

SDIUT'05, UMIACS, University of Maryland, College Park, MD 20742

Denise Best and Dave Doermann – Conference Coordinators

301-405-6444

<http://lamp.cfar.umd.edu/sdiut05>

Day 3 - Workshop on Evaluation of Document Image Processing Technologies

Issues with Automatic OCR Evaluation and Metrics, Kris Concepcion - Mitre

- How do we determine which OCR engine is the best? – new technology requires more sophistication from OCR engines (need for logo detection, signature matching, noise filters, zoning modules, etc)
- What can we do to improve OCR evaluation?
 - Look at ground truth representation, spatial locations, and noise information
 - Develop an extended OCR output DTD
 - Create a working group to help set the standard for both, to get industry on board to get researchers on the problem

Building Arabic Language Training and Testing Sets, M. Walch - Gannon

- Typically from litigation, machine printed documents, but there are handwritten notes in the margins of these documents that are important
- In case of forensics, a good sample collection: The London Letter
- “The Rabbit Letter” – developed markup system for labeling and ground truth for testing purposes (script-driven interactive software to capture words and individual characters) ~2-3 pgs/hour
- Wanted to deal with structural alterations and word segment challenges

Performance Evaluation of Multilingual Document Exploitation Systems, S.Schlosser - NovoDynamics

- ArborScript ES – document exploitation system
- OCR is critical for being able to bring paper documents into the electronic world
- Capture→ Conversion→ Access→ Web Services
- Need a comprehensive vendor-independent evaluation set to communicate solutions and “real” requirements back to the government

- Would like to see: setting up a test facility (Mitre?) to accept vendor software for evaluation and provide comprehensive report/scorecard and accelerated solution transfer process
- Another ambitious goal: design evaluation data so that it supports the assessment needs of developers and users

Ground Truth Representation Used in Testing and Optimization of the Optical Word Recognition System, M. Ladwig – Northrop Grumman Information Technology

- Unify OWR historical collection for generating ground truth and tuning the algorithm
- Ground truth – slow, expensive, some tools help/others don't

Evaluation with Informational Confidence, Stefan Jaeger and D. Doermann - UMCP

- Presented examples of different types of car dealers and showed how these same examples can be applied to evaluation

Metrics for Word Spotting, S. Srihari - CEDAR, University of Buffalo

- srihara@cedar.buffalo.edu
- Spotting from either a text query or image text (script) query
- How to evaluate the retrieval performance of a system: precision/recall, e-measure (effectiveness) and f-measure (harmonic mean)
- Apply word segmentation (under/over) and word shape matching

Performance Evaluation for Text Processing of Noisy Inputs, D. Lopresti - Lehigh University

- All OCR errors are not created equal during latter-stage processing
- Text processing performance: sentence boundaries, tokenization, PoS tagging (white space affects you here), therefore; worst performer - except for "clean" OCR copy where it performs well
- Using clean, light, dark, fax parameters
- Spurious punctuation is causing sentence boundary errors - can't eliminate periods because you still need to detect sentence boundaries, but could take out commas and that would improve your accuracy rates

VACE Video Text Recognition Evaluation Plans, John Garofolo - NIST

- VACE – video analysis content extraction
- Working in 2D video domain
- Detect text in video and provide transcription of the words, publicly available to users
- Workshop
- Evaluation team: NIST, University of South Florida, Advanced Interfaces (PA), UMCP, and BBN/SRI
- Detection, tracking, and recognition in video frame and sequence
- Excluding text: low-readability (newspaper), dynamic (football scoreboard), scrolling (newscast), stylized (logos)
- English-to-Arabic evaluation scheduled for Feb-Mar 06 with results reported at MLMI-2006 (Rich Transcription Workshop, May 4-5, 2006, East Coast)
- Future: explore object-centric text recognition, semantic text clustering and word ordering, clustering of captions with images and speech, and new domains (e.g., meetings, surveillance).
- Vace-info@nist.gov to participate with workshop, sign agreement, get data, participate in teleconferences, etc.



Advanced Arabic Speech Techniques for Scoring, Rami S. - Sakhr

- Incoming faxes to embassies need to be stored and categorized by theme or concept and sent for translation
- Advanced search engine used for scoring (Idrisi – sits on Internet; native Arabic site)

Evaluation Issues in Image Refiner, Kristen Summers - CACI

- Document image enhancement for OCR – carefully applying transformations (e.g., de-speckling, etc).
- Establish ground truth - have the system learn what marks are important to the script
- Available for English, Arabic, and Thai

The Sporadic Nature of OCR Evaluation, Tapas Kanungo - IBM Almaden

- kanungo@us.ibm.com
- History of OCR evaluation programs:
 - UNLV – evaluation program (COTS research - 5yrs)
 - TREC confusion track –(synthetic corrupted txt -2 yrs)
 - UMD Arabic OCR – (COTS - 2yrs)
 - ARDA ACE – text extraction from video
 - TREC Multimedia – closed captions
- Datasets – try and get SAIC Arabic dataset, ACE, TREC, NIST, SUNY
- OCR \$100M market vs. database and storage systems \$30B market (IBM [\$5B of this] vs. Oracle) and there are councils that evaluate these databases (TPC benchmarks). Model is difficult to pursue in OCR.
- Going forward: A plan for OCR
 - Decide on a few OCR challenges (take 3-5 yrs to solve, intellectually challenging, and have business value), leverage other fields (e.g., linguistics, libraries, etc) and solicit input from government, academia, industry, and open source. Build excitement! (e.g., “Deep Blue”)
 - IBM Content MGMT – extraction of entities
 - Google (Luv Vincent) – extraction of metadata
 - Internet Archive (Brewster Kahle) – extraction of references from scanned articles
 - National Library of Medicine (George Thoma) –Extract journal metadata from scanned articles.

- Resource Repository Suggestions: Ind/lab university web sites, LDC (Linguistic Data Consortium), NIST, Brewster (www.opencontentalliance.org), and Google.
 - Gov (NIST, MITRE, SAIC)
 - Univ (UMD, SUNY, JHU)
 - Non-profit (Open Content Alliance, others?)
 - Commercial (BBN, Google)
- Funding
 - DARPA
 - DoD (Army, Navy, etc)
 - NSF
 - NLM/NIH
 - Commercial/non-profit
- How do you generate excitement? Pull in people?
 - Solution: (Brewster Kahle) have an **X-Prize** for OCR, will need to raise funding from industry, DoD, angel investors, others
 - Could come from another country vs. US
 - Brewster – willing to spend \$10M at OCR to fund data/repository creation, etc. and be provider of repository
 - Google – will help by providing scanned books and repository, compute power
 - Yahoo!, Microsoft, and Amazon – TBD





Henry Baird and Dan Lopresti (Lehigh University) – suggest providing a **short course** in Document Image Understanding (DIU) to learn state-of-the-art, latest algorithms, software tools, database, and performance metrics.

- Lehigh Pattern Recognition Research Lab
 - Bethlehem, PA – 1h from NYC, Philly; 3h from DC.
 - baird@cse.lehigh.edu

PANEL DISCUSSION

Dave Doermann begins the panel discussion by expressing his approval of the idea of having competitions and involving outsiders into the community.

 **Dave Doermann:** Should we ask John Garofolo at NIST to have something TREC-like? Need to pose the right problems, could get the research funding for free. Commercial versus smaller challenge problems. Other types of evaluations anyone?


 **Dan Lopresti:** Should go to DARPA who has the wherewithal to handle these types of challenges. If it's not pushing science forward, you have to be careful what path you're leading down.

Tapas Kanungo: The problem should be selected by the team(s) involved. Come up with a committee that decides on the signs and the problems. We haven't had a grand challenge. We could get some pre-research done.

D. Doermann: What catalyst are we missing? Use the X-Prize money to fund the evaluation. Just have to be careful. ICBAR – handwriting recognition last year got a lot of participation.

Unknown Speaker (Grad Student): I attended FRGC. One of the conditions for funding was the actual performance on the challenge. A good result on a challenge would be a pre-requisite to continue to develop the R&D.

Prem Natarajan: You need a community to present the best technical talent and to give you enough interest/integrity. Need a funding source – DoD. The government needs to fund OCR. Let's get the people who would use it and learn their needs otherwise it's a vain effort. Ideas: medical records, NBC has an archive of Arabic records -fine, have someone read and analyze this. Based on models of success in the government, we should have a grand challenge, but have the right people in the room first.

 **Charles Wayne:** What would be a good grand challenge? The government's needs aren't that much different from that of commercial companies. It could be for a prize or for data research. There are other applications besides those of the government. Pick a small number of appropriate challenges, name them, and work on them for a number of years. What are those central challenges you'd be working on?

D. Lopresti: It's always from the perspective of the person with the money. If you went to Google (unconstrained handwriting recognition nah, he'd probably say scanning of

books is what's important). We can say what's solvable in a certain amount of time. Measure of success is critical.

T. Kanungo: All funders produce great science? Not necessarily.

John Garofolo puts up his slide entitled: "NIST Benchmark Test History May '05."

J. Garofolo: Need to plot your history in order to show the way of your funding. Good to show your potential vendors/participants your progress to date. Someone needs to enunciate the goals and to show progress. UNLV's funding got cut in 1993. You need a community to support evaluations.

P. Natarajan: Robust, retargetable focus. Would that be sufficient? Complex images/camera images, are these algorithms retargetable, languages retrainable, etc. But what's the intended use? What are the set of tasks?

Stefan Jaeger: Problem is they don't find it sexy? Character recognition is still sexy because we don't know how to do it. Deep Blue may have been a good project, but it still didn't solve the problem.

T. Kanungo: Computer scan vs. human – how do they compare? Can you beat the human? The point of Deep Blue was to show how the computer beat the human, not whether it did.

Henry Baird: We have people who have built fast OCR systems. When compared to the government and broader uses in the community, they're getting better all the time. Don't focus on doing sign better on another nice program but on trainable system that can be trained on anything. Extreme breadth and versatility and also adaptation.

Barry Budish: There's a major sexy problem – Arabic. There's no uniform way to compare performers against one another. There are different government agencies in here using different tools for different things. One's using one Arabic OCR, the other's using XYZ. Is the grand challenge the only way? Schlosser gave an idea of a single test site -- who's going to do it? Is it a DARPA, MITRE, etc. Government needs to say what they need in terms of uniformity and make it comparable.

Pretar: How do you define a grand challenge problem in the OCR, DI, etc....requires dependence on a particular dataset so that at the end you can determine your success. Community may have to disband and come together again in five (5) years.

Betsy McGrath: Would like to see a focus in handwriting segmentation related to the margins. Is this perhaps more important than the document itself? Is there a button to push rather than having people to read it to capture this important data? I might care about the logo or company name...things like this, concrete items, easily identified but not necessarily easy to solve. (Not appropriate for a grand challenge).

P. Natarajan: Let's assume that it is a challenge candidate. The point, we don't even know enough to define to determine what constitutes a grand challenge. We have mountains of data that needs to be processed but it's very hard to widdle down to two to three (2-3) problems. Is it keyword search? Is it separation? Therefore, if you have a community to discuss a grand challenge, what are the stakes in the ground, what are the areas? Simply bring the people with access to problems to table. (Dave – that's what this [SDIUT] is accomplishing.)

C. Wayne: Users need to think more broadly because they're so focused on their own fields.

Cathy Ball: The Grand Challenge is doing that for seven (7) million documents in a couple of days.

T. Kanungo: DARPA three (3) day workshop 1992 – people broke out into groups and took different topics to look at important topics. Why not re-enact that? Take 5% of budget and put it to free for all...what's the harm in that?

C. Wayne: The risk is here's the 5% and take away the 95%

P. Natarajan: We need the government to tell us what they want. It's competitively driven. ScanSoft no longer doing handwriting recognition research because they lost funding. The government is not funding Arabic handwriting – the money ebbs and flows.

D. Doermann: Hopes the government talks and communicates with us. It's got to be a five (5) year program versus six (6) months.