Classifiers that improve with use

George Nagy
DocLab
Rensselaer Polytechnic Institute

Argument

In-house training sets are never large enough, and never representative enough. We must therefore augment them with samples from actual (real-time, real-world) OCR operation.

We present some methods to this end.

Outline

Non-representative training sets

Supervised learning (continuing classifier education)

"Unsupervised" adaptation

Self-corrective, Decision-directed, Auto-label

Symbolic Indirect Correlation (SIC) *new* ***

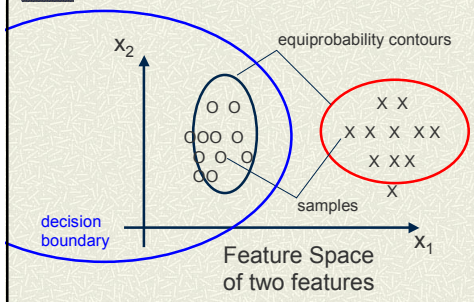
Style-constrained classification

Weakly-constrained data distributions (*new* ***)

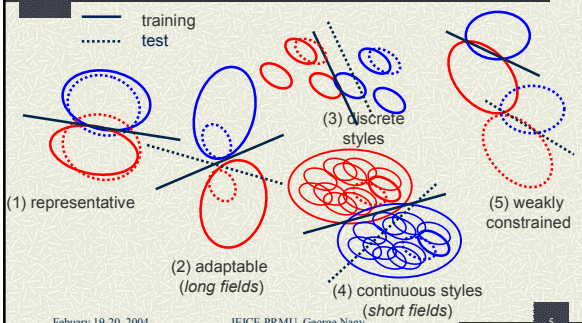
Linguistic context

Recommendations

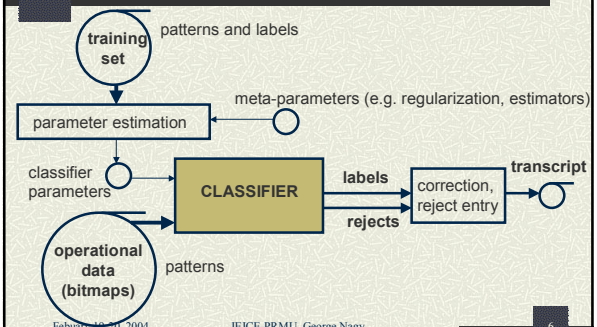
Representation



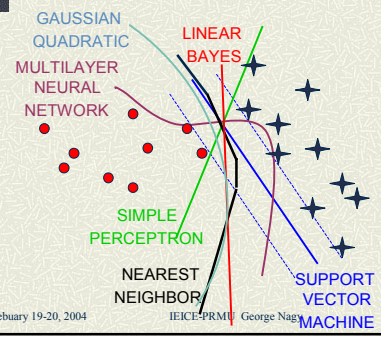
How representative is the training set?



Traditional open-loop OCR System

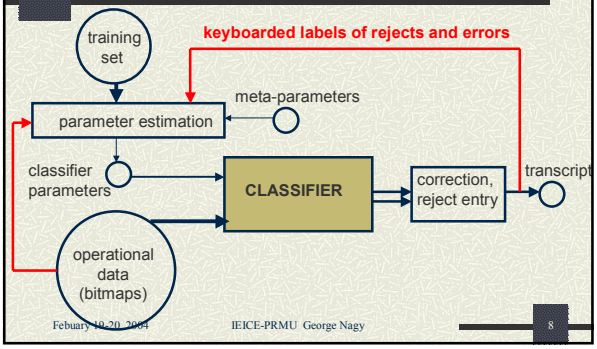


Some classifiers



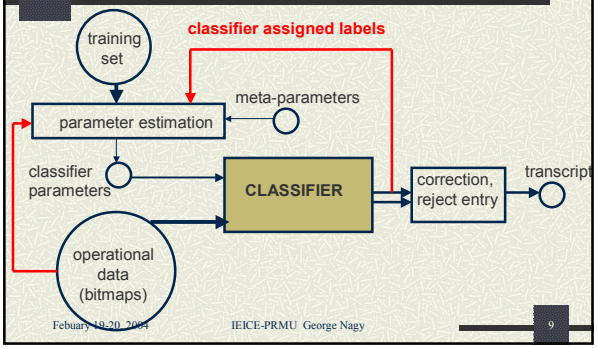
Supervised learning

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

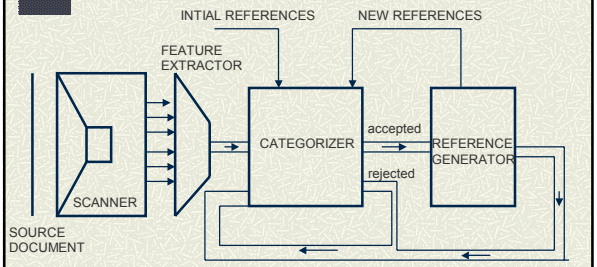


Adaptation (DHS: "Decision directed approximation")

Field estimation, singlet classification



Self-corrective recognition (1966)



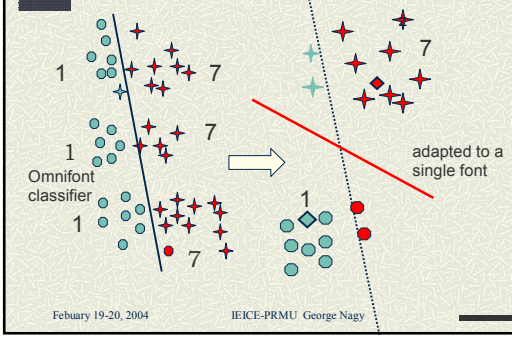
Results: self-corrective recognition (Shelton & Nagy 1966)

```
Training set:  9 fonts,  500 characters/font, U/C
Test set:     12 fonts, 1500 characters/font, U/C
96 n-tuple features, ternary reference vectors

Initial error and reject rates:  3.5%    15.2%
After self correction:           0.7%     3.7%
```


Decision-directed adaptation

aka self-corrective recognition, auto-label adaptation, semi-supervised learning, ...



Results - Baird & Nagy (DR&R 1994)

100 fonts, 80 symbols each from Baird's defect model

(6,400,000 characters)

Size (pt)	Error reduction	% fonts improved	Best	Worst
6	x 1.4	100	x 4	x 1.0
10	x 2.5	93	x 11	x 0.8
12	x 4.4	98	x 34	x 0.9
16	x 7.2	98	x 141	x0.8

Results:
adapting both means and variances

(Harsha Veeramachaneni 2003)

NIST Hand-printed digit classes, with 50 "Hitachi features"

Train	Test	% Error Before	Adapt means	Adapt variance
SD3	SD3	**1.1**	0.7	**0.6**
	SD7	**5.0**	2.6	**2.2**
SD7	SD3	**1.7**	0.9	**0.8**
	SD7	**2.4**	1.6	**1.7**
SD3+SD7	SD3	**0.9**	0.6	**0.6**
	SD7	**3.2**	1.9	**1.8**

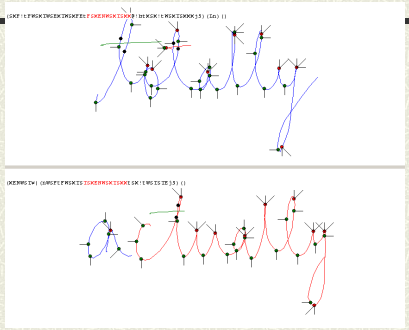
InkLink Adnan El-Nasan (2003)

On-line handwriting recognition

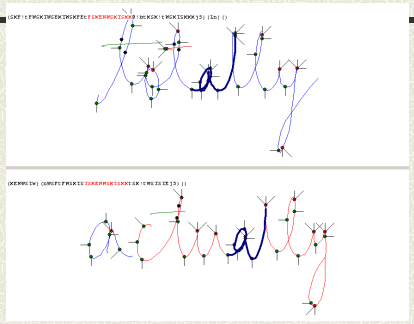
Constrained *localized* polygram matching
one unknown word against many reference words,
using a lexicon of legal words.

The reference set does **not** include most
of the lexicon words!

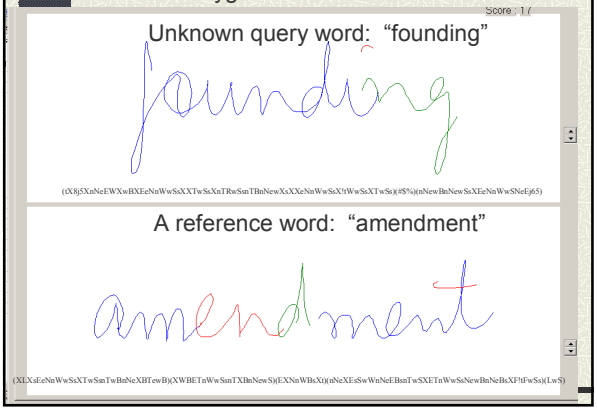
From electronic ink to feature string



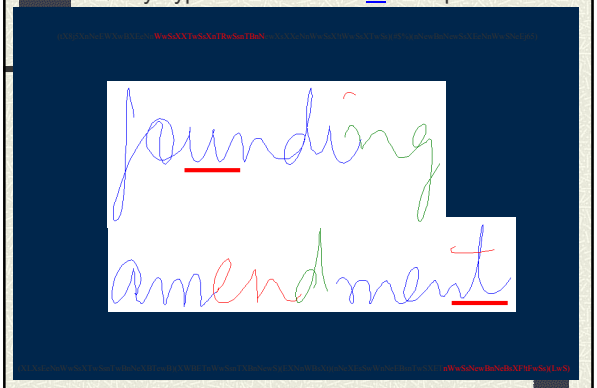
Feature matching



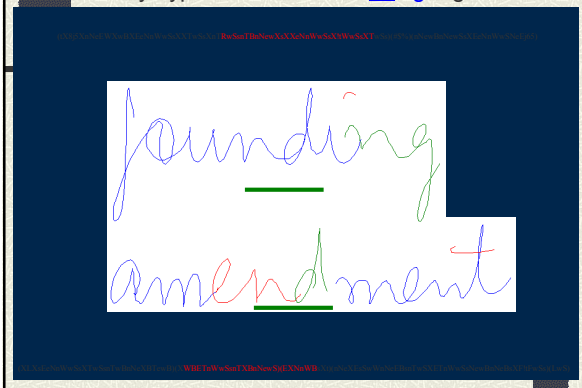
Polygram feature match



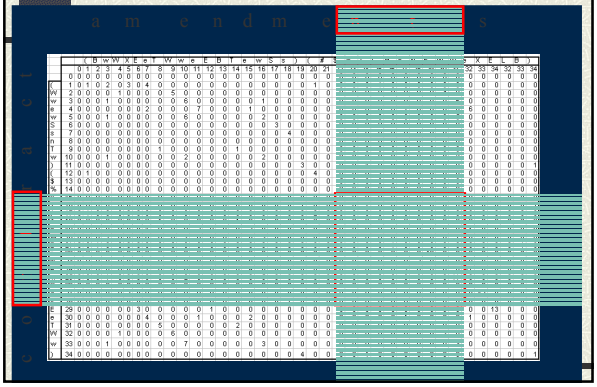
Query hypothesized as "contract": poor match



Query hypothesized as "founding": good match



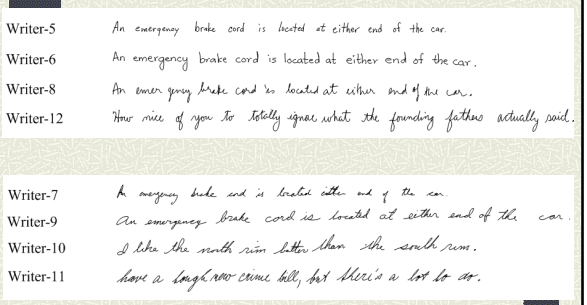
Localized Viterbi trellis search



InkLink classification algorithm

1. The **expected location** where the unknown matches the reference words is pre-computed
2. The features matches of the unknown against the reference words are found by **string matching**.
3. The hypothesis that corresponds best to the **expected length and location** of the matches is chosen.

Our most/least favorite writers



Comparison with external system

(four writers we like) 100-word lexicons

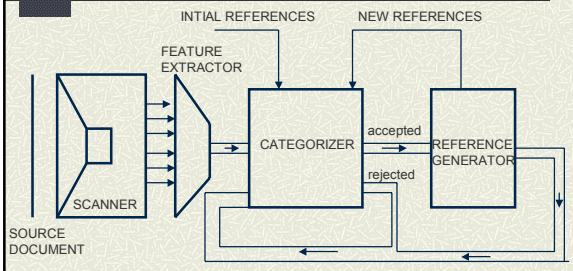
External System

ID / %	4	5	6	8	Average
Top-1	96.7	95.7	99.7	95.3	96.9
Top-10	98.3	95.7	99.7	97.7	97.9

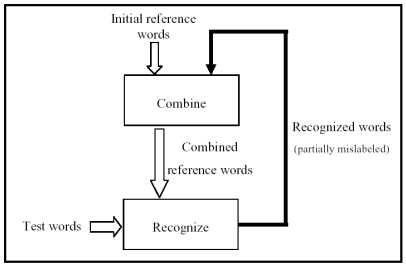
InkLink

ID / %	4	5	6	8	Average
Top-1	94.0	97.3	94.7	94.0	95.0
Top-10	99.0	99.0	99.7	99.0	99.2

Self-corrective recognition (1966)



Auto-label adaptation



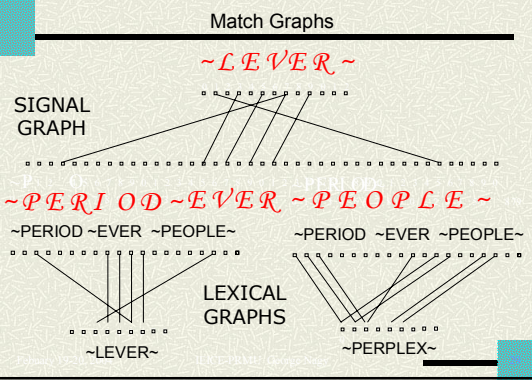
Results of adaptation ("auto-label") in InkLink

- Error rate dropped from 28% to 7% .
- As good with 100 reference words as with 500 reference words without adaptation.

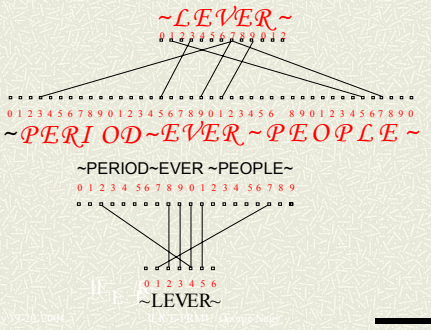
Outline

- Non-representative training sets
- Supervised learning (continuing classifier education)
- "Unsupervised" adaptation
  - Self-corrective, Decision-directed, Auto-label
- **Symbolic Indirect Correlation (SIC)**
- Style constrained classification
- Weakly-constrained data distributions
- Linguistic context
- Recommendations

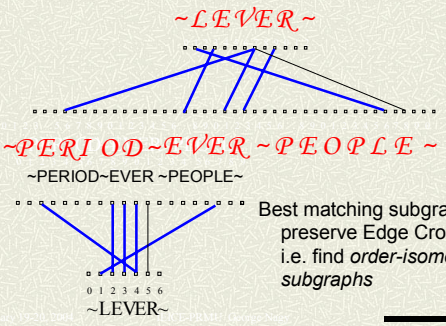
Symbolic Indirect Correlation (SIC)



Signal graph of lever compared to reference signal graph

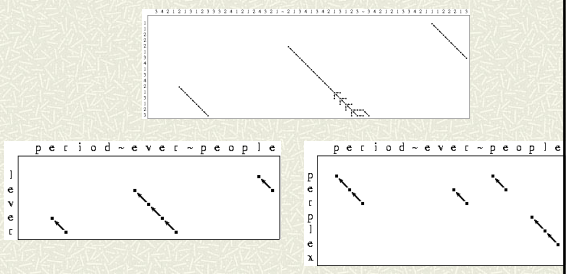


Matching Match Graphs



Best matching subgraphs preserve Edge Crossings, i.e. find *order-isomorphic subgraphs*

Signal and lexial graphs (handwriting)



Results: SIC (simulation only!)

1000 most common words of Brown Corpus

Weighted Noise Model

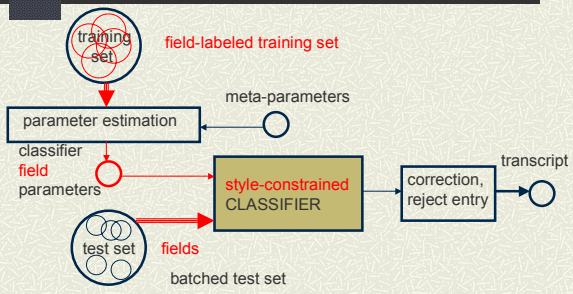
Probability of Wrong Match:	0.4	0.6	0.8	1.0
% Correct Recognition:	100	100	100	96
Reference String Size:	390	500	600	750

Adaptation:
we can add recognized words to the reference string, as in InkLink

Outline

- Non-representative training sets
- Supervised learning (continuing classifier education)
- "Unsupervised" adaptation
  - Self-corrective, Decision-directed, Auto-label
- Symbolic Indirect Correlation
- **Style constrained classification**
- Weakly-constrained data distributions
- Linguistic context
- Recommendations

Style consistency:
Field estimation, field classification



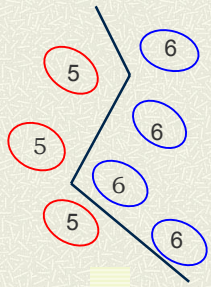
Single-class and multi-class style

	SINGLE CLASS STYLE	MULTI-CLASS STYLE
Source 1:	29/05/1925	25/07/1922
Source 2:	15/05/1990	**05/05/1925**
Source 3:	21/06/1943	**02/06/1943**
Source 4:	05 /29/1945	02/25/1942

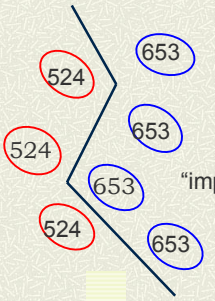
Styles are induced in a collection of documents by multiple sources*.

* fonts, printers, scanners, writers, speakers, microphones, ...

Multi-mode parametric classifier for single-class style constraints



Field classifier for strong style constraints



"implicit font recognition"

Field-trained classification versus style-constrained classification

Field Length = 4

Training set for field classification	Training set for style classification
0000	00
0001	01
0010	02
	98
	99
9998	(10^2 classes)
9999	
(10^4 classes)	

classifier parameters for longer field length computed from pair parameters
(because Gaussian variables defined completely by covariance)

Results: style-constrained classification - short fields (Harsha V.)

Continuous style-constrained classifier, trained on ~ 17,000 characters and tested on ~17,000 characters.
25 top principal component "Hitachi" blurred directional features.

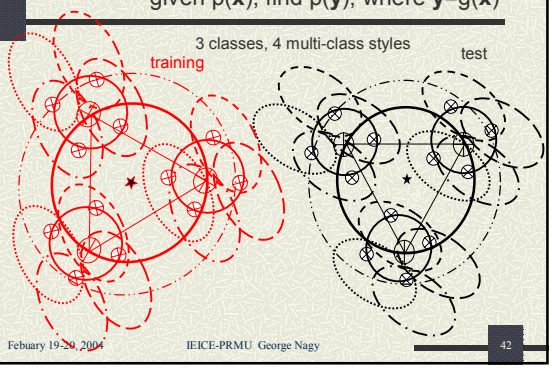
	Field error rate (%)			
Field length:	L=2		L=5	
Test data	w/o style	with style	w/o style	with style
SD3	1.4	1.3	3.0	2.5
SD7	2.7	2.4	5.3	4.5

Outline

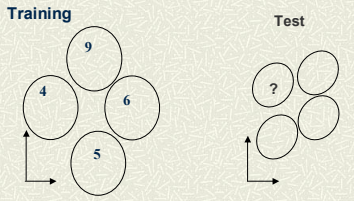
- Non-representative training sets
- Supervised learning (continuing classifier education)
- "Unsupervised" adaptation
  - Self-corrective, Decision-directed, Auto-label
- Symbolic Indirect Correlation
- Style constrained classification
- **Weakly-constrained data distributions**
- Linguistic context
- Recommendations

Weakly-constrained data

given p(**x**), find p(**y**), where **y**=g(**x**)



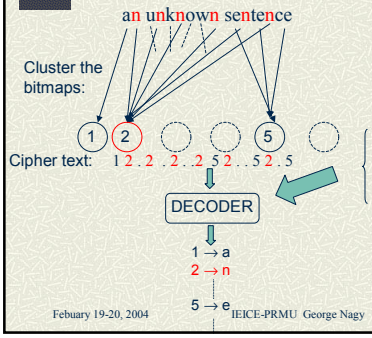
Are weak constraints enough?



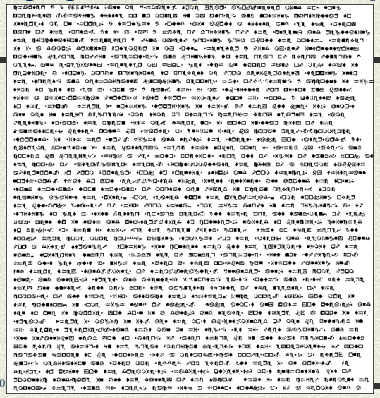
Outline

- Non-representative training sets
- Supervised learning (continuing classifier education)
- "Unsupervised" adaptation
  - Self-corrective, Decision-directed, Auto-label
- Symbolic Indirect Correlation
- Style constrained classification
- Weakly-constrained data distributions
- **Linguistic context**
- Recommendations

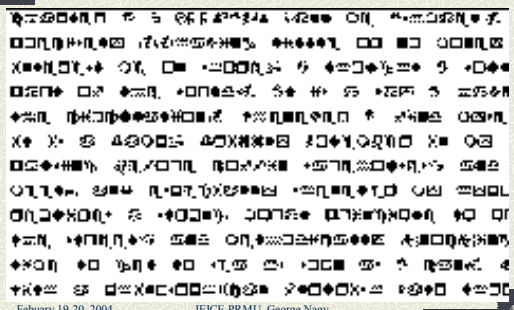
Language context: decoding a substitution cipher



Language context [with Tin Ho 2000]



Text printed with Spitz glyphs



Decoded text

chapter I 2 LOOMINGS
Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have …

chapter i _ bee_inds
_all me ishmaels some years ago__never mind how long precisely __having little or no money in my purses and nothing particular to interest me on shores i thought i would sail about a little and see the watery part of the worlds it is a way i have …

Style context versus Language context

Two digits in an isogenous field: ..... 5 6 .....

with feature vectors: $\mathbf{x}_i$ $\mathbf{x}_j$

and class labels: C_i C_j

Style context: $P(\mathbf{x}_i, \mathbf{x}_j \mid C_i=5, C_j=6) \neq P(\mathbf{x}_i \mid C_i=5)\, P(\mathbf{x}_j \mid C_j=6)$

Language Context: $P(C_i=5, C_j=6) \neq P(C_i=5)\, P(C_j=6)$

Recommendations for OCR systems that improve with use

Never let the machine rest:
design it so that it puts every coffee-break to good use.

Don't throw away edits (corrected labels): use them.

Classify style-consistent fields, not characters:
adapt **on long fields, exploit multi-class *style* in short fields.**

Use order rather than position.

Let the machine guess: lazy decisions.

Make use of all possible context:
language, shape, layout, and function.

Every document can be read by the right reader.

The End

I am grateful to Hitachi CRL for technical and financial support since 1992. Special thanks to Drs. Fujisawa, Liu, Sako, and Shima, and to Messrs. Koga, Marukawa, and Mine.

I also learned much from my former and present students Jung, Sarkar,Veeramachaneni, & El-Nasan.

But any misconceptions in this presentation are entirely my own!

Thank you!

Non-supervised classification is always semi-supervised classification

Only the supervisory constraints are hidden!

Clustering: cardinality, diameter, population, or sequence of clusters.

ML / EM : form of distribution, sigma, number of mixture components.

Trees, neural networks: validation set.

InkLink classification algorithm

$$P\left(c_i \mid l_{ij}\right) = \frac{P\left(l_{ij} \mid c_i\right) P\left(c_i\right)}{P\left(l_{ij}\right)}$$

$$c^* = \arg\max_i \prod_j \frac{P\left(l_{ij} \mid \hat{l}_{ij}, M_{ij}=1\right)}{P\left(l_{ij} \mid \hat{l}_{ij}, M_{ij}=0\right)}$$

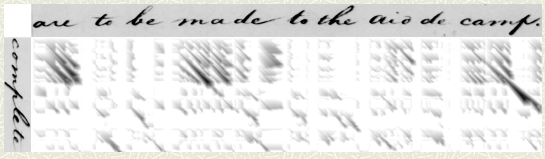
l_{ij} : Observed feature match length

$\hat{l}_{ij}$: Predicted feature match length

M_{ij} : Match

Four classifiers, based on different features sets, are combined by Borda Count

A letter from George Washington



Feature-string matching:

"*complete*" vs. "*are to be made to the aid de camp.*"

CONFERENCE REPORT ON ICDAR '07

George Nagy
Professor Emeritus (??), Rensselaer Polytechnic Institute

OUTLINE:

VENUE
INVITED SPEAKERS
AWARDS
SESSION TOPICS
NOTABLE CONTRIBUTIONS
LARGE-SCALE PROJECTS
OUTLOOK

(Slides from my closing talk at ICDAR 1997 in Ulm, Germany)

INVITED SPEAKERS

(all second generation)

Prof. Tin Kam Ho
Web-wide Voting Network for Weak Classifiers
Exploratory data analysis at Lucent Bell Labs
Dr. Tapas Kanungo
Multimedia Document Devastation Models
DIA at IBM Almaden Research Center
Dr. Rainer Hoch
High-speed Context Switches
Just moved from SAP to the Mannheim Berufsakademie
Dr. Abdelwahab Zramdini
Web Typography
Top secret research at a bank in Geneva
Prof. Omid Kia
Processing Compressed Multimedia Documents
Founded a company, then disappeared!

III. Devices

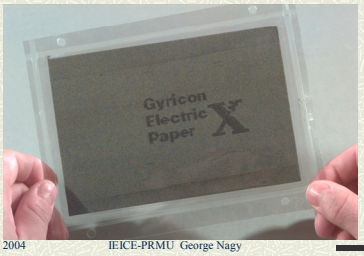
Storage
Docupin, DocuPinCushion (DPC)
750 MB/DP, 18GB/DPC
Not addressable

... イベントのお知らせ "Memory Stick Xmas Magic". 楽しくお得な各種キャンペーンをご紹介します！！... 最新情報誌 Memory Stick Update. ...
SANDISK UNVEILS THE HIGHEST CAPACITY MEMORY STICK PRO STORAGE CARD IN THE WORLD—2 GIGABYTES.

SONY
SANDISK

E-Scap

Cholesteric polymer-dispersed hybrid substrate
10 micron thick (100 sheets per cm)
Ten megacell array provides 300 dpi resolution
10**6 read-write cycles
Ecologically benign (silicon, not carbon, based)
No color so far



Radiotablet

Lambertian high-contrast *reflective* display
Satellite cyberband channel
Pentameric (soft, 4mm thick elastomere) vary-keyboard
100 MB flash-cache
Docupin port (USB)

Tablet PCs
Toshiba's Convertible Tablet PC paves the way ...

VI. Oral Documents

Speech-to-print
Print-to-speech
Convergence of speech and document image processing - began in 1990's with adoption of HMM in OCR

AUDIO DOCUMENTS

So far, primarily for entertainment (music)
(almost everything is primarily for entertainment!)
Some natural sounds (bird calls, frogs, dolphins)
Oral history (eye-witness accounts)
Personal note-taking (key-chain recorders)
Instructional audios ("prompter")
???

handwriting

Now used mainly for *self-communication*

letters already replaced by email

forms being replaced by e-forms

Children learn to type in pre-school!

Therefore personal electronic ink recognition
is becoming more important than
scanned handwriting recognition.

Other questions?