

# *DRR IS A TEENAGER!*



George Nagy

DocLab

Rensselaer Polytechnic Institute

nagy@ecse.rpi.edu





00001111!



*You are  
bright,  
outgoing,  
cooperative,  
imaginative,  
inventive!  
The extended family is proud of you!*

# Precursors:

**1991**

## ***High-Speed Inspection Architectures, Barcoding, and Character Recognition***

SPIE Volume 1384

Editor: **Michael J. W.Chen**

Table recognition for automated document entry system

Haruhiko Kojima; Teruo Akiyama

Japanese document recognition and retrieval system using programmable

SIMD processor: Sueharu Miyahara; Akira Suzuki; Shunkichi Tada;

Takahiko Kawatani

**1992**

## ***Machine Vision Applications in Character Recognition and Industrial Inspection***

SPIE Volume 1661

Chairs/Editors: **Donald P. D'Amato, Wolf-Ekkehard Blanz,  
Byron Dom, Sargur N. Srihari**

# #1: *Character Recognition Technologies*

February 1-2, 1993



Chair: **Donald D'Amato**

# Volume 1906

Contributed paper, pp. 134-145:

Gary Kopec and Phil Chow

*Document image decoding  
using Markov source models*



# #2 Document Recognition 1994

Luc Vincent



Theo Pavlidis



Volume 2181

# KEYNOTE

Gary Kopec and Phil Chou

*Communication theory framework for  
document recognition*

**Document Image Decoding:**

Whole page recognition with stochastic attributed  
context-free grammars.

Based on earlier work on a text-image editor.

Image EMACS

# Other gems from 1994

- **Table recognition:** Rahgozar, Fan, Rainero
- **Latent character shape coding:** Spitz
- ***the*:** Khoubyari, Hull
- **Word recognition by collocation:** Hong & Hull
- **Information metrics:** Nartker
- **Error classification:** Esakov, Lopresti, Sandberg
- **Segmentation metrics:** Randrimasy, Vincent, Wittner
- **Post-correction:** Taghva, Borsack, Condit
- **Music-notation recognition:** Fahmy & Blostein
- **Fax restoration:** Handley, Dougherty

# More gems

- **Line drawings:** Kasturi et al.
- **Skew detection:** Besho, Ejiri, Cullen
- **Recognition w/o segmentation:** Al Badr, Haralick
- **Saddle features:** Rocha, Sakoda, Zhou, Pavlidis
- **Survey of DIA:** Ablameyko, Bereishik
  
- **Arabic OCR:** Allam  
(98.5%-99.7% for typewritten, 98.1%-99.3% for typeset)

*NB: Every 1994 paper dealt with scanned docs*

# DRR at 7 (2000)

- Chinese and Persian OCR in addition to Arabic & Devanagari
- Full-blown character shape coding by Larry Spitz
- Richard Fateman on mathematical equations
- ISRI releases its DOE database to researchers
- Document segmentation:
  - pre-zoned, perspective, color, multi-scale, “medium independent”
- *Text categorization*

# Adolescence brings expanding interests:

- On-line OCR
- Specialized DBs: *MEDLINE*, *FR*, *LSS*
- Photographs of text, digital video
- More Asian languages
- Preprocessing → Restoration
- IR for HW
- Document authentication
- Access authentication (CAPTCHAs)

# Last Year

- *Real* OCR (Istvan Marosi)
- ICR (better than SCR!)
- Camera, multispectral and tablet input
- Coded symbol data
- Transcript mapping for Arabic
- Document CBIR and content inventories
- Digital publishing and libraries
- *Shape descriptors*

# Some regulars



1/27/2008

SPIE/DRR SJ 2007

In addition to inventive  
and faithful participants,

a good conference takes

- a diligent program committee
- dependable session chairs
- competent administration
- reliable tech support
- long breaks and plenty of coffee
- and, above all, dedicated chairs!

- III Luc Vincent & Jonathan Hull
- IV Luc Vincent
- V Dan Lopresti & Jianying Hu
- VI Dan Lopresti & Jianying Hu
- VII Dan Lopresti & Jianying Hu
- VIII Paul Kantor, Tapas Kanungo, Jianying Hu
- IX Paul Kantor, Tapas Kanungo, Jianying Hu
- X T. Kanungo, E. Barney Smith, J. Hu, P. Kantor
- XI Elisa Barney Smith, Jianying Hu, James Allen
- XII Elisa Barney Smith, Kazem Taghva
- XIII Kazem Taghva, Xiaofan Liu
- XIV Xiaofan Liu, Berrin Yanikoglu
- XV Berrin Yanikoglu, Kathrin Berkner

DR



DRR

# Recognition **versus** Retrieval

Page *scanned* at 300 dpi grayscale ~ 10 MB  
(B/W ~ 1MB, compressed ~ 30 KB)

Page of *encoded* text ~ 3500 chars incl blanks  
(compressed ~ 1 KB

.doc ~ 30 KB

.pdf ~ 10 KB

.rtf ~ 10KB

.txt ~ 4 KB)

# Recognition **versus** Retrieval (cont'd)

Therefore until recently OCR and page layout analysis experiments were typically conducted on 100s or 1000s of pages (e.g. U.Wash. db)

Information retrieval experiments often require databases larger by more two orders of magnitude (e.g. TREC).

# Trendline



Compression efficiency on scanned documents is approaching the results obtainable by encoding via OCR.

At the same time, increasing processor speed and storage capacity are reducing the need for compression.

This augurs well for DR & R.

THE 2008  
DOCLAB AWARD  
OF EXCELLENCE

FOR SUSTAINED RESEARCH ON

R & R

IS HEREBY CONFERRED ON

The Information Science Research Institute  
team  
of the University of Nevada – Las Vegas

Julie Borsack  
Allen Condit  
Kazem Tahgva  
Thomas Nartker  
...



# Multi- and inter-disciplinary

We *are*, and have always been.

We come from physics, engineering, computer science, communications, library science, biology, remote sensing, cognitive science, web science, ...

This results in wonderful synergies, but the **vocabulary** is daunting!

# Different words for the same concept

% Correct, Error, Reject (cf. C.K. Chow 1970)

Precision, Recall, F1

Type I, Type II

False Alarm, Miss

False Positive, False Negative

Error of Omission, Commission





## Dangerous words

Concept

Context

Model

Semantics

Ontology

Un-, non-, and semi-supervised

Adaptive, self-adaptive

Learning

*Document*

On a personal note ...



I am still dabbling in DIA research on:

Style

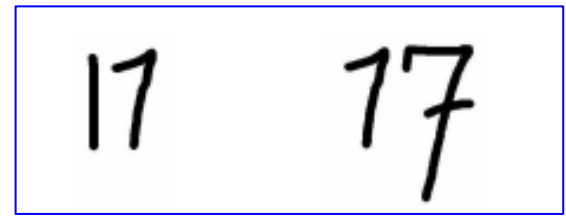
Symbolic Indirect Correlation (*S/C*)

Table Analysis for Semiautomatic

Generation of Ontologies (*TANGO*)

Mark sense ballots (*Cyber Trust*)

# Style



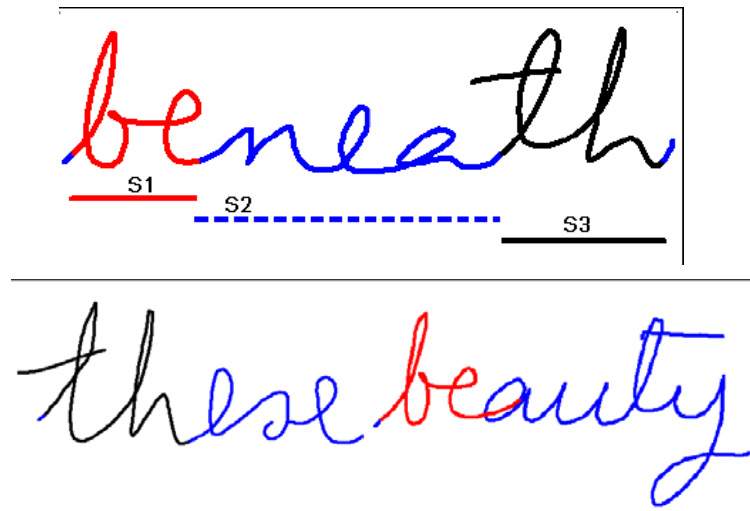
with Prateek Sarkar, Harsha Veeramachaneni,  
Srinivas Andra, and Xiaoli Zhang

The quick red fox jumps over the lazy  
brown dog.

Instead of classifying individual characters,  
we classify *fields* of patterns

# Symbolic Indirect Correlation

With Dan Lopresti, Sharad Seth, and Ashutosh Joshi



**Feature matches between query *beneath* and a 2-word reference string.**

# TANGO

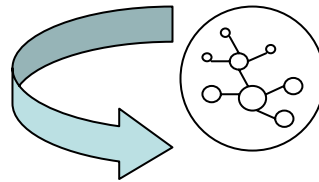
with David Embley, Deryle Lonsdale, and students

repeat:

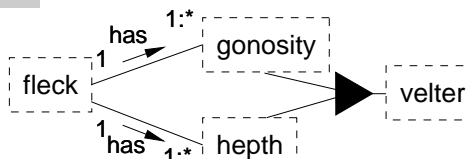
- interpret table
  - discover constraints
  - match with ontology
  - merge
  - adjust
- until ontology developed

fleck	velter	
	gonsity (ld/gg)	hepth (gd)
burlam	1.2	120
falder	2.3	230
multon	2.5	400

fleck	fleck	velter	velter
fleck	fleck	gonsity (ld/gg)	hepth (gd)
fleck	burlam	1.2	120
fleck	falder	2.3	230
fleck	multon	2.5	400



**TANGO in a nutshell: TANGO repeatedly turns raw tables into conceptual mini-ontologies and integrates them into a growing ontology.**



1/27/2008

SPIE/DRR SJ 2007

27

We attempt to combine information from multiple tables in a restricted domain

# DIA for trustworthy elections



with Dan Lopresti,  
Elisa Barney Smith & students

## Objectives:

Create public test data

Determine accuracy of detecting marks as a function of **size**, **location**, **color/contrast**, **noise**, and **shape**

Speed up and reduce bias in (human) audits

OFFICIAL BALLOT		STATE GENERAL ELECTION	
Judge _____		COUNTY NAME	
Judge _____		NOVEMBER 7, 2006	
INSTRUCTIONS TO VOTER			
To vote, completely fill in the oval(s) next to you			
FEDERAL OFFICES		STATE OFFICES	
UNITED STATES SENATOR VOTE FOR ONE		SECRETARY OF STATE VOTE FOR ONE	
<input type="radio"/> CANDIDATE INDEPENDENCE		<input checked="" type="radio"/> CANDIDATE INDEPENDENCE	
<input type="radio"/> CANDIDATE REPUBLICAN		<input checked="" type="radio"/> CANDIDATE REPUBLICAN	
<input checked="" type="radio"/> CANDIDATE DEMOCRATIC-FARMER-LABOR		<input checked="" type="radio"/> CANDIDATE DEMOCRATIC-FARMER-LABOR	
<input type="radio"/> CANDIDATE Party or Principle		<input checked="" type="radio"/> vote "I" only	
<input type="radio"/> vote "I" only		STATE AUDITOR VOTE FOR ONE	
UNITED STATES REPRESENTATIVE DISTRICT (NUMBER) VOTE FOR ONE		<input checked="" type="radio"/> CANDIDATE INDEPENDENCE	
<input checked="" type="radio"/> CANDIDATE INDEPENDENCE		<input checked="" type="radio"/> CANDIDATE REPUBLICAN	
<input type="radio"/> CANDIDATE REPUBLICAN		<input checked="" type="radio"/> CANDIDATE DEMOCRATIC-FARMER-LABOR	

# DRR XV

will surely expand my horizons.

So let us stop this palaver, and get on with  
the important stuff!

*Dear DRR, we wish you  
many happy returns!*



# Bib

I have on my shelves

- II 1994 2181
  - VI 1999 3651
  - VII 2000 3967
  - XI 2004 5296
  - XII 2005 5676
  - XIII 2006 6067
  - XIV 2007 6500
- 
- I'm very much looking forward to your invited talk! :-)
- 
- I just searched my bookshelves and I found the following volumes you are missing:
  - III 1996 2660
  - IV 1997 3027
  - V 1998 3305
  - VIII 2001 4307
  - XI 2004 5296
- 
- I also have duplicates of these volumes you already have:
  - VI 1999 3651
  - VII 2000 3967
  - XII 2005 5676
  - XIII 2006 6067