DRR IS A TEENAGER!

George Nagy
DocLab
Rensselaer Polytechnic Institute
nagy@ecse.rpi.edu
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. Srijari
#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
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
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!
Thank you all!
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).
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

1/27/2008 SPIE/DRR SJ 2007 19
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 (SIC)
Table Analysis for Semiautomatic Generation of Ontologies (TANGO)
Mark sense ballots (Cyber Trust)
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.
repeat:
• interpret table
• discover constraints
• match with ontology
• merge
• adjust
until ontology developed

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

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
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!
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