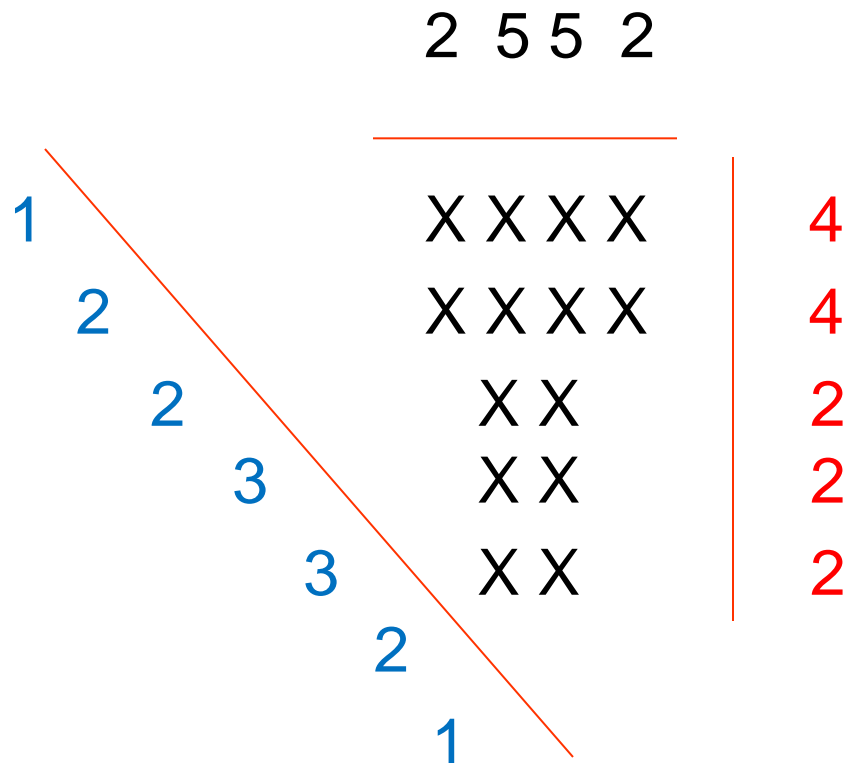
e a

convex deficiency



Pixel projections

Directional Projections

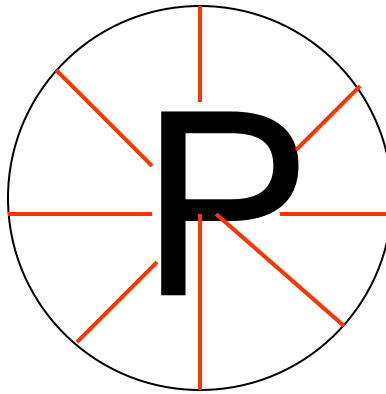


$$\mathbf{V} = (2 \ 5 \ 5 \ 2 \ 4 \ 4 \ 2 \ 2 \ 2 \ 1 \ 2 \ 2 \ 3 \ 3 \ 2 \ 1)$$

3/2/2012 <-- Future

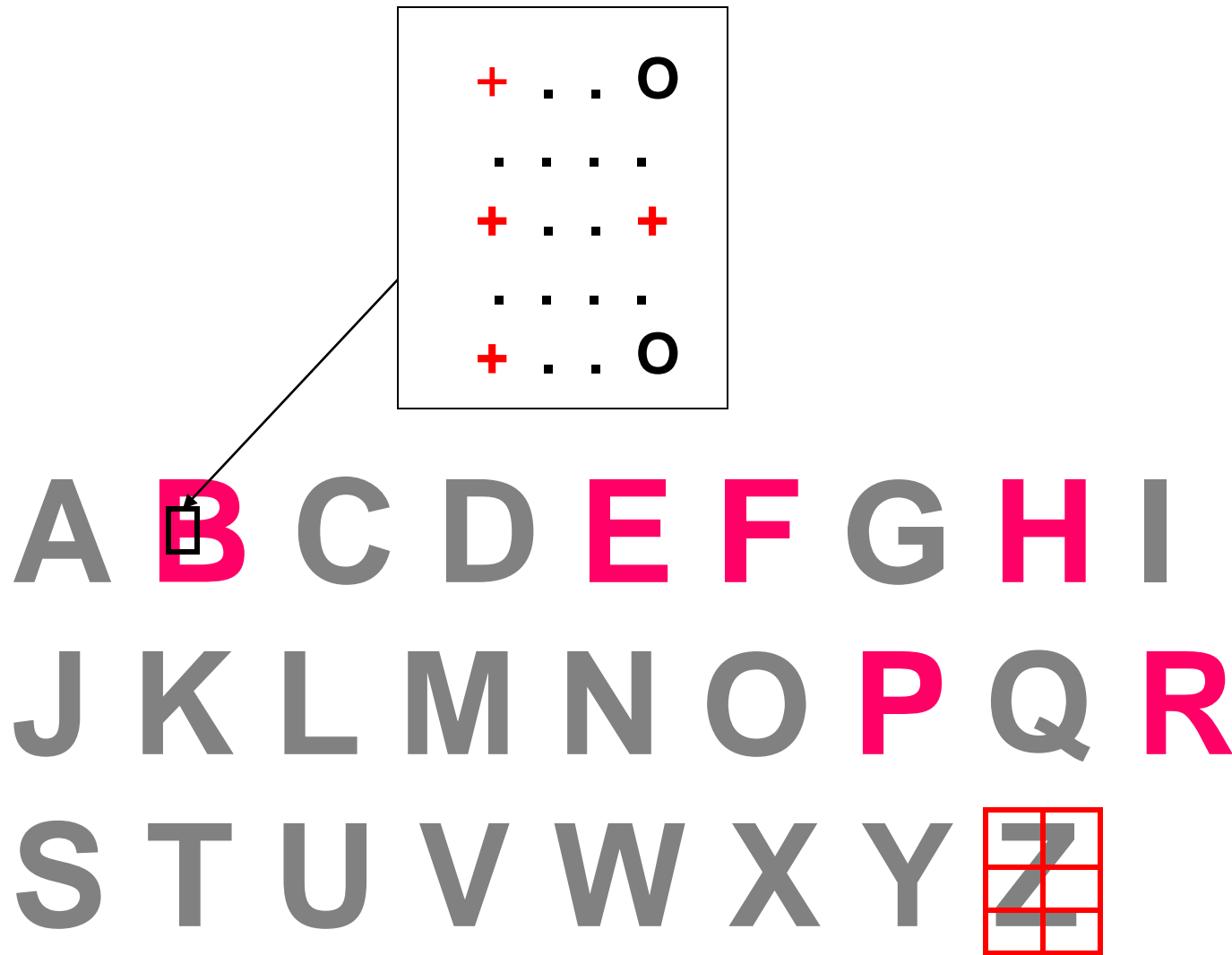
Perimetric profile

Circle centered on
centroid or median

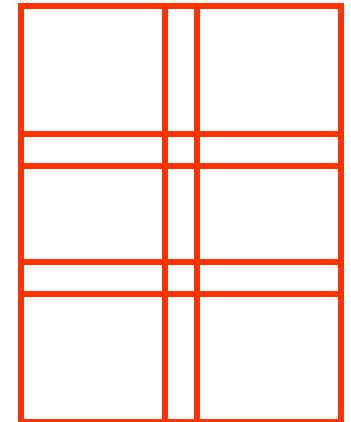
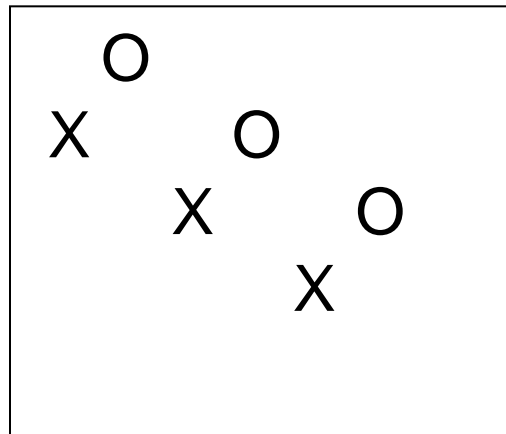
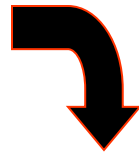
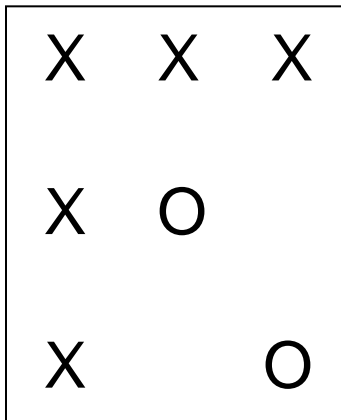
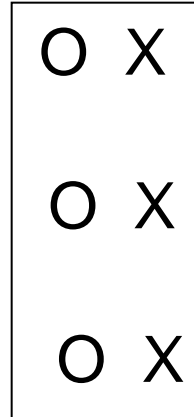
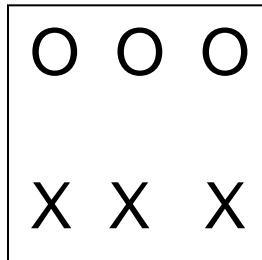
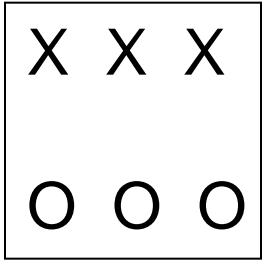


$$V = (5 \ 6 \ 7 \ 8 \ 9 \ 5 \ 8 \ 7)$$

N-TUPLE FEATURES (OR CLASSIFIERS?)



Directional n-tuples



$$6 \times (2 + 2 + 8 + 4) = 96$$

Why are there so many different kinds of features?

Because what features are best depends not only on the **data** but also on the **classifier**.

So we need application-specific, automated feature design!

Current automated algorithms do only **feature selection**:

- (a) Add the feature that decreases the error rate most, or
- (b) Eliminate the feature that has least effect on the error rate.

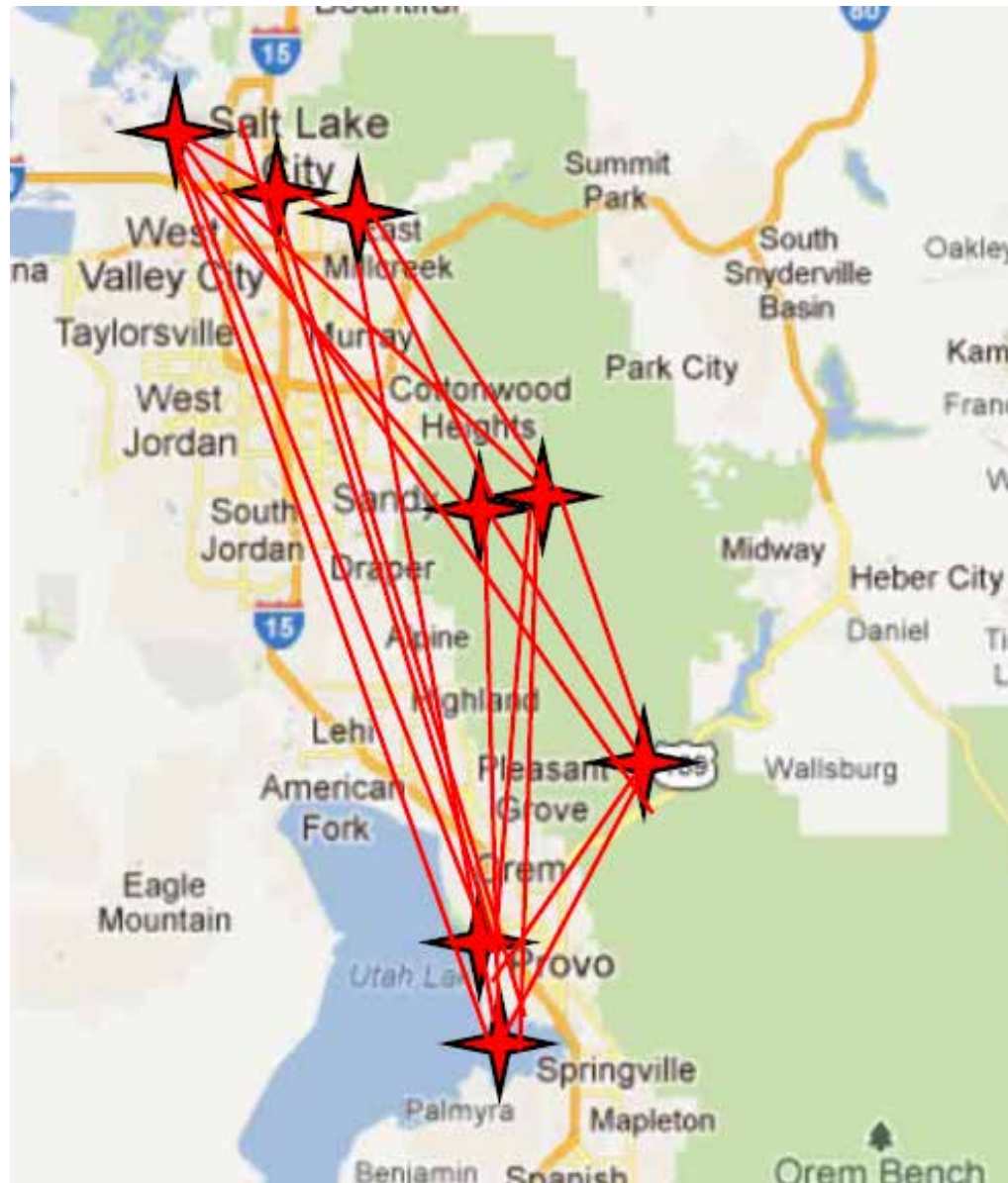
But the strongest team may not consist of the best individual players.

Even feature selection is NP-complete L

Furthermore, too many features increase the error rate.

This is the *Hughes phenomenon*.

N = 8



Automated feature design - is there hope?

The traveling salesman problem is also NP-complete.
Yet many transportation networks are quite efficient.
There exist also theoretically sound bounded approximations.

Academics shy away because a universal optimal solution is unlikely, and experimentation requires a plethora of data and classifiers. Nevertheless, it is a wonderful problem for supercomputers!

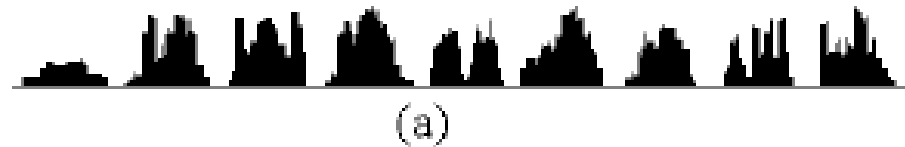
So far we have discussed only the classification of single, isolated characters!

2. SEGMENTATION

CLASSIFICATION

PRINTED CHARACTERS

二值图像的噪声消除



print



following mid zone



Methods:

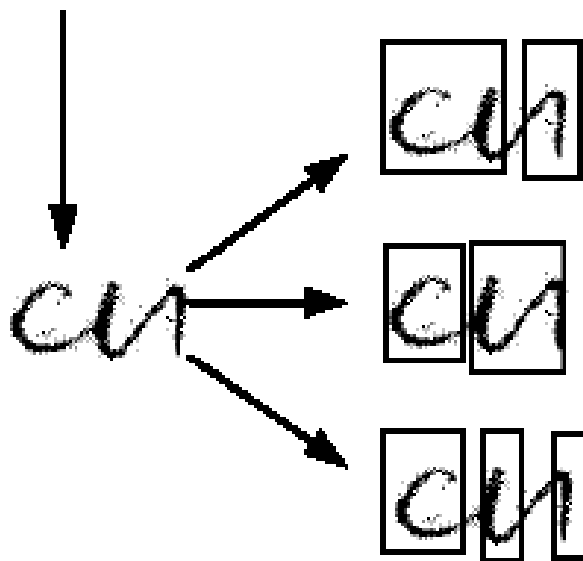
Projections on baseline
Connected components
Profile analysis
Whitespace analysis

...

Li OT
LT

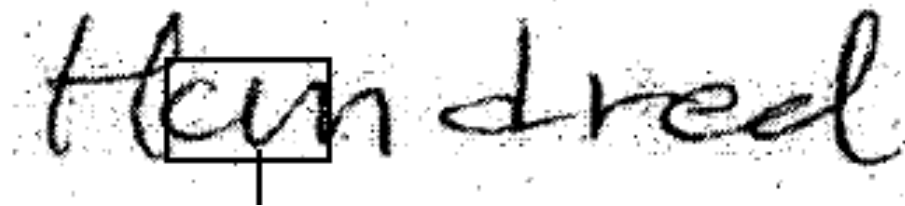
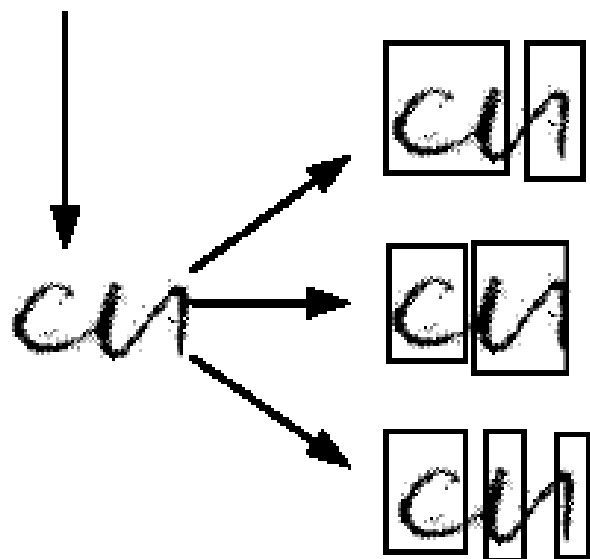
Handwriting

A segment from a word:



"w", "ui", "iu", or "iii" ?

A segment from a word:



NOT "w", "ui", "iu", or "iii" !

CHARACTER RECOGNITION SYSTEMS

A Guide for Students and Practitioners

Mohamed Cheriet

École de Technologie Supérieure/University of Quebec, Montreal

Nawaf Kharm

Concordia University, Montreal

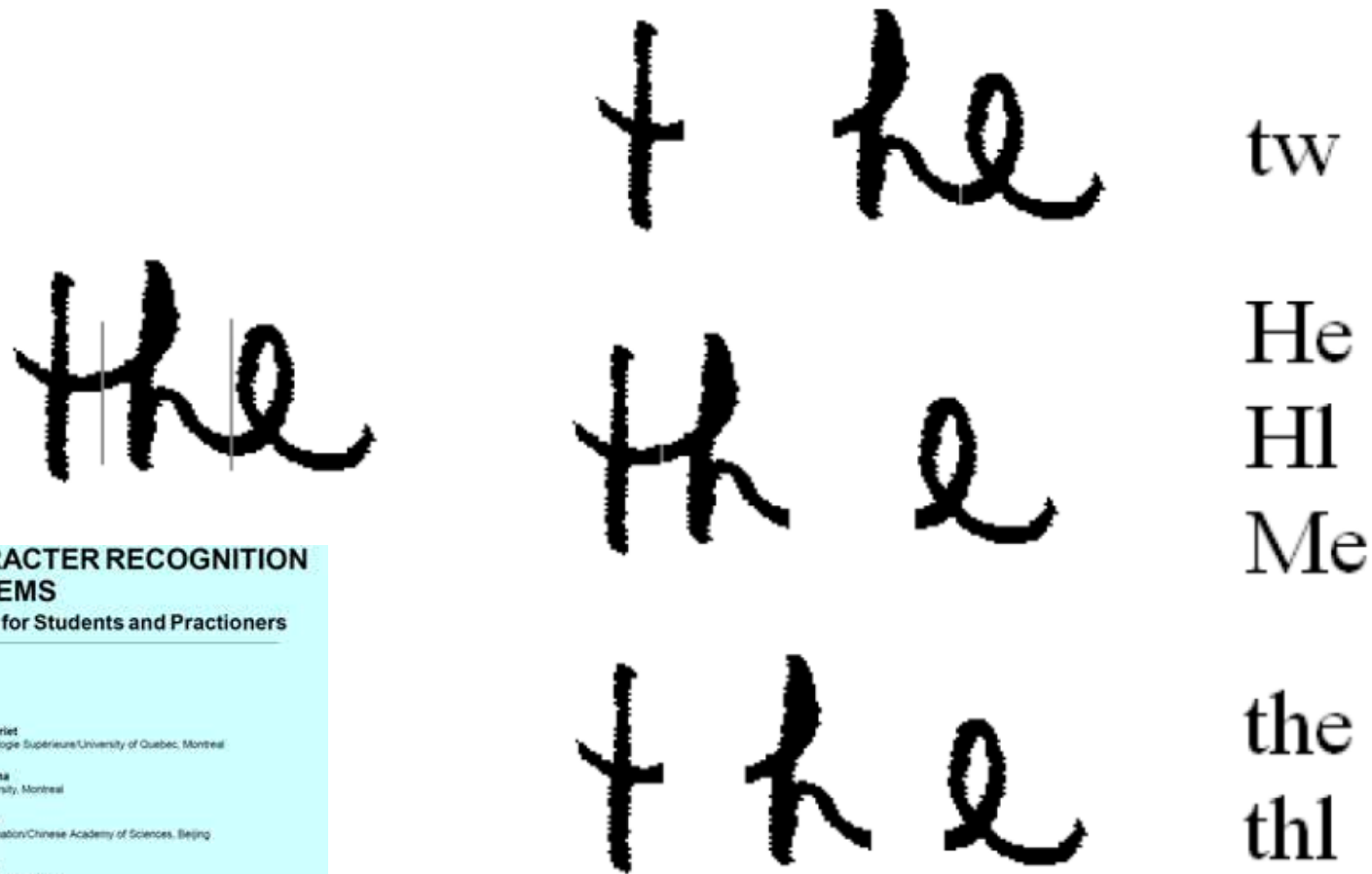
Cheng-Lin Liu

Institute of Automation/Chinese Academy of Sciences, Beijing

Ching Y. Suen

Concordia University, Montreal

Plausible segmentations and labels



CHARACTER RECOGNITION SYSTEMS

A Guide for Students and Practioners

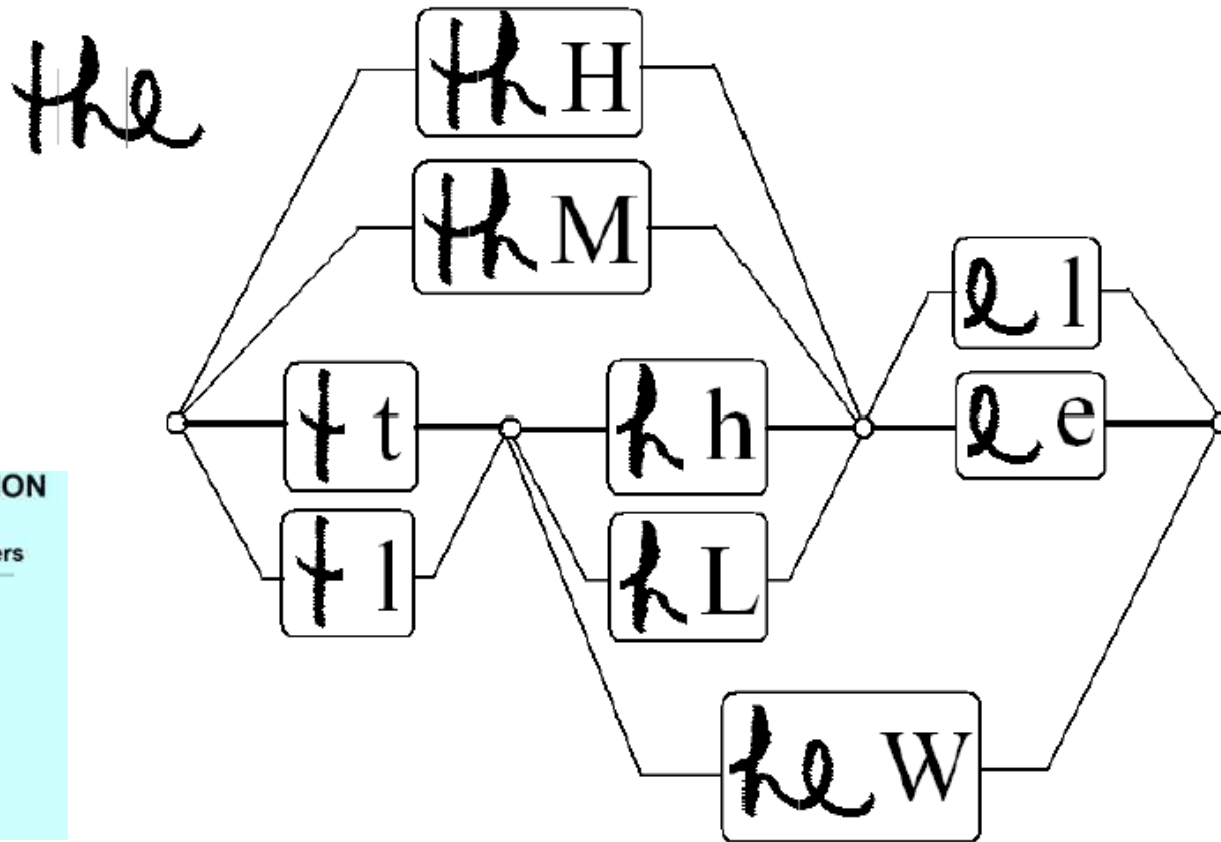
Mohamed Cheriet
École de Technologie Supérieure/University of Quebec, Montreal

Nawwat Khanna
Concordia University, Montreal

Cheng-Lin Liu
Institute of Automation/Chinese Academy of Sciences, Beijing

Ching Y. Suen
Concordia University, Montreal

SEGMENTATION AND LABEL LATTICE



CHARACTER RECOGNITION SYSTEMS

A Guide for Students and Practitioners

Mohamed Cheriet
École de Technologie Supérieure/University of Quebec, Montreal

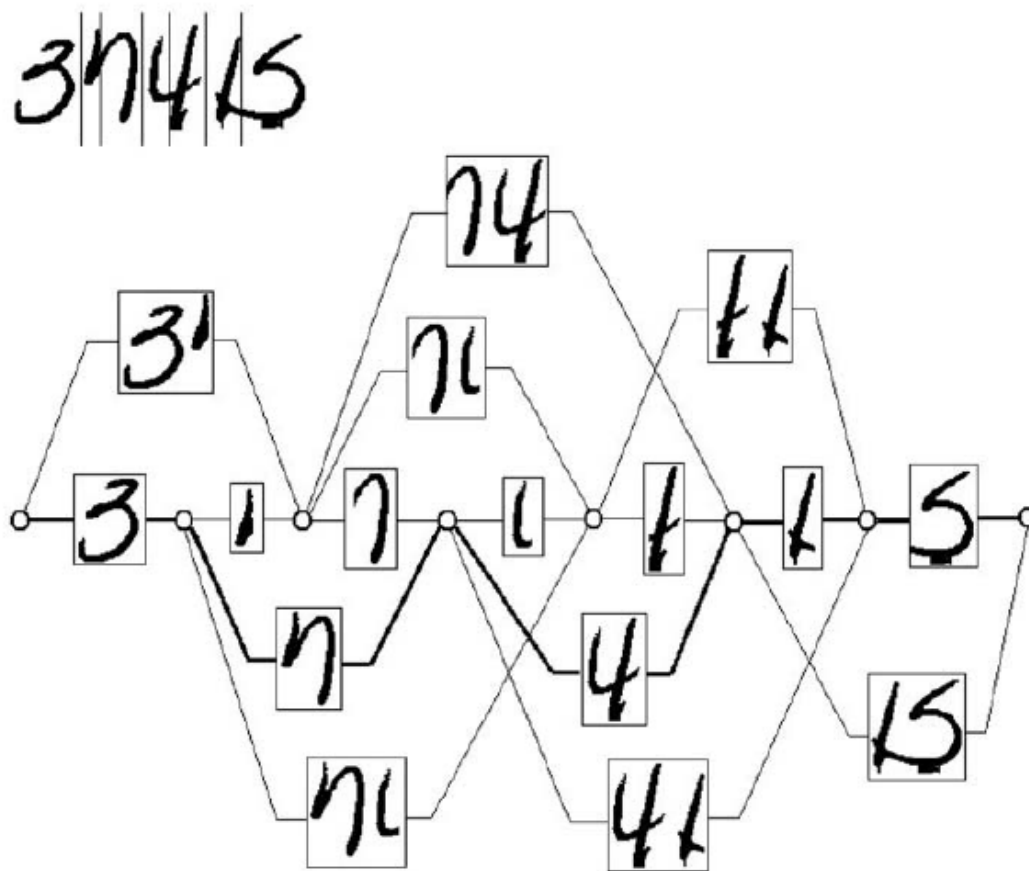
Nawaf Khanna
Concordia University, Montreal

Cheng-Lin Liu
Institute of Automation/Chinese Academy of Sciences, Beijing

Ching Y. Suen
Concordia University, Montreal

Over - segmentation

Segmentation Lattice



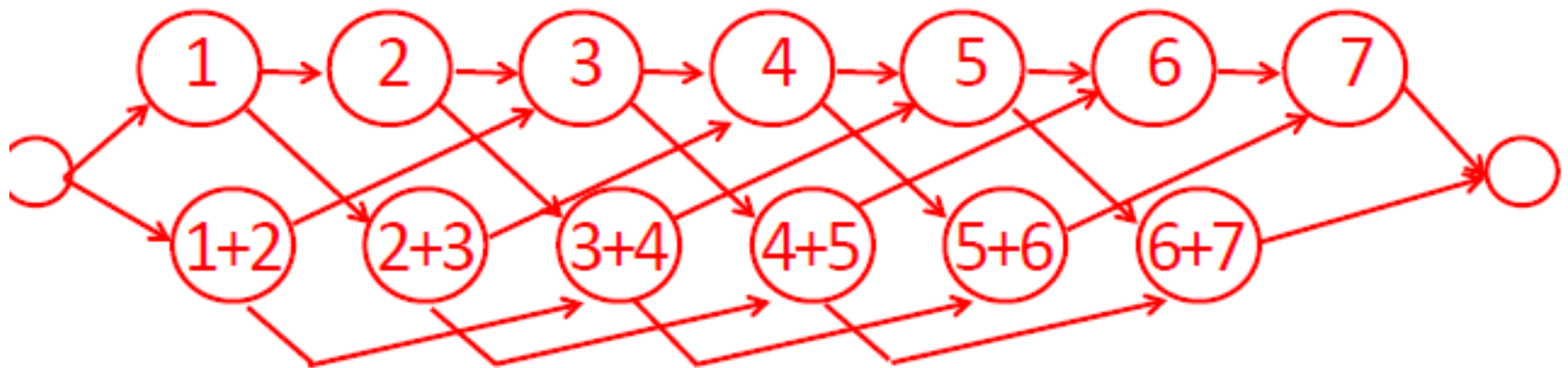
LIU ET AL.: EFFECTS OF CLASSIFIER STRUCTURES AND TRAINING REGIMES ON INTEGRATED SEGMENTATION AND RECOGNITION OF... PAMI November 2004

Another segmentation candidate lattice

rnhnm

Over-segmented:

1 2 3 4 5 6 7
rnhnm

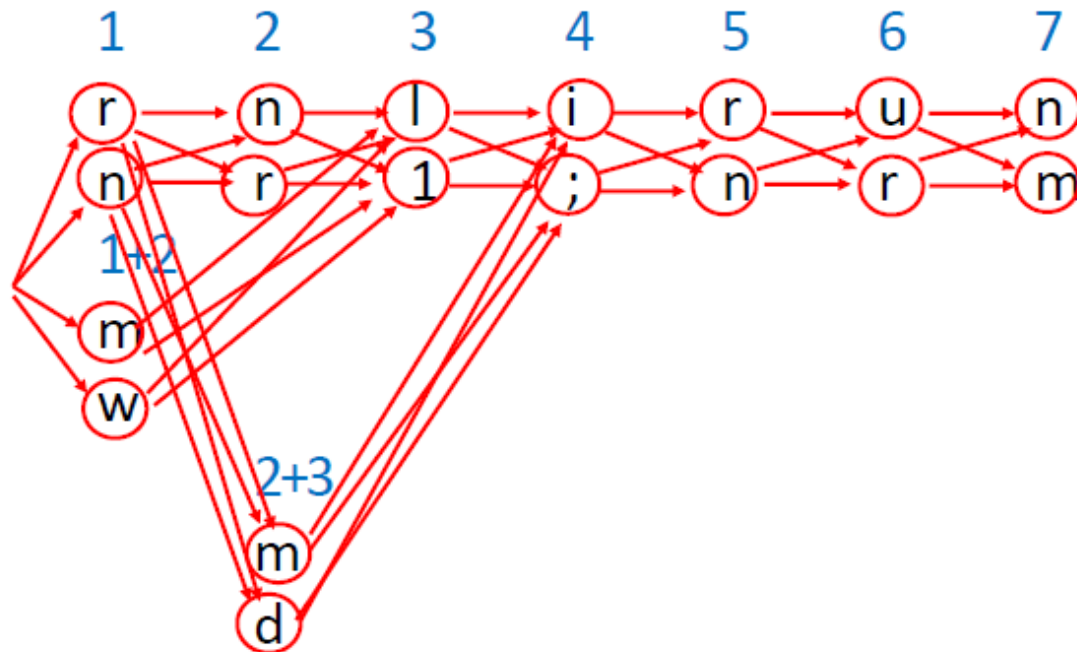


Partial Segmentation and Label Lattice

Only top-2 label candidates per segment

rnhnm

Segmented: rnhnm



Strategy

- **DP or Beam search** examines most promising paths
- Optimal path (both segmentation and labels) found by maximizing joint **posterior probabilities**
- Calculation takes into account *context*
 - language model:** letter or word n-grams or lexicon
 - style constraints:** quadratic discriminant field classifier, discrete style classifier, confidence transformation
- May also include a geometry model (g vs. 9; s vs. S)

A page of handwritten Chinese text

Q-F Wang, F. Yin, C-L Liu,, Handwritten Chinese Character Recognition
by Integrating Multiple Contexts, to appear in PAMI 2012

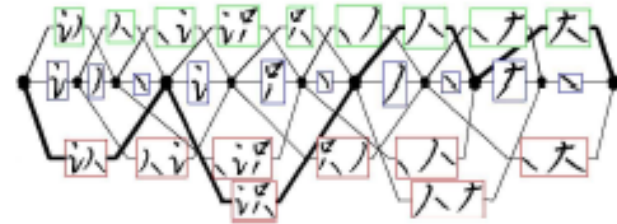
将进酒 唐·李白

君不见黄河之水天上来，奔流到海不复回。
君不见高堂明镜悲白发，朝如青丝暮成雪。
人生得意须尽欢，莫使金樽空对月。天生
我材必有用，千金散尽还复来。烹羊宰牛且
为乐，会须一饮三百杯。岑夫子，丹丘生，
将进酒，杯莫停，与君歌一曲，请君为我
倾耳听。钟鼓馔玉不足贵，但愿长醉不复
醒。古来圣贤皆寂寞，惟有饮者留其名。
陈王昔时宴平乐，斗酒十千姿欢谑。主人
何为言少钱，径须沽取对君酌。五花马，千
金裘，呼儿将出换美酒，与尔同销万古愁。

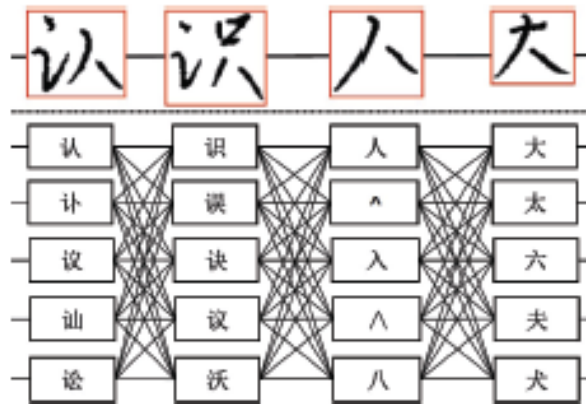
- (a) Over-segmentation to a sequence of primitive segments
- (b) Segmentation candidate lattice of part of (a)
- (c) Character candidate lattice of optimal segmentation path in (b)
- (d) Word candidate lattice of (c)



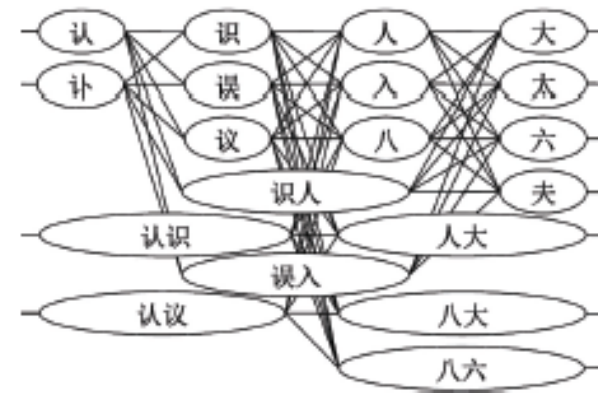
(a)



(b)



(c)



(d)

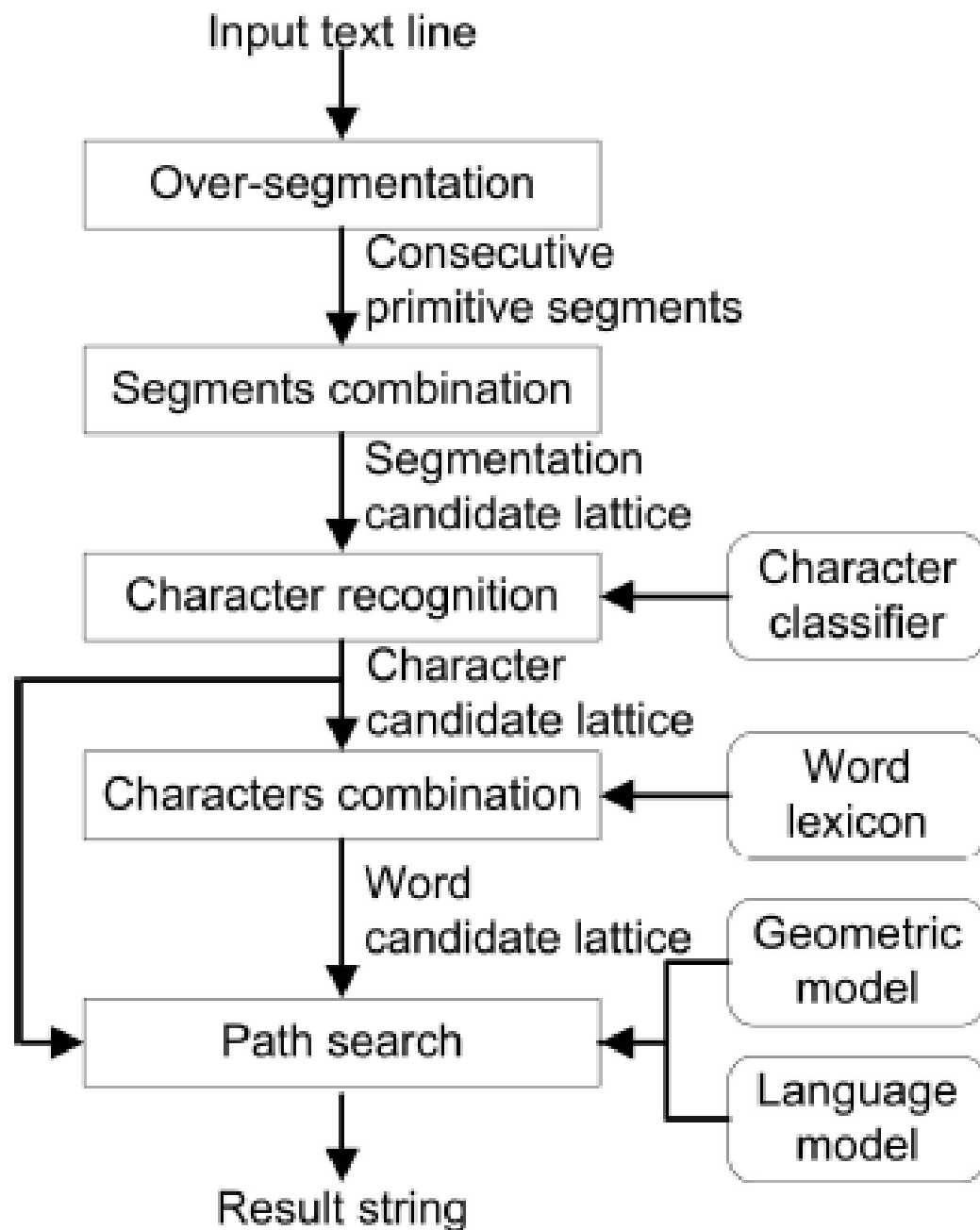
Q-F Wang, F. Yin, C-L Liu,, Handwritten Chinese Character Recognition by Integrating Multiple Contexts, to appear, PAMI 2012

Complete OCR (ICR?) system

9% error on
207,000 character
text by 204 writers.

**5% due to mis-
segmentation**

Q-F Wang et al., ibid.



Although integrated segmentation/recognition has made immense progress since its inception in the mid-seventies, segmentation errors still dominate in print and handwriting recognition. Further research is necessary for full integration of classification with style and language contexts.

3. GREEN INTERACTION

The economics of OCR have changed*

OCR Equipment:

Scanners and cameras

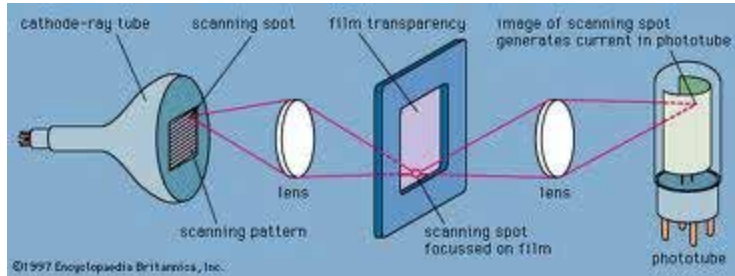
Storage

Networks

Processors

Printers and displays

*But OCR research has not



\$300,000 0.2 ppm



RAMAC: 5MB

3/2/2012

<-- Future

Scanners



\$40-\$120K 60 ppm
Sheet-feed scanners 150 ppm



550 ppm

Memory

USB: 256 GB



65



Modem

Hayes: 300 baud, \$300,
V.32: 14,400 bps \$800 in 1991



10Mbps



1982 10 MHz

CPU

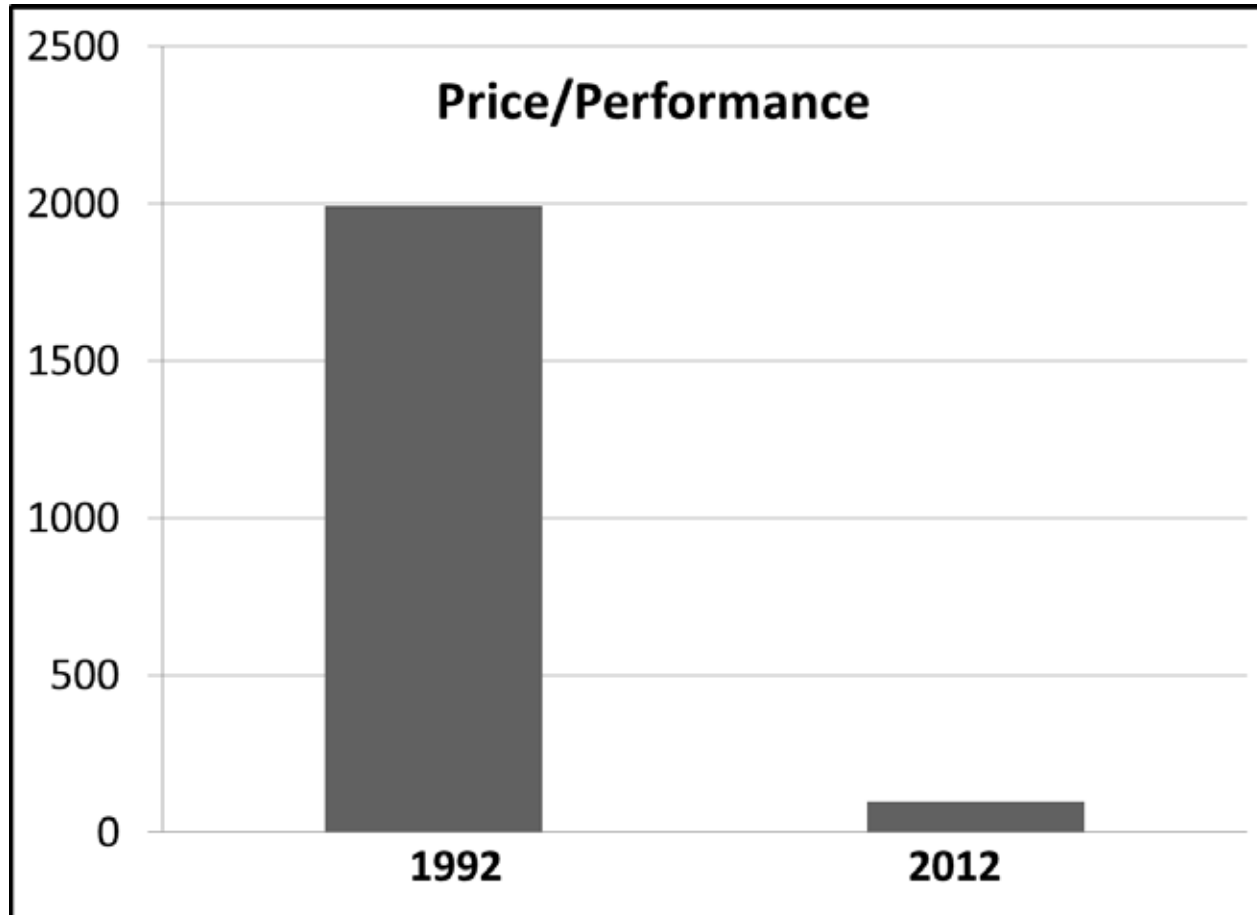


1970

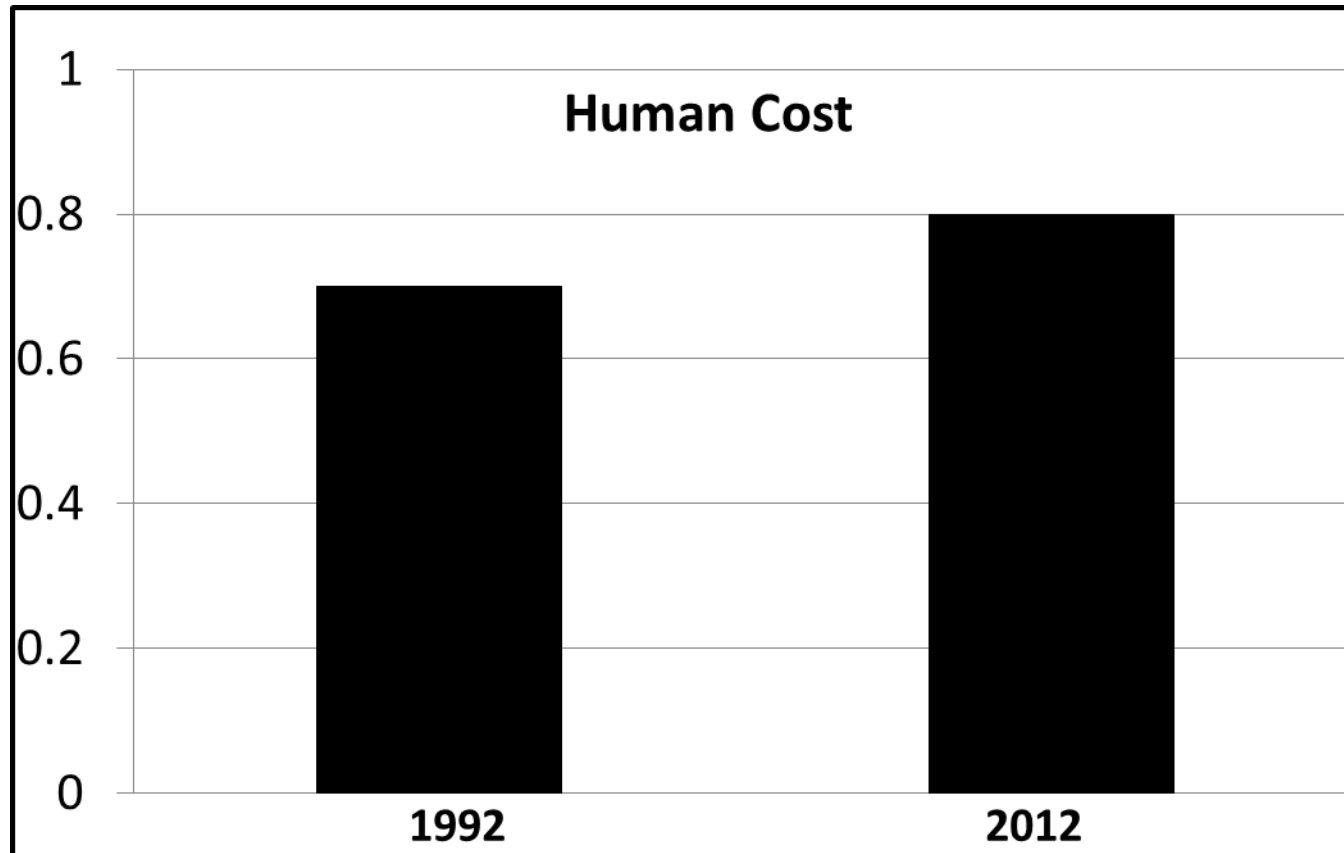
Laser
printer



Equipment Costs for OCR



Labor Costs for OCR



Don't waste human interaction: recycle it!

Advertised digitization and OCR prices*

Digitization	1-3 ¢/page
Digitization + OCR	4-10 ¢/page
Bound volume digitization + OCR	35-70 ¢/page
Bound volume digitization + OCR + correction	220 ¢/page
Editorial services	\$2-\$5/page
Mass digitization (mostly destructive)	\$6-10 /book

* May-June 2011

A confirmed cost figure (Sept. 2011)

Abbyy had 19 bankers' boxes (~ 100,000 pages) of my own OCR/DIA memorabilia digitized at a service bureau.

The searchable PDF files will be made publicly available on request as soon as Abbyy finishes indexing them.
(Copyright restrictions prevent posting it on the web.)

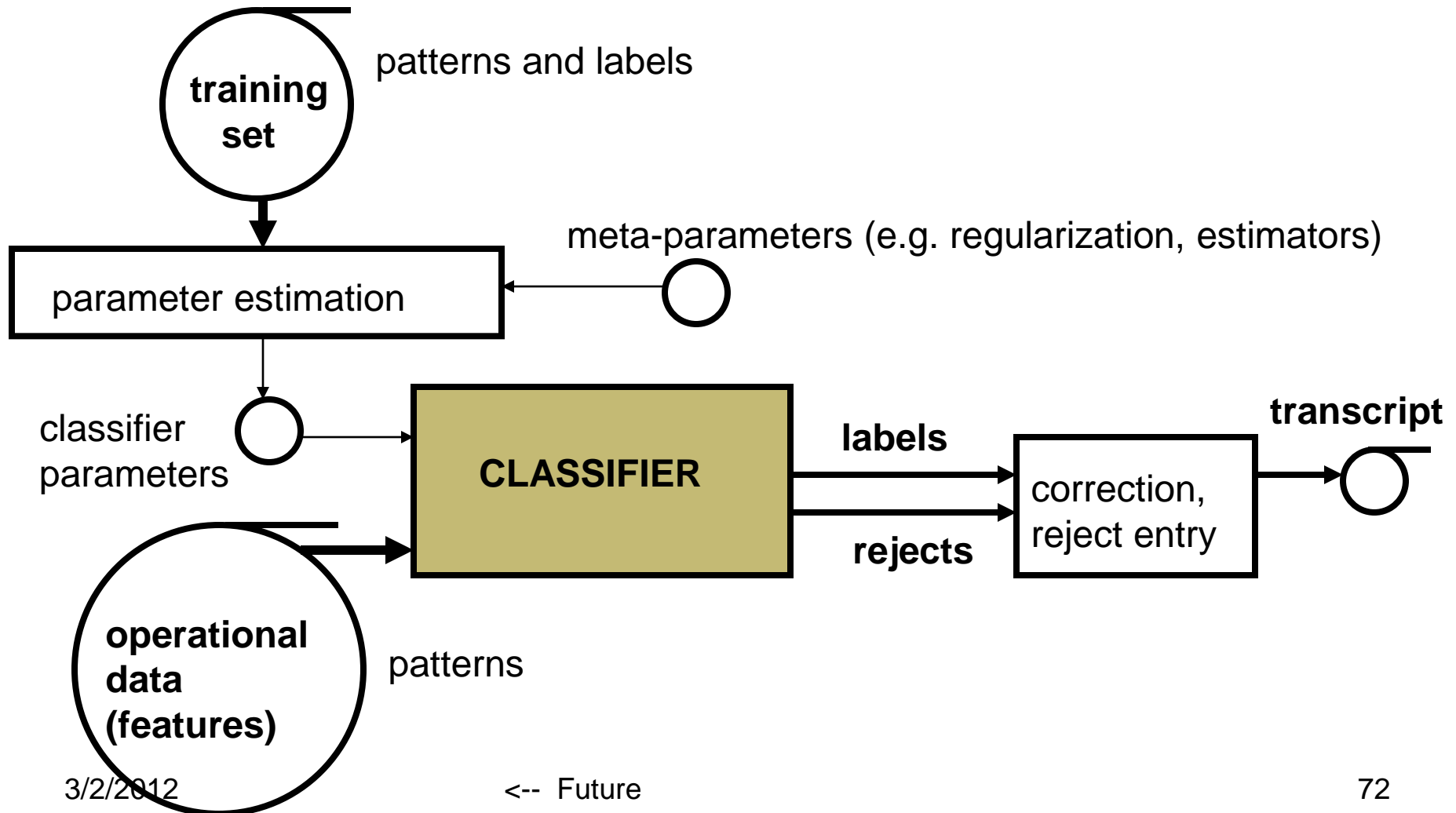
The average cost was **18 ¢/page**
including both bound and unbound material.

Observations

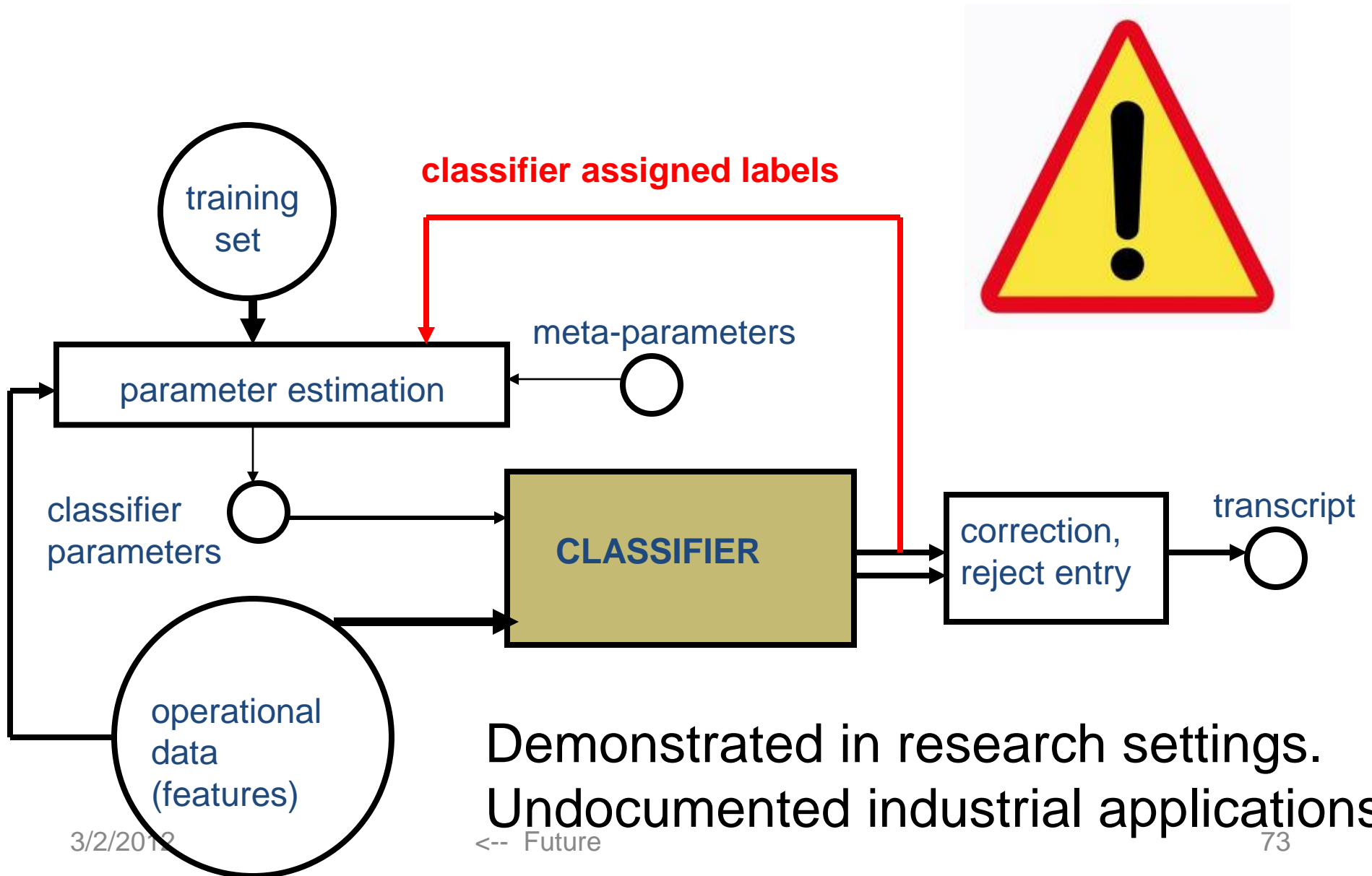
- There is always more within-application similarity than between-applications similarity.
- Classifiers best when trained on *representative* samples.
- The amount of data processed operationally is much *larger* than the training sets used for design.
It is also more *representative*.
- The best possible way to train a classifier is via the processed and corrected document stream.
- Retraining should be continuous and transparent to the operators.



Traditional open-loop OCR System

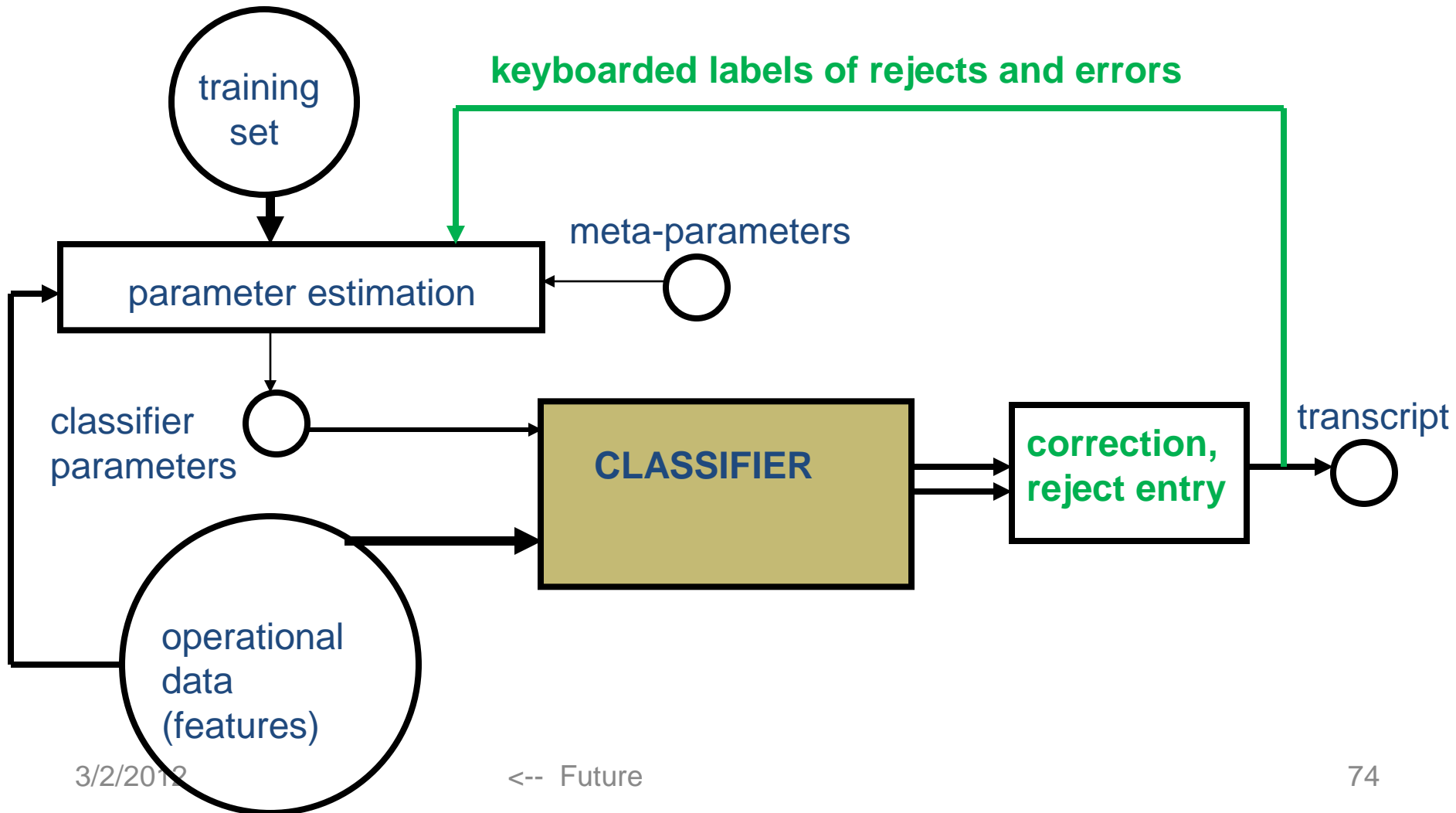


Adaptation (Jain, PAMI 00: “Decision directed classifier”)



Green Interaction

Generic OCR System that makes use of
post-processed rejects and errors



Green interaction will be particularly effective for multiple isogenous (same-source) page streams:

Operations with a limited number of document sources but unpredictable order of arrival

E.g.: Books from a finite number of publishers;
Birth/Marriage/Death certificates from the same jurisdictions (and especially from the same clerks);
Census forms by the same enumerators

Delays between the OCR stage and post-processing could cause problems. However, OCR and post-processing could be automatically interleaved.

Combine human & computer capabilities

HUMAN

Figure-ground separation
Part-whole relationships
Saliency
Extrapolate from few training samples
Exploit broad context
Gauge *relative* size and intensity
Detect *significant* differences
Colored noise; Texture
Non-linear feature dependence
Global optima in low dimensions

MACHINE

Measure
Count, sort and search
Store and recall many reference patterns
Estimate statistical parameters
Apply Markovian properties
Estimate decision functions from samples
Evaluate a complex sets of rules
Compute geometric moments
Orthogonal spatial transforms (e.g. wavelets)
Connected components analysis
Rank-order items according to a criterion
Filter out additive, white noise
Find local extrema in high dimensions

Agenda for OCR:

Automate feature design by hook or crook,
integrate segmentation and classification,
exploit downstream corrections.

Let the machine use all possible context,
and never-ever let it sleep.

Above all, avoid the curse of optimality!

Thank you!



<http://www.ecse.rpi.edu/~nagy/>

Paths through Segmentation Candidate Lattice (max two merged)

rnhnm

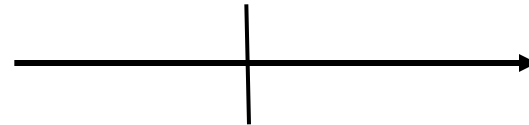
Segmented:

1 2 3 4 5 6 7
r n h n m

1	2	3	4	5	6	7
1+2		3	4	5	6	7
1	2+3		4	5	6	7
1	2	3+4		5	6	7
1+2		3+4		5	6	7

Feature dimensionality

- Number of features: $D = 30-300$
- Feature values: $0-1, 0-9, 0-2^{32}$
- Number of points: $2^D, 10^D, 2^{32D}$
- One feature \rightarrow 1-D
 - Feature or *weak classifier*?



Example (moments)

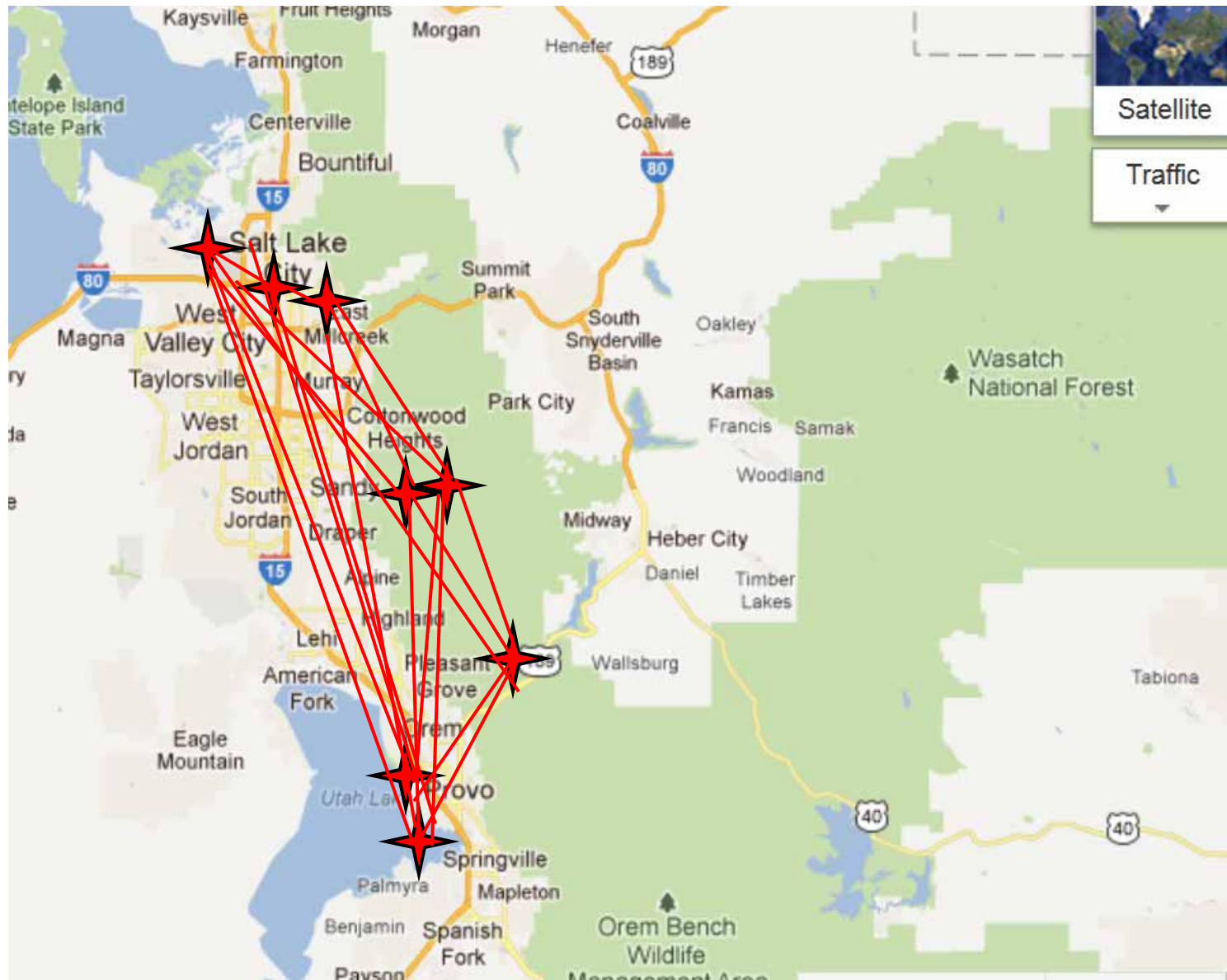
X	X	X	X	X
	X	X		
		X		
		X	X	
X	X	X	X	X

=

1	1	1	1	1
0	1	1	0	0
0	0	1	0	0
0	0	1	1	0
1	1	1	1	1
0	0	0	0	0

$$M^{00}=15, \quad m^{01}=3, \quad m^{10}=4; \quad m^{02}=1.47, \quad m^{20}=2.93, \quad m^{11}=-0.13$$

N = 8



Thank you!

<http://www.ecse.rpi.edu/~nagy/>