

## DRR IS A TEENAGER

George Nagy  
DocLab  
Rensselaer Polytechnic Institute  
[nagy@ecse.rpi.edu](mailto:nagy@ecse.rpi.edu)

### Perspective

I am pleased to have this opportunity to comment on the evolution and role of the Document Recognition and Retrieval symposia from the perspective of a member of the OCR and DIA community who has attended more than half of these annual meetings.

DRR has a valuable and comfortable niche among the dozen conferences and workshops devoted to related topics. It attracts industrial and academic participants, researchers and practitioners, and a steady stream of visitors from Asia and Europe. The meetings typically lasts two days, and most of the participants take the opportunity to drop in on some of the other twenty or so IS&T/SPIE Electronic Imaging symposia. Attendance typically runs under one hundred, but many participants return year after year and have come to know each other well. The venue and scheduling are convenient, with ample time for off-line discussion. The printed proceedings are of a manageable size to carry and even read.

The first symposia focused on Optical Character Recognition (OCR), which made its commercial debut at in 1955 at *Readers Digest*. In the intervening decades what is now called Document Image Analysis (DIA) supported OCR with column, paragraph, line and character segmentation, and the detection of non-text regions. Format recognition was limited to relatively consistent families of documents like postal envelopes, social security earnings reports, bank checks and patents. Complete systems included specialized hardware and cost tens or hundreds of thousands of dollars.

I personally date the beginning of DIA in a larger sense to 1984, when I first read the work of Larry Spitz and others. Independent research projects were soon launched on line-drawing recognition, document component classification, and table recognition. Effective form processing methods were introduced a little later. Information retrieval (R) on coded text is of course much older: I recall an IBM project in about 1966 on finding relevant prior art in (keypunched) patents.

### The first symposia

Precursor symposia, like *High-Speed Inspection Architectures, Barcoding, and Character Recognition*, chaired by Michael J.W. Chen in 1991, and *Machine Vision Applications in Character Recognition and Industrial Inspection* (Donald P. D'Amato, Wolf\_Ekkehard Blan, Byron Dom, Sargur N. Srihari, 1992) attracted papers on all three topics. For instance, the 1991 symposium included "Table recognition for automated document entry

system” by Kojima and Akiyama, and “Japanese document recognition and retrieval system using programmable SIMD processor” by Miyahara, Suzuki, Tada and Kawatani.

The next year one of the chairs, Donald D’Amato (a nuclear physicist who became a leading expert on OCR, image processing, and biometrics technologies) co-founded and chaired the first EI symposium devoted exclusively to OCR, titling it *Character Recognition Technologies* (February 1-2, 1993, San Jose, SPIE Vol. 1661). Among other notable papers was contributed was Kopec’s and Chow’s “Document image decoding using Markov source models.”

The chairs – Luc Vincent and Theo Pavlidis – and the organizing committee of the 1993 symposium broadened its scope and title to *Document Recognition*. By then the community became aware of the originality and potential impact of Kopec’s and Chou’s work, and they were invited to present the keynote (titled “Communication theory framework for document recognition”). Document Image Decoding introduced many new concepts in a consistent formal framework, including the notion of a document image as a message encoded for a noisy channel, separation of the model from the recognition engine, the incorporation of Knuth’s side-bearing character model, a 2-D hidden Markov model, and segmentation-free page recognition. Refinement and application of this work continued at PARC even after Gary Kopec’s untimely death in 1998.

The second DRR included other contributions of interest, several with lasting value. Some are listed below in arbitrary order.

- *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
- *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)

Expanded versions of many of these projects were subsequently published in archival journals like IEEE-PAMI, IJDAR, PR, and PRL.

The success of DRR is shared by the central administration of SPIE, the faithful, diligent and inventive conference participants, and the dedicated chairs and proceedings editors. We take pleasure in listing them here:

- I Donald D'Amato
- II Luc Vincent & Theo Pavlidis
- 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

Chairs with industrial or commercial affiliation are listed in sans-serif type. The mixture is fairly even. My apologies for any classification errors. Particular credit is due to Jianying Hu for her long and distinguished service.

## Recognition *and* Retrieval

Sometime during the Lopresti/Hu regime (in 1997?), the title and scope of the conference was expanded to *Document Recognition and Retrieval*. Although recent symposia attracted several papers on information retrieval, there have been few that combined recognition *and* retrieval. We surmise that this can be attributed to the following consideration.

A page of 3500 characters of text scanned at 300 dpi grayscale requires about 10 MB (bi-level it would be about 1MB, which could be compressed by JBIG or DjVu to about 30 KB. The same page, encoded as an ASCII .txt file would be under 4KB, which could be compressed to 1 KB (in MS DOC format it might take 30 KB, and either PDF or RFT about 10 KB). Therefore until recently OCR and page layout analysis experiments were typically conducted on hundreds thousands of pages (e.g. University of Washington database), whereas information retrieval experiments often require databases larger by more two orders of magnitude (e.g. TREC).

However, compression efficiency on scanned documents is approaching the results obtainable by encoding via OCR. At the same time, increasing processor speed and storage capacity is reducing the need for compression. This augurs well for DR & R.

In spite of the difficulties of combining experiments that combine both document image recognition and information retrieval, DRR has been successful in attracting such papers. As an example, we single out the sustained contributions of Julie Borsack, Allen Condit, Kazem Tahgva, Thomas Nartker and their co-authors from the Information Science Research Institute of the University of Nevada in Las Vegas.

## Cross-disciplinary aspects

DRR is, and has always been, multi- and inter-disciplinary. Participants come from physics, engineering, computer science, communications, library science, biology, remote sensing, cognitive science, web science. Much of the interest of the meetings arises from the mix of goals and approaches. This results in wonderful synergies, but the vocabulary is daunting!

We often use different words for the same concept. Consider performance evaluation, on which a fundamental result was published by C.K. Chow in 1970. (He was invited to review the topic at DRR in about 1995.) DRR participants must be ready for any of the following terminology:

*% Correct, Error, Reject;*  
*Precision, Recall, F1;*  
*Type I, Type II;*  
*False Alarm, Miss;*  
*False Positive, False Negative;*  
*Error of Omission, Commission;*

All of them need some elaboration. For example, is “correct” in the first line the fraction of all the data, or of the data that was actually classified, i.e., excluding rejects? While on this topic, we must point out that many researchers still compare classifiers on the same data according to raw error rates. More sensitive tests can be based on analysis of a 2x2 contingency table that reports how many classifiers were correctly classified by both classifiers, how many were misclassified by both, and how many were correctly and incorrectly classified by either.

Sometime the same word is used for different concepts. A short list of the words that set a caution flag for me is: *Concept, Semantics, Ontology, Un-, non-, and semi-supervised Learning, Adaptive, Document.*

## Databases for DRR

In 1994, every DR paper dealt with scanned documents. In 2006, in addition to some excellent papers on OCR (including a splendid keynote by Istvan Marosi!) and ICR (?), we saw camera, multispectral, tablet, and coded symbol data, transcript mapping for Arabic, document CBIR and content inventories, digital publishing and libraries, and even an outlier on shape descriptors.

## Conclusion

DRR has served as a venue for face-to-face interchange for researchers and practitioners interested in OCR and Document Analysis. Researchers new to the area stop in, perhaps while attending one of the other EI meetings. Eventually they might submit a paper, and return year after year.

Under mixed industrial and academic leadership, the focus of the symposium has shifted from OCR and related preprocessing methods to wider issues, including not mostly-text documents. Document sources are no longer restricted to page scanners, but include digital cameras, sensors beyond the visible spectrum, tablet computers, and page composition software. Although there are exciting projects underway to scan the vast amount of existing hardcopy, the most documents already on the web were created in digital form as a result of the exponential growth of information. We therefore expect a further shift in emphasis towards extracting information automatically (or, to begin with semi-automatically or interactively) from the predominant PDF and HTML/XML web document population.

I personally wish many happy returns to DRR, and look forward to future conferences. I am grateful for the opportunity DRR has given me to find about new approaches, results, problems and applications. I treasure the many technical and personal discussions that I have had with participants, and anticipate many more both with the old hands and with new arrivals. Photos of some of the regulars appear below as an exercise in the hot topic of face recognition. Many others are either camera shy, or Google image deliberately ignores them.

