A PRESCIENT ABSTRACT:

In the past 30 years, various ideas have been presented for machines which could recognize spatial patterns. With the advent of the digital computer and its use in data processing, there has been a great increase of interest in the automatic conversion of human language to language understandable by a machine. The translation to machine language of spatial symbols---pattern recognition---is important to this conversion.

Papers presented at the the March 3-5, 1959, Western Joint Computer Conference
San Francisco, California Pages: 291-294
W. H. Highleyman *(RPI 1955*), Bell Telephone Labs., Inc., Murray Hill, N.J.
L. A. Kamentsky Bell Telephone Labs., Inc., Murray Hill, N.J.

Document Systems Analysis: Testing, Testing, Testing...

A Short History of Document Test Data

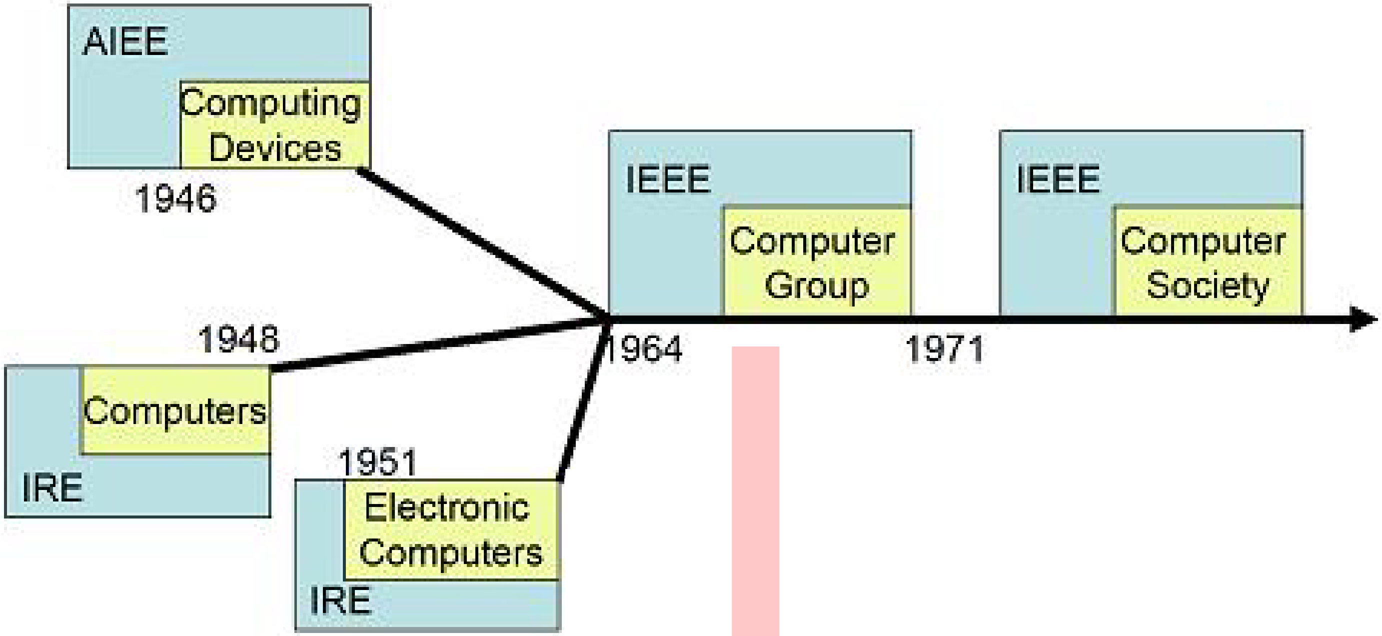
by

George Nagy

DocLab, RPI

PG-50

Organizations & Committees ~ 1967



ICPR	1973
IAPR	1978
PAMI	1979
ICDAR	1991
DAS	1994
IJDAR	1998
DocEng	2001

IEEE-CG TC on Pattern Recognition

Subcommittee on Reference Data Sets

IBM T.J. Watson Research Center 1965

Gardiner Tucker

Don Rosenheim

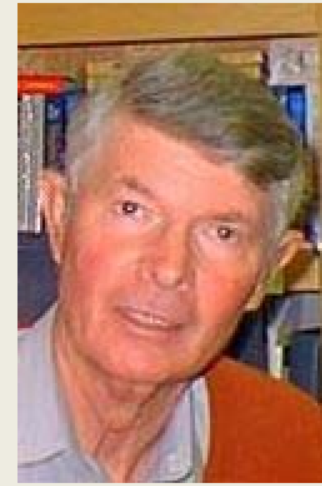
A. Hoagland

Don Streeter

Glenn Shelton, Jr.

George Nagy & Dick Casey

First Pattern Recognition Meeting Oct. 1966 - chaired by Al Hoagland



About a third of the presentations about das

Abend, Ball, Chow, Cover, Duda, Freeman, Groner, Hall, Julesz, Kanal, Kirch, Minsky, McCormick, Munson, Prewitt, Roberts, Papert, Rabinow, Roberts, Rosenfeld, Sammon, Specht, Stanat, Sheinberg, Sutherland, Specht, Uhr, Widrow, Zadeh... (52, 15 from U's)

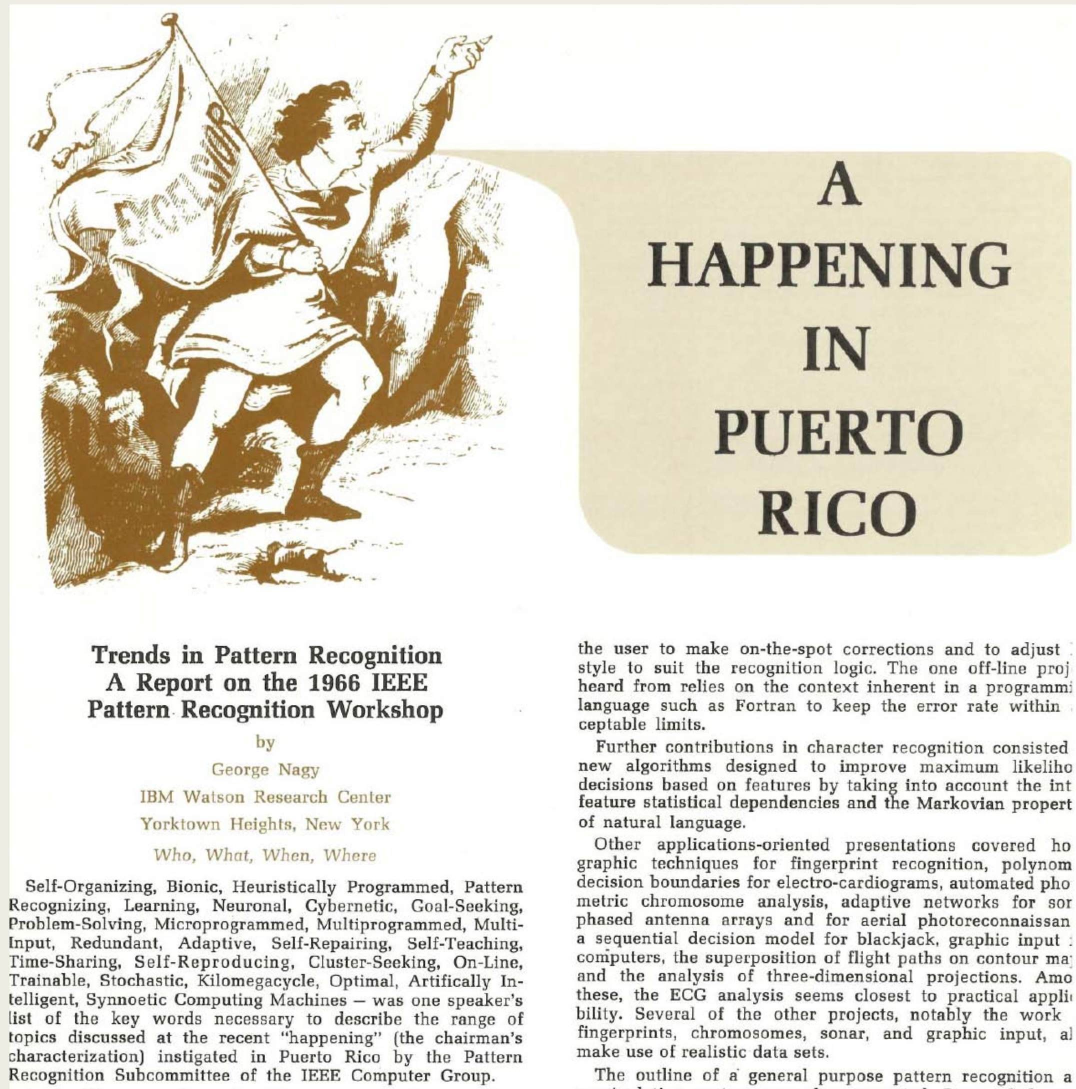
p 16

COMPUTER GROUP NEWS

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS

Vol. 1 No. 4

January, 1967



Trends in Pattern Recognition A Report on the 1966 IEEE Pattern Recognition Workshop

by

George Nagy

IBM Watson Research Center

Yorktown Heights, New York

Who, What, When, Where

Self-Organizing, Bionic, Heuristically Programmed, Pattern Recognizing, Learning, Neuronal, Cybernetic, Goal-Seeking, Problem-Solving, Microprogrammed, Multiprogrammed, Multi-Input, Redundant, Adaptive, Self-Repairing, Self-Teaching, Time-Sharing, Self-Reproducing, Cluster-Seeking, On-Line, Trainable, Stochastic, Kilomegacycle, Optimal, Artifically Intelligent, Synnoetic Computing Machines – was one speaker's list of the key words necessary to describe the range of topics discussed at the recent "happening" (the chairman's characterization) instigated in Puerto Rico by the Pattern Recognition Subcommittee of the IEEE Computer Group.

the user to make on-the-spot corrections and to adjust
style to suit the recognition logic. The one off-line proj
heard from relies on the context inherent in a programmi
language such as Fortran to keep the error rate within
ceptable limits.

Further contributions in character recognition consisted
new algorithms designed to improve maximum likeliho
decisions based on features by taking into account the int
feature statistical dependencies and the Markovian propert
of natural language.

Other applications-oriented presentations covered ho
graphic techniques for fingerprint recognition, polynom
decision boundaries for electro-cardiograms, automated pho
metric chromosome analysis, adaptive networks for sor
phased antenna arrays and for aerial photoreconnaissan
a sequential decision model for blackjack, graphic input
computers, the superposition of flight paths on contour ma
and the analysis of three-dimensional projections. Amo
these, the ECG analysis seems closest to practical appli
bility. Several of the other projects, notably the work
fingerprints, chromosomes, sonar, and graphic input, al
make use of realistic data sets.

The outline of a general purpose pattern recognition a

IEEE Computer Group News, pp. 16-19, January 1967

reprinted in *Spectrum*

June 9, 2010

Excerpts from DocLab Archives (1967-1974)

Minutes of the

IEEE Computer Group

Technical Committee

on Pattern Recognition (*PRC*) and its

Subcommittee on Reference Data Sets and Performance Evaluation Standards (*SCDS*)

PRC April 10, 1967 NYC

Al Hoagland, A. Hamburgen, L. Kanal, A. Rosenfeld, B. McCormick, B. Widrow, New: D. Brick, C.K. Chow, H. Freeman, J. Sammon, G. Nagy (+ visitors: H. Ostreicher, M. Watanabe)

- PR Subcommittee upgraded to full TC
- Puerto Rico Workshop proceedings:
  - **Wiley** would typeset and publish for $10 per copy
  - **Spartan** would print “as is” for $ 5-6 per copy
  - **McGraw Hill** not interested, **Addison Wesley** maybe
  - 8000 copies of OPTICAL CHARACTER RECOGNITION (Spartan 1960) sold
  - IEEE SMC PG-35 interested (also in taking over PG -16)
- *Pattern Recognition Society* established by R. Ledley

April 10, 1967, continued

- Participation at forthcoming meetings:
  - Washington, DC (UMD, ONR, USPS) June 1967- Rosenfeld
  - NCC Chicago September 1967 – C.K. Chow, J. Munson
  - Systems Science Hawaii January 1968 – Watanabe

    Specifically, Nagy's controversial short paper on Unsupervised Learning, presented at the Puerto Rico worship, should be expanded to a full paper (???)
  - IFIPS Edinburgh August 1968 – Herb Freeman

  - IEEE G-SSC TC on Pattern Recognition and Learning Systems
  - IEEE G-AC TC on Adaptive and Learning Systems

(MMS+SSC→SMC

After the meeting, Al Hoagland asked me to look into reference data.

MEMORANDUM

To: A. S. Hoagland, Chairman, IEEE Pat Rec Committee
From: G. Nagy
Date: August 28, 1967
Subject: **Reference Data Sets**

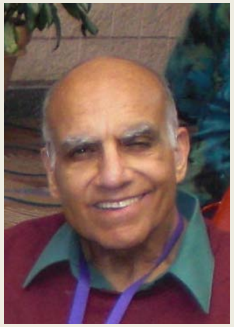
Contacted universities, OCR manufacturers, industrial and non-profit research labs, and the United States Government.

"The consensus of opinion appears to be that the establishment of reference material of this sort is long overdue."

Alternatives:

1. Establish a small subcommittee to publish a catalog of data sets which may be obtained from their respective owners;
2. Co-sponsor with a government agency (NIST, then NBS);
3. Endorse a non-profit research org or a consultant to make reference data available to the public for a fee;
4. **A means within IEEE to acquire, maintain and distribute the data, in a manner similar to the Computer Group Repository, which in fact already encourages the submission of original data.**

PRC October 13, 1967 Boston



- Procs. of Puerto Workshop mailed to Thompson (Kanal).
- Nagy's suggestion for a subcommittee on reference data sets adopted. Requests to add pictorial and time-varying signals, as well as multi-dimensional feature data.

  [**CK Chow,** Munson (SRI), Spooner (CAL), Samson (RDC), Gibbons (PO)].
- Duda designated Committee's Contributing Editor
  - (to *Computer Group News*).
- PRS is establishing a new journal: *Pattern Recognition*

  (editors from PRC invited).
- Explore cooperation in IP with Optical Society of America.
- Delft PR workshop in August 1968 (Verhagen).
- Pisa two-week summer school in September 1968 (Grasselli).
- San Francisco SSC Conf in October 1968.

SCDS March 21, 1968 NYC

- C.K. Chow received enthusiastic IEEE endorsement and offer to cooperate.
- Harry Huskey will accept data tapes as items of his (PGEC) repository: **IEEE Test Data Sets in *Computer Group News*, March 1968, p. 28.**
- Harburgen obtained legal advice: IEEE should copyright data, identifying and giving full credit to source. IEEE cannot restrict use of data to researchers.
- **MP:** Spooner has found several MP sources. **REI** has large data base, but reluctant to release it except to researchers. C.K. Chow said **IBM** may release 10,000 chars of MP. **CAL** has 40,000 alphanumeric chars in 24x24 binary form with 64 levels. Expect soon to have 150,000 chars including IBM Executive and Remington Rand Elite.
- **HP:** Highleyman; Knoll; Munson
- **Cursive:** Harmon, Murray Eden, Gibbons
- **Speech ; EKG, EEG, EMG; Seismic; Radar/Sonar; Fingerprints; X-Rays; Bubble Chamber, Microscope, Aerial, Celestral Photos; Maps; Line Dwgs (none!); Property lists (Medical diagnostics, Taxonomy);**

W.H. Highleyman

PRC **May 15, 1969** Boston

- 203 copies of **Pattern Recognition** sold!
- Should IEEE repository handle data set distribution (efficiency and cost)?
- No formal efforts will be made to create an umbrella organization to coordinate pattern recognition activities among IEEE committees and other groups (*the autonomy of the Computer Group with the IEEE is now under discussion.*)
- Delft Workshop report by Chandrasekaran, Kanal and Nagy published by the IEEE.
- Hoagland thanked and sent up, Chow welcomed as TC chair.
- Workshop in Honolulu proposed.

SCDS - May 15,1969

The problem of verifying the description of data on magnetic tape was seen to be a formidable one because of incompatibilities among different computer systems.

Hamburgen will check IEEE Computer Center facilities.

PRC **November 17, 1969** Las Vegas

- Four documented data sets submitted to IEEE HQ.
  Requires IEEE ADCOM and Tech Ad Board approval:
  Chow wrote to McCluskey
- Twelve facilitators plus two international coordinators appointed for data bases
- Subcommittee renamed: *Reference Data Bases and Evaluation Procedures*
- Hawaii workshop with SMC should have no session on OCR or statistical techniques, sessions should not be organized by subject, and most sessions should be devoted to future methods of pattern recognition. (Objection: this leaves out most engineers who work in pattern recognition)

PRC **November 18, 1970** Houston

- Spooner to chair SCRD
- Hamburgen will organize a subcommittee on character recognition
- Still 672 copies left of Pattern Recognition!
- Still looking for a Service Bureau to copy tapes
- Preparations for 1972 Workshop on PR in Hot Springs (Ed Parrish)

SCDS May 20, 1971 Atlantic City

1. Data Sets sent from IEEE HQ to Computer Society two months ago, but have not yet arrived
5. Should advertise in *IEEE Spectrum*
6. Publish names of data set users
7. Rabinow to be asked to obtain a set of alphanumeric characters that have been systematically degraded
9. **"It was first agreed that the first order of business was to get the data sets in hand and advertise in Computer so that we can point to a significant milestone in the Committee's accomplishments."**

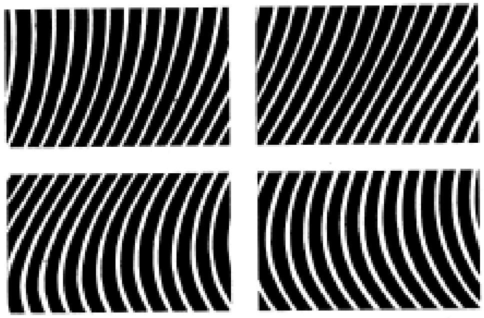
PRC **November 17, 1971** Las Vegas

- Five data sets are ready and will be advertised in the January-February 1972 Issue of *Computer* magazine.
- Production cost estimated at $100.
- PRC should consider expanding its scope to

  image processing and ***artificial intelligence***.
- Al Klinger requested to prepare a Glossary of Pattern Recognition.

Data Sets advertised in IEEE *Computer*

January 1972 (6 data sets)



PATTERN RECOGNITION
DATA BASES AVAILABLE

In order to encourage research in the field of pattern recognition, the IEEE Computer Society's Technical Committee on Pattern Recognition has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the bottom of the page.

Discounts off the data base list prices are available to IEEE members and members of the American Federation of Information Processing Societies' constitutient societies.

When ordering, you may elect to send us your own blank tapes; if you do, be sure they are in good condition and have no other data recorded on them.

1.1.1 Machine Imprinted Alphanumeric Characters – Dr. H. F. Ryan, Cornell Aeronautical Laboratory, Inc/U.S. Postal Service

An alphanumeric character data base of 100,000 samples of 66 character classes. (Also known as the CAL-U.S. Postal Service Alphanumeric Character Data Base.) Thresholded binary images of segmented, centered, mixed-font, machine-imprinted characters. Resolution is 24 x 24. Magnetic tape, 9 track, 2 reels, 1600 BPI.

Price: $112.50 ($68.75 with furnished tapes)
Member's discount price: $90, ($55, with furnished tapes)

1.2.1 Handprinted Numeric Characters – Dr. A. L. Knoll, Honeywell Information Systems, Data Systems Division

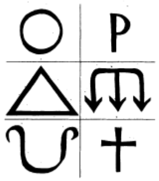
The data base consists of 50 samples of each numeric character generated by 9 different authors. Simple printing rules were specified but not always followed. The samples were selected from those contributed. The images are binary with a resolution of 25 x 21. Punched cards.

Price: $37.50
Member's discount price: $30.

4/COMPUTER/JANUARY/FEBRUARY 1972

January 1973

PATTERN RECOGNITION
DATA BASES AVAILABLE



In order to encourage research in the field of pattern recognition, the IEEE Computer Society's Technical Committee on Pattern Recognition has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the bottom of the page.

Discounts off the data base list prices are available to IEEE members and members of the American

but not always followed. The samples were selected from those contributed. The images are binary with a resolution of 25 x 21. Punched cards.

Price: $41.25
Member's discount price: $33.

1.2.2 Handprinted FORTRAN Alphanumeric Characters – Dr. John H. Munson, Stanford Research Institute
The data base consists of two parts, with each part on a reel. The first part contains 3 alphabets of 46 characters, corresponding to the non-blank character set of the basic FORTRAN language, hand-printed by each of 49 authors making a total of 3 x 46 x 49 = 6,762 patterns.
The second part has 2,999 characters printed by a single author. There are 920 characters made up of 20 alphabets of 46 characters each; the remaining 2,079 characters are taken from fragments of actual coding sheets. The images are binary with a 24 x 24 resolution. Magnetic tape, 7 track, 2 reels, 556 BPI.

alphanumeric characters. The images are binary with a resolution of 12 x 12. Punched cards.

Price: $55.
Member's discount price: $44.

1.2.4 Handprinted Numeric Characters – Hiroshi Genchi, Tokyo Shibaura Electric Co., Ltd/Toshiba Research and Development Center
The data base consists of 10,000 hand written numeric characters collected from live as well as experimental mail throughout Japan. The data base is contained on two magnetic tapes with each tape having 5,000 characters. Images are binary with a resolution of 36 x 50. A single pattern consists of 56 words. Magnetic tape, 7 track, 2 reels, 556 BPI, 6 bits per character.

Price: $115.25 ($68.75 with furnished tapes)
Member's discount price: $93. ($55. with furnished tapes)

1.3.1 Cursive Script – Dr. L. D. Harmon, Bell Telephone Laboratories

April 1976 (10 data sets)

Pattern Recognition
DATA BASES

1.1.1 Machine Imprinted Alphanumeric Characters – Dr. H. F. Ryan, Calspan Corp./U.S. Postal Service

An alphanumeric character data base (normalized version) of 100,000 samples of 66 character classes. (Also known as the CAL-U.S. Postal Service Alphanumeric Character Data Base.) Thresholded binary images of segmented centered mixed-font, machine-imprinted characters. Resolution is 24 x 24. Magnetic tape, 9 track, 2 reels, 1600 BPI.

Price: $123.75 ($75.50 with furnished tapes)
Member's discount price: $99 ($60 with furnished tapes)

1.1.1A Machine Imprinted Alphanumeric Characters – Dr. H. F. Ryan, Calspan Corp./

In order to encourage research in the field of pattern recognition, the IEEE Computer Society's Technical Committee on Machine Pattern Analysis has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the back of the issue.

Discounts off the data base list prices are available to IEEE members and members of the American Federation of Information Processing Societies' constituent societies.

When ordering, you may elect to send us your own blank tapes; if you do, be sure they are in good condition and have no other data recorded on them.

If you have a data base that you wish to

To order: Use the multipurpose order form at the back of the issue.

magnetic tapes with each tape having 5,000 characters. Images are binary with a resolution of 36 x 50. A single pattern consists of

Announcement

In order to encourage the field of pattern recognition, the IEEE Computer Society's Technical Committee on Machine Pattern Analysis has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the back of the issue.

Discounts off the data base list prices are available to IEEE members and members of the American Federation of Information Processing Societies' constituent societies

When ordering, you may elect to send us your own blank tapes; if you do, be sure they are in good condition and have no other data recorded on them.

If you have a data base that you wish to contribute to the Technical Committee on Machine Pattern Analysis, please contact Dr. J. B. McFerran, Sperry-Univac, P. . Box 3525, St. Paul, Minnesota 55165.

PRC **June 17, 1972** Atlantic City

- 20 Data Sets purchased!
- Ed Parrish to chair SCDS
- 41 @ Hot Spring PR Workshop. Income – Expenses = $705.92.
- S. Yau: Expand scope of PRC and change name, or form another committee :

Pattern Recognition	*Complex Information Processing*
Pattern Recognition and Machine Intelligence	*Heuristic Problems*
Pattern Recognition and Artificial Intelligence	*Cognitive Technology*
Models of Cognition (or Intelligence)	*Artificial Intelligence*

Machine Pattern Analysis `

SCDS

Suggestion: find a "willing expert " to verify data sets.
McFerran explained Univac's computer-generated character data bases.

PRC **Jan 4, 1973** Disneyland

> 100 attendees at 2-DSP Conf at U. Missouri in August 1972

37 Data Sets purchased!

Machine imprinted alphanumeric characters 1.1.1	7
Handprinted numeric characters 1.2.1	7
Handprinted FORTRAN alphanumeric characters 1.2.2	10
Handprinted alphanumeric characters 1.2.3	7
Handprinted numeric characters 1.2.4	3
Cursive script 1.3.1	3
	37

PRC **June 7, 1973** NYC

60 Data Sets purchased!

Will be advertised regularly in *Computer*

Should software (programs) be added?

PRC **February 26, 1974** San Francisco

Lengthy discussion of acceptability of artificially generated data.

Collaborate with ACM SIGART?

PRC **May 7, 1974** Palmer House

Task force on publishing mechanisms in PR-oriented journals

PRC **Sept 11, 1974** Mayflower

Copenhagen	IJCPR June 1974 : 400 attendees
Silver Spring	IPR Workshop November 1974
Asilomar	PR Workshop March 1975
El Coronado San Diego	ICPR Fall 1976

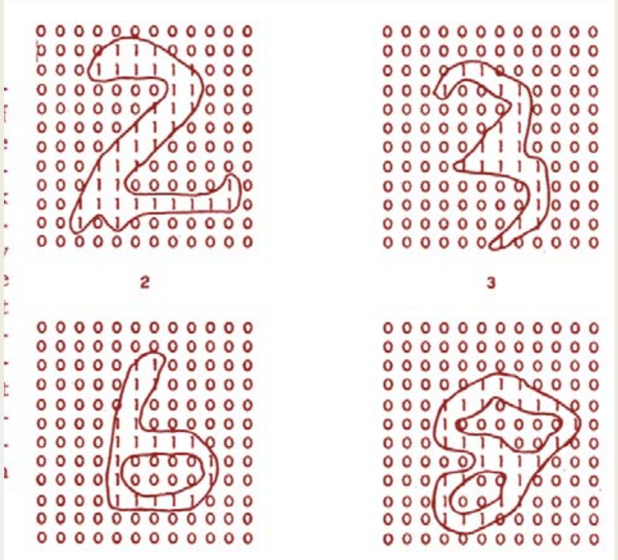
First Machine Pattern Analysis TC Newsletter *Octob*er 1974

1974 PRC

Agrawala, Brick, Butler, Butterfield, Chien, Chow, Deutsch, Fischler, Frank, Freeman, Fukunaga, Gibbons, Hamburgen, Harlow, Hoagland, Kanal, Klinger, Lainiotis, Lambert, Lederer, McFerran, Mathur, Meisel, Nadler, Nagy, Parrish, Patrick, Robinson, Rosenfeld, Samit, Sammon, Shapiro, Sklansky, Spooner, Stoffel, Swonger, Watanabe, Weinstein, Wilson, Widrow, Yau

Most often used dataset ever

(50 alphabets)



1.2.3 Handprinted Alphanumeric Characters – Dr. W. H. Highleyman, Sombers Associates.

There are approximately 2300 samples of alphanumeric characters. The images are binary with a resolution of 12 x 12. Punched cards.

Price: $55
Member's discount price: $44

Data for Character Recognition Studies*

A few years ago during my initial work in the problem of pattern and character recognition, I reduced several samples of hand printing and machine printing to matrix form. I later used this data to obtain experimental results which appeared in two papers.[1,2] My intent in reducing this data

* Received February 7, 1963.

[1] W. H. Highleyman, "An analog method for character recognition," IRE Trans. on Electronic Computers, vol. EC-10, pp. 502–512; September, 1961.

[2] W. H. Highleyman, "Linear decision functions, with application to pattern recognition," Proc. IRE, vol. 50, pp. 1501–1514; June, 1962.

over 50 requests by August 1967

IEEE Computer, April 1976, p. 83

SRI

1.2.2 Handprinted FORTRAN Alphanumeric Characters

Dr. John H. Munson, Stanford Research Institute

The data base consists of two parts, with each part on a reel. The first part contains 3 alphabets of 46 characters, corresponding to the non-blank character set of the basic FORTRAN language, hand-printed by each of 49 authors making a total of 3 x 46 x 49 = **6,762 patterns**.

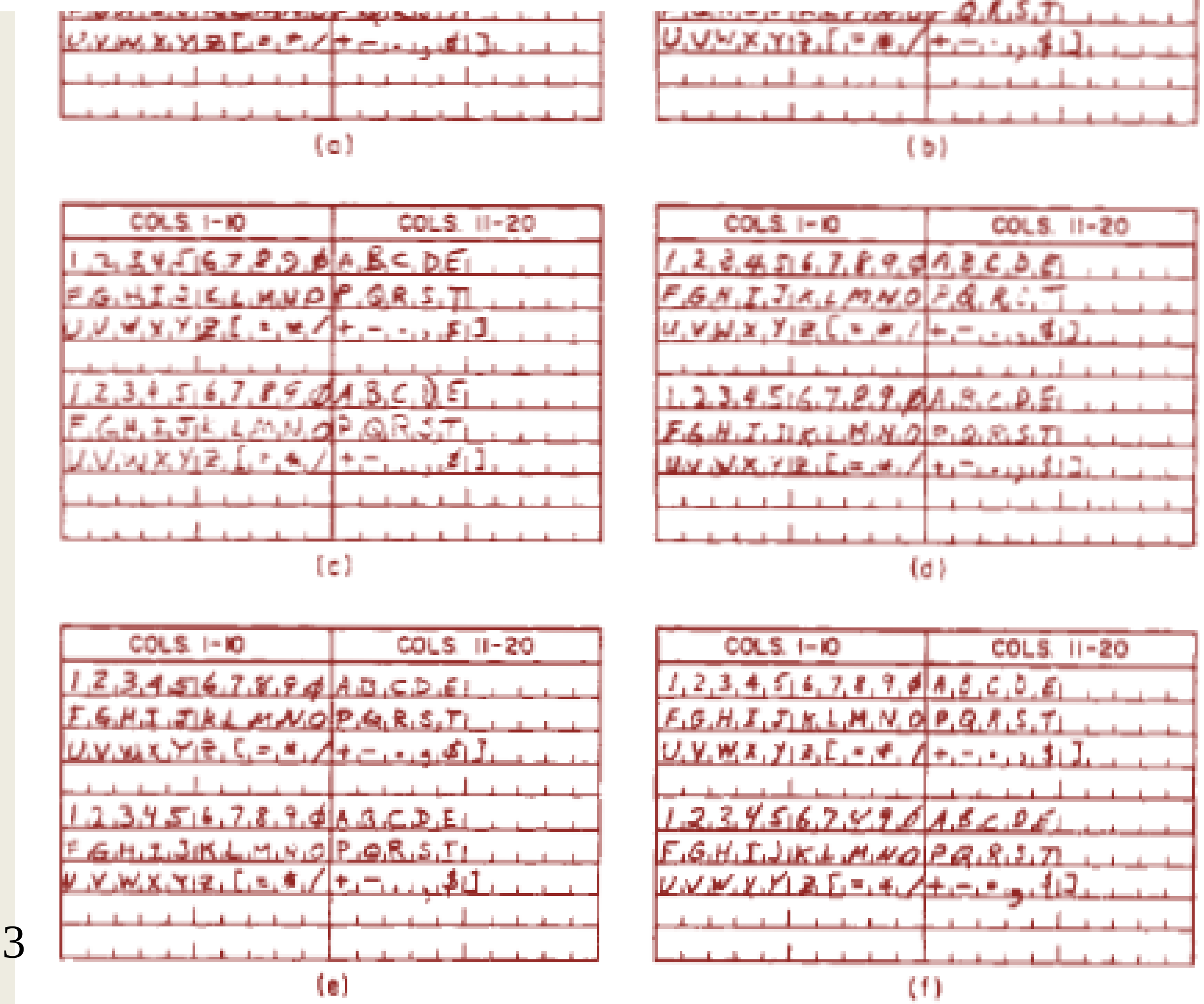
The second part has 2,999 characters printed by a single author. There are 920 characters made up of 20 alphabets of 46 characters each; the remaining 2,079 characters are taken from fragments of **actual coding sheets**. The images are binary with a 24 x 24 resolution. Magnetic tape, 7 track, 2 reels, 556 BPI.

Price: $116.75 ($75.75 with furnished tapes)

Member's discount price: $93 ($60 with furnished tapes)

IEEE Computer, April **1976**, p. 83

Fig. 14. Handprinted characters on FORTRAN coding sheet. An augmented FORTRAN alphabet is shown by each of twelve different writers. The range of variation is considerable even though the writers were in no particular hurry. These data were collected at the Stanford Research Institute under sponsorship of the U. S. Army Electronics Command.



IEEE Computer, April **1976**, p. 83

Post Office

1.1.2 Machine Imprinted Multilevel Characters

- James S. Gibbons, Electronic Sciences Division, U.S. Postal Service.

An alphanumeric character data base of 32,000 (16 level) multi-level characters, extracted from 2100 test mail pieces each in a 24 x 24 array covering an area of 0.144 x 0.144 square inches. Complete information as to upper/lower case, font style, percent reflectance, color print and background accompanies each character on magnetic tape. Suitable for machine independent research as well as OCR testing. Magnetic tape, 4 reels test characters, 1 reel set up characters, 9 track, 800 BPI.

Price: $262.50 ($150.00 with furnished tapes)

Member's discount price: $210.00 ($120.00 with furnished tapes)

Honeywell

1 1.2.1 Handprinted Numeric Characters -

Dr. A. L. Knoll, Honeywell Information Systems, Data Systems Division

The data base consists of 50 samples of each numeric character generated by 9 different authors. Simple printing rules were specified but not always followed. The samples were selected from those contributed. The images are binary with a resolution of 25 x 21. Punched cards.

Price: $41.25

Member's discount price: $33

Toshiba

1.2.4 Handprinted Numeric Characters - Hiroshi Genchi, Tokyo Shibaura Electric Co., Ltd./Toshiba Research and Development Center
The data base consists of 10,000 hand written numeric characters collected from live as well as experimental mail throughout Japan. The data base is contained on two magnetic tapes with each tape having 5,000 characters. Images are binary with a resolution of 36 x 50. A single pattern consists of 56 words. Magnetic tape, 7 track, 2 reels, 800 BPI, 6 bits per character.

Price: $115.25 ($68.75 with furnished tapes)
Member's discount price: $93 ($55 with furnished tapes)

Bell Labs

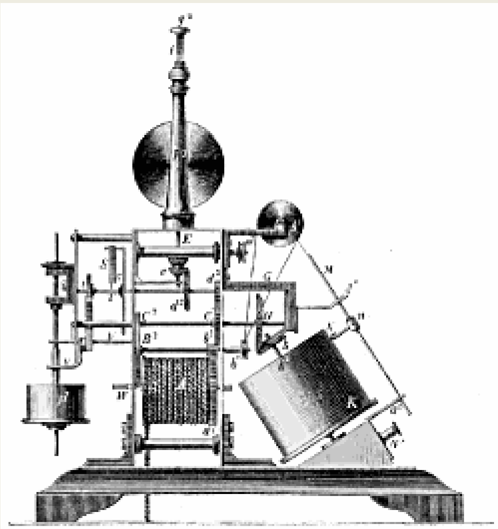
1.3.1 Cursive Script - Dr. L. D. Harmon, Bell Telephone Laboratories
The data consists of 52 cursive script sentences. The resolution for each sentence is 256 (vertically) x 2048 (horizontally). The images are binary. Magnetic tape, 7 track, 1 reel, 200 BPI.

Price: $60.50 ($42.25 with furnished tapes)
Member's discount price: $50 ($33 with furnished tape)

To order: Use the multipurpose order form at the back of the issue.

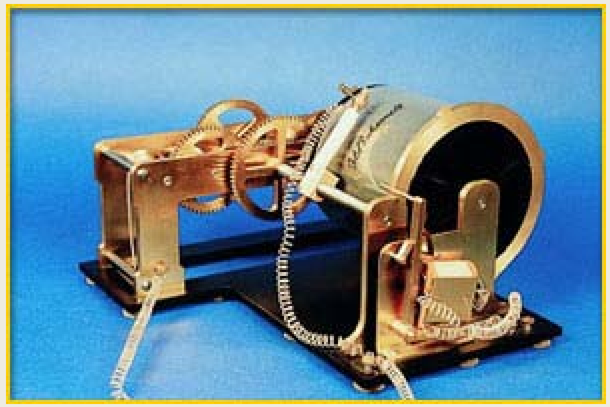
IEEE Computer, April **1976**, p. 83

OCR infrastructure: Scanners

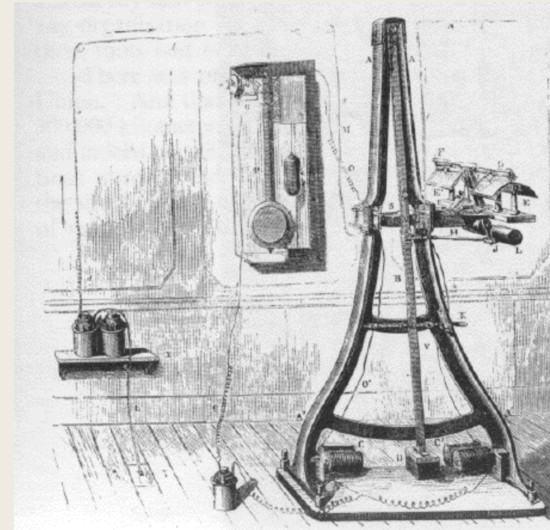


Bain 1841

Drum scanners
for telegraphy

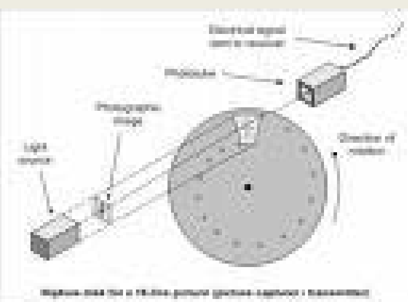


Bakewell 1847

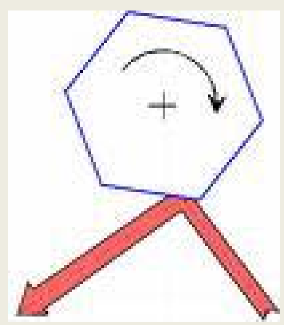


Caselli's facsmile telegraph transmitter: one of the first to send examples of hand writing and simple pictures over telelgraph wires. (From L. Figuier "Les Mervieilles de la Science," Paris, about 1866 and cited "From Semaphore to Satellites," ITU, 1965.)

Caselli 1861



Nipkow disk



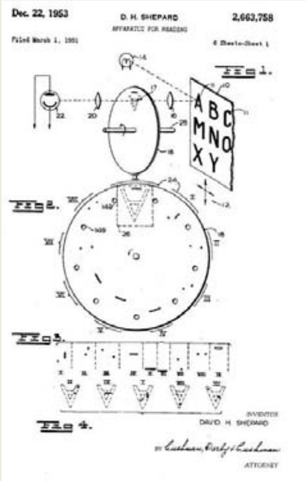
Rotating mirror

CRT

OCR GENEOLOGY <1960

GUSTAV TAUSEK PATENT 1929

(MECH. TEMPLATES + PHOTODETECTOR)



DAVID SHEPARD 1951

INTELLIGENT MACHINES RESEARCH

READERS DIGEST, STANDARD OIL 1955

IMR → FARRINGTON → COGNITRONICS

JACOB RABINOW 1952

NATIONAL BUREAU OF STANDARDS 1952

US & CANADA POST 1953

US AIR FORCE (BORROUGHS) 1959

OCR Geneology contrinued

1960 Farrington Philco-Ford NDP Sperry Univac

1962 Rabinow Engineering *HP* → Control Data

> 1960 Fujitsu Hitachi NCR IBM RCA GE Solatron Scan-Data REI ... *> 50 OCR companies in the United States*)

(40 docs & 10,000 chars per second, $,$$$,$$$)

> 1980 Kurzweil → Xerox → Scansoft → Nuance

↑

Palantir → Calera → Caere ← Recognita

THOCR Expervision Fuji Sanyo RAF IRIS ABBYY...

> 2000 HP → Tesseract → OCRopus GOCR (GNU)

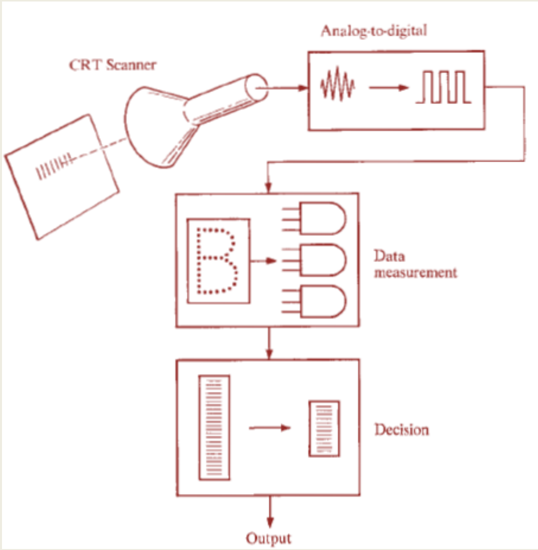
1962 Ray Bonner:
A “Logical Pattern” Recognition Program

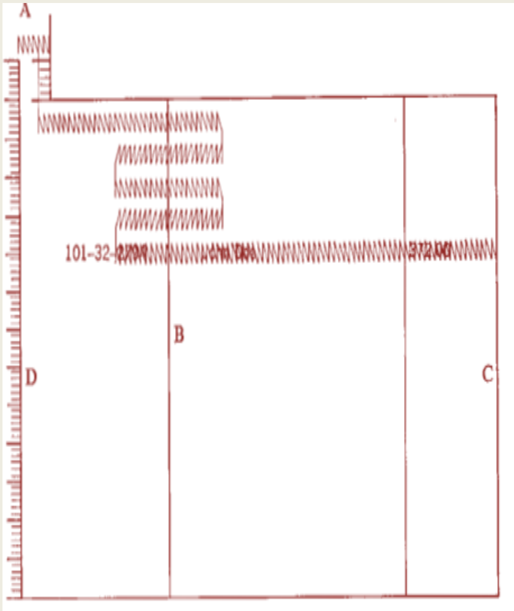
27,519 *unique* 7x10 bitmaps
culled from **1,000,000** samples

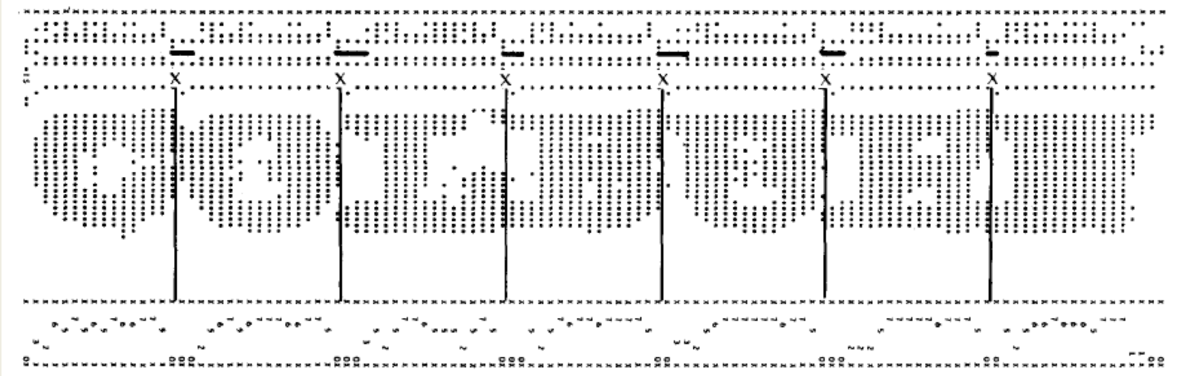
2 substitution errors and 14 rejects

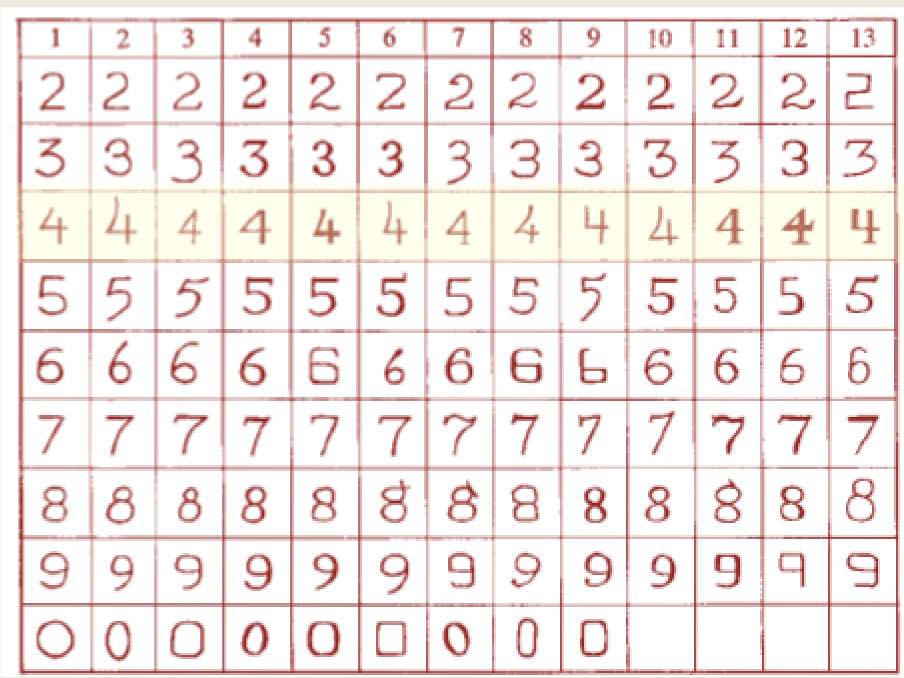
1966: IBM-1975 Social Security Page Reader

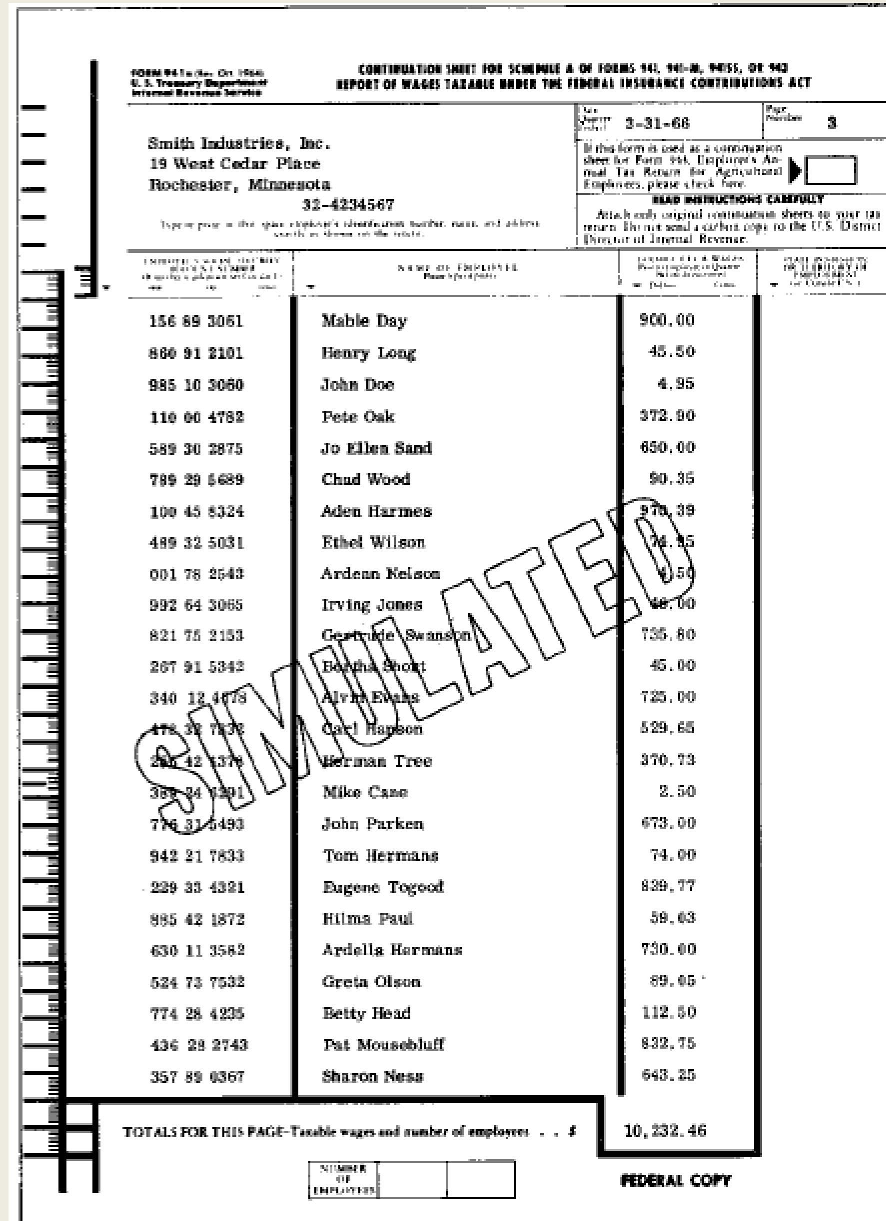
- 500 lines per minute ; 200 fonts
- Last Quarter, 1965:
  1,300,000 pages = 33 million lines
- 16 million lines recognized
- 4 million lines with one or more rejects
- 12 million lines on rejected pages
- OCR confusion probabilities +
  SSA master file of 150 million names
  (one million distinct names with frequencies)
  (no cross-checking between names and SSN's)











FORM 941a (Rev. Oct. 1964) U. S. Treasury Department Internal Revenue Service

CONTINUATION SHEET FOR SCHEDULE A OF FORMS 941, 941-M, 941SS, OR 943
REPORT OF WAGES TAXABLE UNDER THE FEDERAL INSURANCE CONTRIBUTIONS ACT

Smith Industries, Inc.
19 West Cedar Place
Rochester, Minnesota
32-4234567

Quarter ended 3-31-66 — Page number 3

If this form is used as a continuation sheet for Form 943, Employer's Annual Tax Return for Agricultural Employees, please check here.

READ INSTRUCTIONS CAREFULLY

Attach only original continuation sheets to your tax return. Do not send a carbon copy to the U.S. District Director of Internal Revenue.

156 69 3061	Mable Day	900.00
860 91 2101	Henry Long	45.50
985 10 3060	John Doe	4.95
110 00 4782	Pete Oak	372.90
589 30 2875	Jo Ellen Sand	650.00
789 29 5689	Chud Wood	90.35
100 45 8324	Aden Harmes	970.39
489 32 5031	Ethel Wilson	74.85
001 78 2543	Ardenn Nelson	4.50
992 64 3065	Irving Jones	40.00
821 75 2153	Gertrude Swanson	735.60
267 91 5342	Bertha Short	45.00
340 12 4678	Alvin Evans	725.00
478 32 7838	Carl Hanson	529.65
2[illegible] 42 1378	Herman Tree	370.73
38[illegible] 24 6291	Mike Cane	2.50
776 31 6493	John Parken	673.00
942 21 7833	Tom Hermans	74.00
229 33 4321	Eugene Togood	839.77
885 42 1872	Hilma Paul	59.63
630 11 3582	Ardella Hermans	730.00
524 73 7532	Greta Olson	89.05
774 28 4235	Betty Head	112.50
436 28 2743	Pat Mousebluff	832.75
357 89 0367	Sharon Ness	643.25

TOTALS FOR THIS PAGE—Taxable wages and number of employees . . . $ 10,232.46

NUMBER OF EMPLOYEES

FEDERAL COPY

SIMULATED

PATTERN ANALYSIS AND MACHINE INTELLIGENCE

JANUARY 1979 VOLUME PAMI-1 NUMBER 1

A PUBLICATION OF THE IEEE COMPUTER SOCIETY

Reference data sets in the modern era

ETL-8 **1979**
~150,000 kyoiku-kanji characters.

Computer Recognition and Human Production of Handwriting
Eds. R. Plamondon, C. Y. Suen & M. L. Simner

HANDPRINTED CHINESE CHARACTER DATABASE

KAZUO TORAICHI and RYOICHI MORI
Institute of Information Sciences and Electronics, University of Tsukuba, 305 Japan

IWAO SEKITA
Doctoral Program in Engineering, University of Tsukuba, 305 Japan

KAZUHIKO YAMAMOTO and HIROMITSU YAMADA
Information Sciences Division, Electrotechnical Laboratory, 305 Japan

In order to develop and evaluate recognition methods, a common character database is necessary. The authors collected a database of handprinted Chinese characters on a magnetic tape. The database consists of 48,000 characters (4,000 categories × 12 sets). Each character is quantized to 1 bit and is sampled as 64 (horizontal) × 63 (vertical) frames. In this paper, the database is introduced and its fundamental topological features are evaluated. In particular, the number of connected components and that of holes are compared with features obtained from the handprinted Japanese Educational *Kanji* character database (ETL-8).

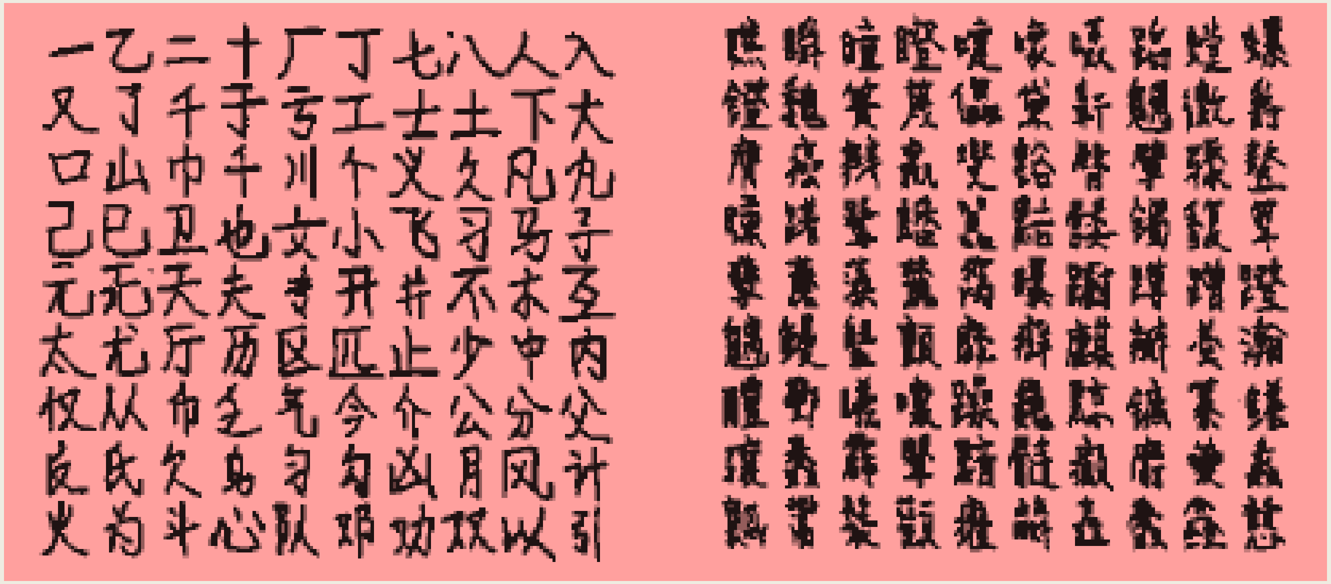
1. INTRODUCTION

1982 ETL-9(B)

T. Saito, H. Yamada and K. Yamamoto,
On the Data Base ETL9 of Hand-printed Characters in JIS Chinese Characters and Its Analysis",
IECE, Vol. J.68-D. No.4, p. 757-764, 1985 (in Japanese)

200 samples each of first level JIS (71 Hiragana + 2965 Kanji)

32 x 32 bitmaps   607,200 characters



Fin-de-siècle sources of western test data

NBS-NIST	1975 -	300K HP alphanumeric
		91K phrases of real census data
CENPARMI	1981 -	100K HP 64x32 alphanumeric
ISRI	1993 - 96	Zoned page images
International Unipen Foundation	1992 -	*On-line* cursive words
CEDAR	1994 -	15K HW “words” from mail
UW I,II,III	1992 - 95	Zoned page images + Synthetics + Graphics
IAM	1999 -	HW sentences

cf. Guyon, Haralick, Hull, Phillips,
Data Sets for OCR and Document Image Understanding Research,
Chapter 30, Handbook of Character Recognition and Image Analysis, Bunke & Wang,
1997

1996

List of Available ISRI Test Datasets

(Nartker, Rice, and Lumos *SPIE/IS&T* 2005)

Ground-truth Test Datasets		Number of		Image Resolution						Used in Annual Test
Sample Name	**Description**	**Pages**	**Characters**	**200 dpi bin**	**300 dpi bin**	**400 dpi bin**	**300 dpi grey**	**Fine-mode Fax**	**Std-mode Fax**	
Sample 2	DOE Sample 2	460	817,946		x					1993 & 94
Sample M	Magazine Sample	200	666,134	x	x	x	x			1994 & 95
Sample N	Newspaper Sample	200	492,080	x	x	x	x			1995
Sample B	Business Letter Sample	200	319,756	x	x	x	x	x	x	1995 & 96
Sample L	Legal Document Sample	300	372,098	x	x	x	x	x	x	1996
Sample S	Spanish Newspaper Sample	144	348,091	x	x	x	x			1995 & 96
Sample 3	DOE Sample 3	785	1,463,512	x	x	x	x			1995 & 96
Sample R	Annual Report Sample	300	892,266	x	x	x	x			1996
Sample Z	Magazine Sample 2	300	1,244,171	x	x	x	x			1996
Totals		**2889**	**6,616,054**							

NIST OCR DATABASES

NIST Scoring Package

NIST Structured Forms Reference Set of Binary Images

NIST Structured Forms Reference Set of Binary Images II

NIST Machine-Print Database of Gray Scale and Binary Images

NIST Miniform Training Database

NIST Miniform Training Database II

NIST Miniform Test Database

NIST Handprinted Forms and Characters

NIST Scientific and Technical Document

NIST Federal Register Document Image Database

Metadata/Text Retrieval Conference 1997 (METTREC)

A new project at NIST (cosponsored by DOD) intended to bring developers of Optical Character Recognition (**OCR**) and Information Retrieval (**IR**) technologies together to study and evaluate the automated production and usability of large-scale, on-line collections of digital documents.

1. How should these collections be constructed in an automated way using OCR technology?
2. What impact do OCR errors have on the accuracy of IR?
3. What types of information should be provided from OCR processing (in addition to text) that will facilitate and enhance IR?
4. What types of data (in addition to text) should be indexed and retrieved to improve the usability of these collections?
5. What types of "real-world" OCR/IR integrated applications can be solved today, solved next year, solved in five years?

To be based on 68,000 scanned pages of a year's worth of the **Federal Register** plus GPO typesetting files (SGML)

What is METADATA?

Foil 3 Mike Garris 1997

- o Non-text elements (physical or logical) of a document
  Fonts, page layout, equations, figures, tables, document type, language, title, author, dates, ...

- o Metadata may be used to construct queries and/or it may be part of the information retrieved.

- o Working definition of metadata will be developed by the planning committee.

PROPOSED SIZE OF DATA SETS

Training Set: 2,000 FR94 page images
2,000 FR94 ground truth files
100 of the 2,000 designated for evaluation
5 known-item queries

Testing Set: 10,000 FR94 page images
10,000 FR94 ground truth files (IR participants only)
100 known-item queries

REASONS FOR CLOSING DOWN THE METTREC PROJECT

RECOGNIZING AND USING METADATA

Our primary interest in METTREC has been to pursue the use of automatically recognized metadata and measure its impact on information retrieval.

Based on our experience over the passed year, we have observed:

a.) An organized "OCR research community" no longer exists. There only remains a small number of commercial vendors competing in a "shrinking“ market.

b.) Very little research has been developed into technology tools for automatically detecting metadata in legacy paper documents that can be used for IR.

c.) There is little motivation for OCR participation.

d.) No one in the IR community is actively researching the use of metadata. It is acknowledged that metadata is interesting and might be useful, but no one is actually trying to exploit it. ....

Conclusion: Metadata cannot be readily detected with existing OCR technology, and the IR community is not prepared to address the use of metadata. Therefore, an OCR/IR metadata evaluation conference is not practical.

1998


R.I.P.

METTREC resurrected: NIST Special Database 25

Volume 1

Federal Register Document Image Database

....roughly 250 issues, comprised of nearly 69,000 pages, published in the Federal Register in 1994.

....The database includes scanned images, SGML-tagged ground truth text, commercial OCR results, and image quality assessment results. ....This volume of the database contains

4711 page images scanned binary

at 15.75 pixels per millimeter (400 pixels per inch). Cost of the database: $210.00

Standard Reference Data

National Institute of Standards and Technology

100 Bureau Dr., STOP 2310, Gaithersburg, MD 20899-2310

CASIA Database (~1990)

Chinese Academy of Science Institute of Automation

3,755 level-1 set of GB2312-80, 300 writers = 1,126,500

Cheng-Lin Liu:

High Accuracy Handwritten Chinese Character Recognition Using Quadratic Classifiers with Discriminative Feature Extraction. ICPR (2) 2006: 942-945

HCL2000 Database

3,755 level-1 set of GB2312-80, 1000 writers = 3,755,000

H. Zhang, J. Guo, G. Chen, C Li:

A Large-scale Handwritten Chinese Character Data base for Handwritten Character Recognition, ICDAR 2009

TUAT HANDS on-line Databases

~ 3,000,000 patterns in ~4000 Kanji + Symbol categories

S. Jaeger, K. Nakashima, ICDAR 2001

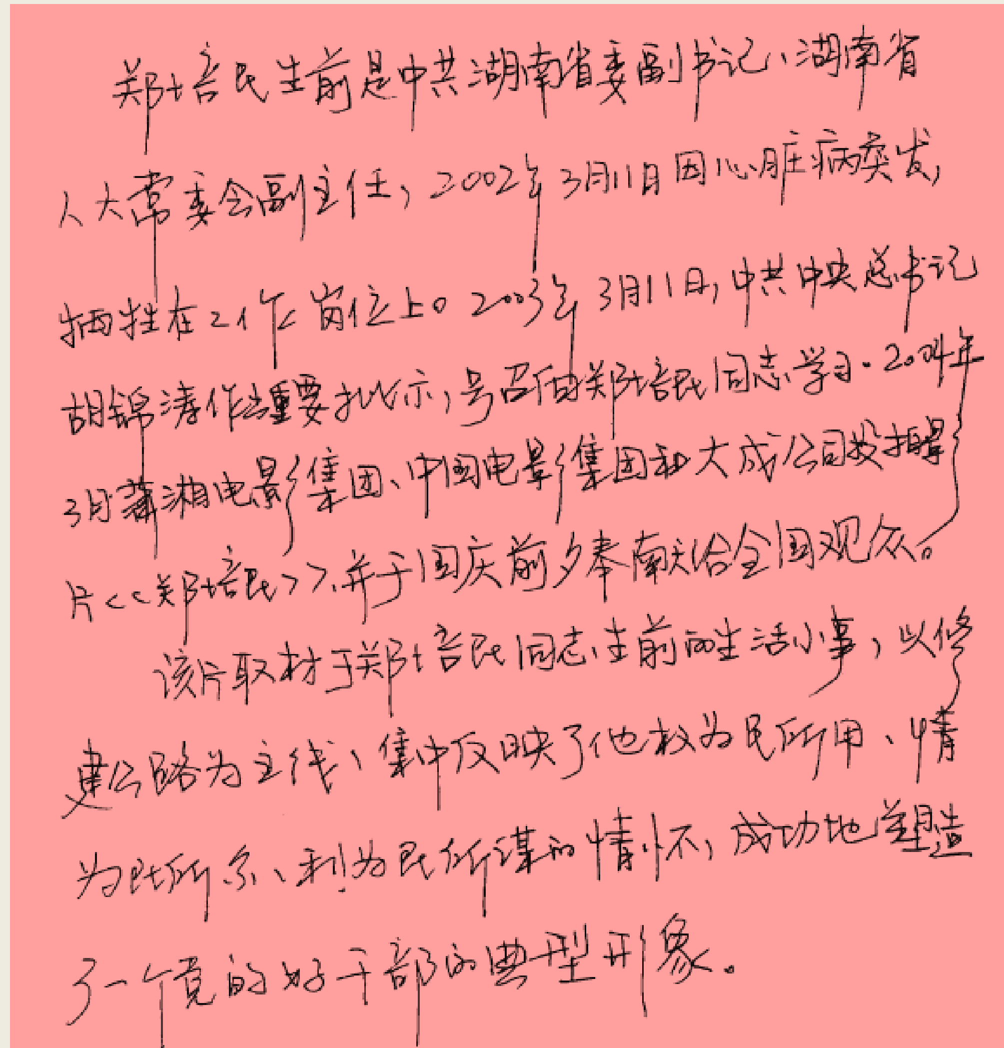
2006: Su, Zhang, Guan: HIT-MW

Unconstrained Chinese handwritten characters without preprinted boxes.

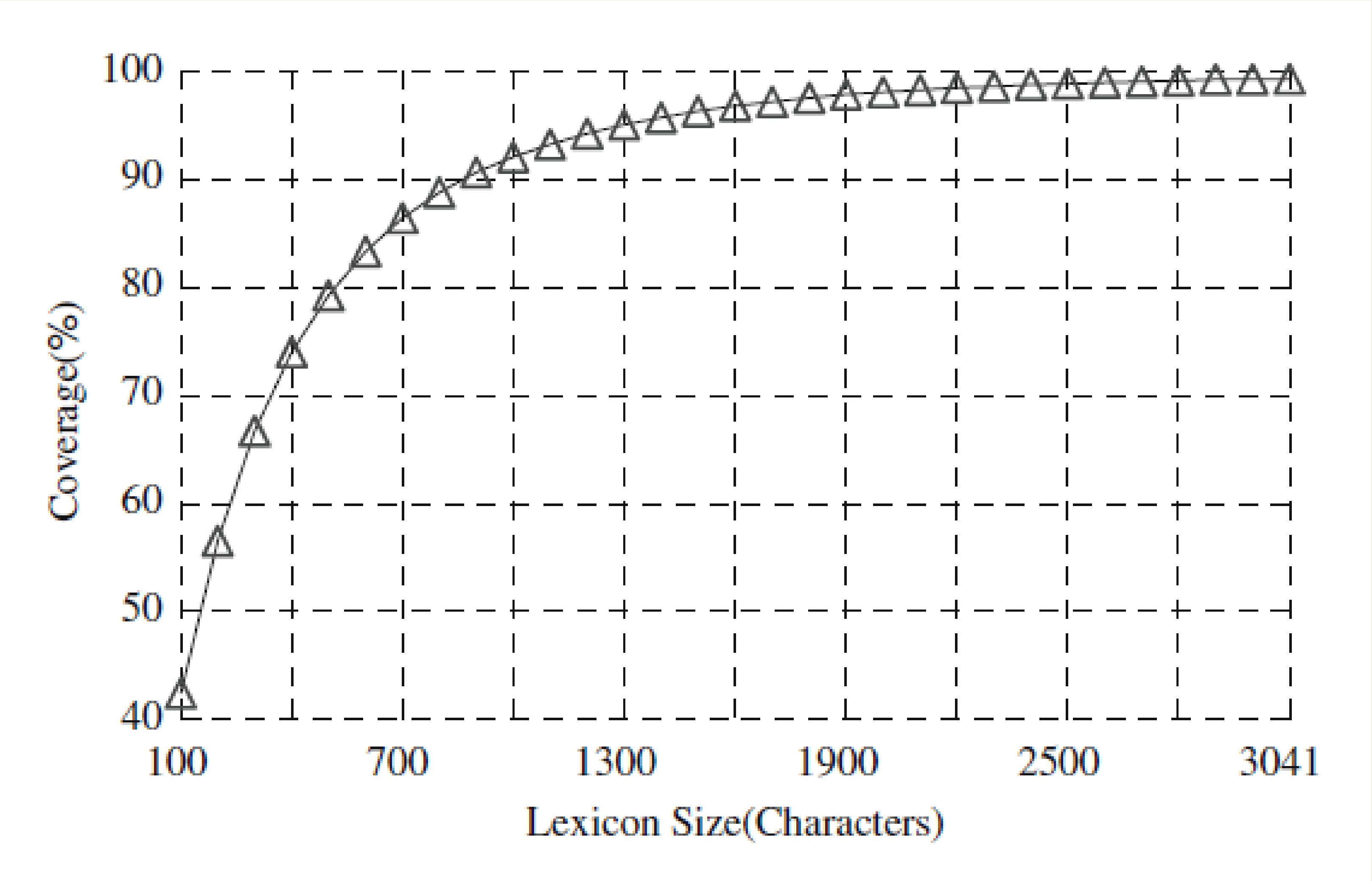
853 forms and 186,444 characters

Gray-scale 300dpi BMP

Texts from China Daily


郑培民生前是中共湖南省委副书记、湖南省
人大常委会副主任，2002年3月11日因心脏病突发，
牺牲在工作岗位上。2003年3月11日，中共中央总书记
胡锦涛作出重要批示，号召向郑培民同志学习。2004年
3月潇湘电影集团、中国电影集团和大成公司投拍影
片《郑培民》，并于国庆前夕奉献给全国观众。
该片取材于郑培民同志生前的生活小事，以修
建公路为主线，集中反映了他权为民所用、情
为民所系、利为民所谋的情怀，成功地塑造
了一个党的好干部的典型形象。

HIT-MW coverage



Su & Wang: "Corpus-based HIT-MW database for offline recognition of general-purpose Chinese handwritten text"

International Journal on Document Analysis and Recognition

Volume 10, Number 1 / June, **2007**

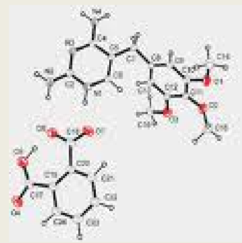
Database	Language	Unit	Year	Source
Highleyman	English	Alphanum	1961	[8]
Munson		Alphanum	1968	[20]
Suen		Numeral	1972	[28]
CENPARMI		Postcode	1992	[30]
CEDAR		City name	1994	[9]
CAMBRIDGE		Sentence	1994	[26]
IAM		Sentence	1998	[16]
IAAS-4M	Chinese	Character	1985	[15]
ITRI		Character	1991	[31]
HCL2000		Character	2000	[36]
HK2002		Character	2002	[5]
ETL-8	Japanese	Character	1976	[19]
ETL-9		Character	1985	[24]
PE92	Korean	Character	1992	[11]
KU-1		Character	2000	[23]
IRONOFF	French	Character	1999	[32]
GRUHD	Greek	Character	2000	[10]
ISI	Indian	Alphanum	2005	[2]

A taxonomy of test data sets

DAS

Engineering drawings |← 40”→|

Schematic diagrams



Maps

Formulas + equations

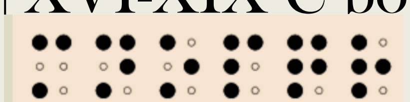
Signatures, letters, diaries

L☺g☺s

Ballots

Illuminated MS, Incunabula,

XVI-XIX C books … … …



T	A	B	L	E	S

Forms -------

CAPTCHAs

NON-DAS

Text (IR, TREK) ?

Cloud, bubble, & spark chambers

Sky (star) pics

Faces

Finger/palm/foot prints

X-rays, CT, MR, PET, ...

Micrographs (cells)

Plants and flowers

Features (U C Irvine)

… … …

Four slides from the NSF-III PI Workshop, April 2010

The Scholarly Practice of Information Integration and Informatics

Haym Hirsh

National Science Foundation

Director, Information and Intelligent Systems (CISE)

UC Irvine Machine Learning Repository

- Over 150 Data Sets for Machine Learning and Data Mining
- Median age: 15 years

- Should data sets have an expiration date?

Benchmark Data: Reproducible, but Good?

- Benchmark data allow reproducibility
- Reproducibility also depends on software, parameters, etc.
- Benchmark data sets must be representative of the sorts of problems our algorithms will see in practice
- Benchmark data sets must stay timely as technological and scientific advances allow our ambitions to grow
- The pace of data growth makes this difficult

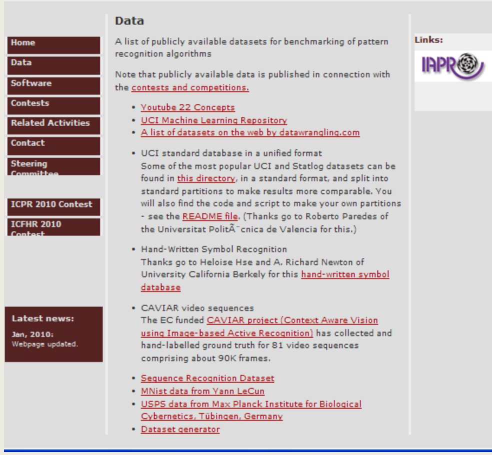
Proprietary Data/Information

- Many innovative ideas involve proprietary or otherwise restricted data/information
- Do we publish innovations that use generally unavailable data/information?

- What can we learn from the medical community?

2010

IAPR TC 5: Benchmarking & Software



Data

A list of publicly available datasets for benchmarking of pattern recognition algorithms

- A list of datasets on the web by datawrangling.com
- UCI standard database in a unified format
- Hand-Written Symbol Recognition
- CAVIAR video sequences
- Sequence Recognition Dataset
- MNist data from Yann LeCun
- UCI Machine Learning Repository
- Youtube 22 Concepts
- USPS data from Max Planck Institute for Biological Cybernetics
- Dataset generator

2010

Machine-print OCR

- APTI: Arabic Printed Text Image Database

Handwriting

On-line

- IAM On-Line Handwriting Database
- UNIPEN database
- Kuchibue & Nakayosi Reference: "Collection of on-line handwritten Japanese character pattern databases and their analysis," International Journal on Document Analysis and Recognition, Vol. 7 No. 1, pp.69-81 (2004).

Off-line

- CEDAR Off-line Handwriting CDROM1
- IAM Database - A full English sentence database for off-line handwriting recognition.
- MARG- Medical Article Records Groundtruth ([1]) is a freely-available repository of document page images and their associated textual and layout data. The data has been reviewed and corrected to establish its "ground truth". Please contact Dr. George Thoma (thoma@lhc.nlm.nih.gov) at the National Library of Medicine for more information.
- Hindi font samples by Andras Kornai, June 5 2003

Miscellaneous Kanji handwritten OCR databases

- IPTP CD-ROM2

comments on test data

from [*www.nagy.virtual.blog*](http://www.nagy.virtual.blog)

When is N large enough? (Gauss, Chernoff, Guyon, Makhoul, Schwartz, Vapnik)

$$n \approx \left(\frac{z_a}{\beta}\right)\frac{(1-p)}{p} \approx \left(\frac{z_a}{\beta}\right)\frac{1}{p} \approx \frac{2\ln\alpha}{\beta^2 p} \approx \frac{100}{p}$$

So keep sampling until 100 errors have occurred
for a 5% (α) probability that the true error rate is no more than 20% (β) greater than the observed value *p*.

Which is the better classifier?

Track exclusive errors v_1 and v_2 by each classifier.
Then the first classifier is probably better if

$$\hat{p}_1 - \hat{p}_2 \geq \frac{z_\alpha}{n}\sqrt{v_1 - v_2}$$

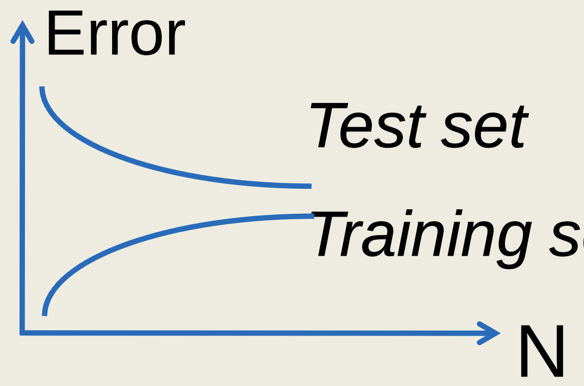
Training on Test Data

- Blatant — e.g. estimating language model from a *transcript* of the test data.
- Subtle — e.g. increasing *P*(*q* _) after finding 38 occurrences of ***Iraq*** in the test data
- Subconscious — failing to avert eyes from test data

If you make more than one run on the test set,

report the maximum error!

Testing on Training Data:



Data Partitioning

- Training – Test (50-50 or 90-10 ?)
- Training – Validation – Test
- Training – Validation 1 – Validation 2 – Validation 3 ... Test
- Leave-one-out, Leave-N-out (cross-validation)
- Non-stationary data
- Sampling with or without replacement

80K samples: 400 sources (writers or fonts), 200 sentences

25% sample: 100 sources, 200 sentences

200 sources, 100 sentences

400 sources, 50 sentences ?

Representative of what?

- Real data for HP so far primarily from envelopes and checks. Predicting real-world accuracy requires a document census or data from *actual operations*.

- Copying is different from writing or calculating!

  Vocabulary, spelling, and grammar are as personal as shape: writer demographics are important (cf. NIST SD3/7)

- In the real world, uniform distributions are rarer than a hen's tooth! (cf. Benford's Law for *digits*, Zipf's Law for *words*)

- Ground truth depends on *context*.

Reporting classifier "performance"

Two-class Confusion Matrix: *5 independent values!*

Correct, Error, Reject (C+E+R=1)

Recognition Rate

Precision & Recall (F-score/measure)

Sensitivity & Specificity

Type I &Type II error

False Alarm & Miss

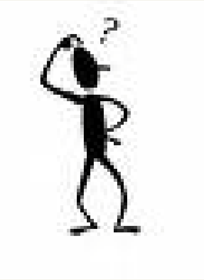
False Positive & False Negative

(+ is good in war, bad in medicine)

Errors of Omission & Commission

Reject-error curve (C.K. Chow 1970) ROC

N=100		Classified		
True		A	B	Reject
	A	*50*	***3***	*7*
	B	***4***	*30*	*6*



Preprocessing considered harmful

Please don't

binarize,

normalize,

segment,

deskew,

denoise,

thin, fatten,

or otherwise "improve" test data!

Instead, save your algorithm and parameter settings in the database.

***Wizard Words* mean exactly what *you* want them to mean:**

Concept
Model
Semantics
Context
Ontology

Knowledge
Information
Data
Interpretation
Understanding

as in: *“We situate our semantic concept models in an ontological context.”*

Un-, non-, semi-supervised;
adaptive, self-adaptive,
teaching, training, learning

Document

Noise

- Real noise is *never* i.i.d.
- Bits don't flip by themselves.
- Random-phase noise is unavoidable but benign.
- Bleed-through and leaky margins are infectious.
- Document noise is different from digitization noise.
- A good scanner is worth three noise filters.
- Adding noise to develop robust algorithms is like ear-training with a gong!
- Every *n*th document should be a calibration target.

Proofreading in context is error prone

This is a rcvicw, from an intuitivc rathcr than a mathcmatical pcrspcctivc, of thc statistical foundations of adaptivc rccognition systcms. Kcy considcrations in adaptivc classification arc priors, samplc sizc and sampling stratcgy, labcls, statistical dcpcndcncics, and dimcnsionality. Thc small-samplc bias and variancc of maximum likclihood, maximum *a postcriori* and Baycs cstimators arc comparcd in a small concrctc casc. Itcrativc cxpcctation maximization for cstimating thc sufficicnt statistics of mixturcs is illustratcd in a simplc sctting. It is shown that corrclation among fcaturcs is somctimcs unjustly maligncd. A countcrintuitivc incrcasc in thc crror ratc aftcr adding a sccond fcaturc is traccd to thc cursc of dimcnsionality. Adaptivc classification is prcscntcd in thc contcxt of both paramctric and non-paramctric (ncarcst ncighbors and ncural ncts) cstimation. Somc rcccnt thcorctical rcsults and not-so-rcccnt cxpcrimcntal obscrvations on hybrid classification (bascd on both labclcd and unlabclcd samplcs) arc summarizcd.

Ack: X-ing a Paragrab, Edgar Allan Poe, 1850

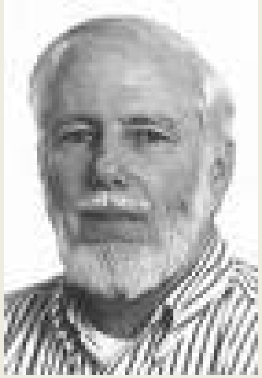
Homogenous class display for OCR

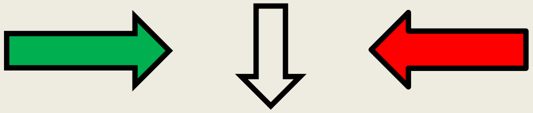
c c c e c c c c c c c c c e c c c c c c e c
c c c c c c c e c c c c e c c c c c c c c c
c c c c c c c c c c e c c c c c c c c c c c
c e c c c c c c c c c c c c c c c c c e c c
c c c c c e c c c c e c c c c c c c e c c c
c c c c e c c c c c c c e c c c c c c e c c

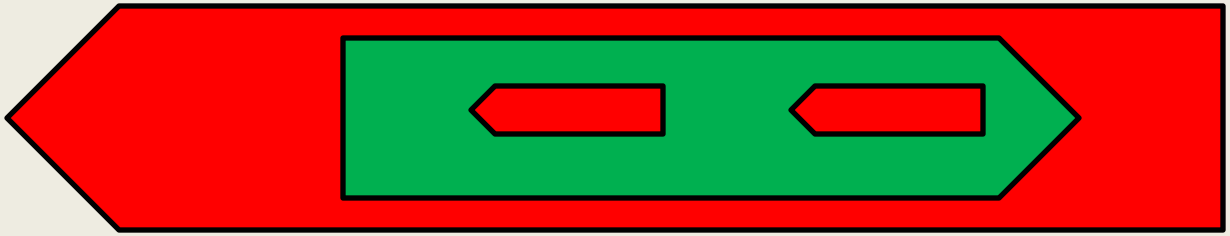
A A A R A A A A A A A A A O A A A A A A J A
A A A A A A A T A A A A O A A A A A A A A A
A A A A A A A A A A U A A A A A A A A A A A
A F A A A A A A A A A A A A A A A A A D A A
A A A AA A A O A A A A Q A A A A A A A H A A
A A A A A P A A A A A A A R A A A A A A A A

11%

Foreign is relative



- Language model
- Σχριπτ
- Allographs and variants
- Diacritics (English has only *tittles*)
- Numerals, punctuation
- 
- Pure or mixed



Keep on testing, testing, testing:

progress in document analysis has long been driven by sound experiments on carefully prepared test data.

Thank you