BOBCAT UMD Work plan

December 26th, 2007
(Updated January 18, 2008)

Project Objectives:

To produce:

- Documented methods and procedures, automated where feasible, to measure OCR accuracy; to assess tools to find logos, signatures, and key words in Arabic documents; to assess tools for language ID, to assess tools to separate handwritten documents/page regions from machine printed documents/page regions for automatic routing: to human (handwriting) vs. algorithm (machine print), or to appropriate recognition algorithms (high vs. low-noise processes).
- A tool to facilitate image groundtruthing.
- Documented test designs, data analysis procedures, and interpretation guidelines for specific assessments conducted in BOBCAT-DI in FY07-08.

Work plan:

A number of specific objectives will be carried out to provide the tools necessary to integrate into the evaluation test bed

1) Survey of available datasets, ground truth, evaluation metrics and competitions
    a. Delivery: Feb 15, 2008
    b. Overview: We will provide a site for the search and cataloging of existing methods and data for evaluation. Users will be able to upload information about existing datasets and search for data and contact information for publicly reference datasets.
2) Ground Truth Dataset - Arabic
    a. Delivery April 30, 2008 - Tobacco, June 30, 2008 - Arabic
    b. Overview: We will provide the 1000 zone level ground truthed document images form the USF Tobacco litigation corpus for testing zone identification, logo detection and signature verification. We will provide 500 ground truthed document images from the Anfal Collection (including text lines, logos and signatures).
3) GEDI Enhancements
    a. Delivery May 31, 2008
    b. Overview: We will enhance existing tools to incorporate evaluation metrics and visualization. The GEDI tool will add capabilities for ground truthing handwritten data including polygons shapes, and baseline representations.
4) Metrics for Evaluation of Data on Handwritten Documents
    a. Delivery: May 31, 2008 (draft), End of Project (final)
    b. Overview: After surveying the information in 1) above, and discussing needs with sponsor and the MADCAT project, we will provide a set of recommendations for evaluation of government data.
5) Evaluation Survey
    a. Delivery: August 31, 2008
    b. Overview: A journal survey will survey evaluation protocols, methods and techniques. To be submitted for publication
6) Prototype Portal
    a. Requires more discussion

Progress Through December 26, 2008

1) Began survey of datasets and metrics
2) Hired student to build online portal.
3) Began implementing online portal for data search and presentation to the community
4) Contacted previous evaluation organizers including NIST, and ICDAR to inquire about possible

Related Notes:

During SPIE this year, David Doermann will visit the Google Books project, and report on the availability of additional data, and metrics for handwriting.