Statement of Work

Bobcat-DI Phase II

Background

Phase I of Bobcat-DI focused on developing the capability to assess the accuracy of OCR and other document image processing tools for the US Government. Although a number of independent evaluations have been run by various academic organizations, all focus on slightly different problems. In recent years, the University of Maryland has developed a number of tools aimed at supporting generic annotation and evaluation of document (and video) data. These image processing and analysis metrics and methods are needed to enable assessments that are reliable, robust, and scientifically defensible. To this end, the project

1. developed the PETS tool kit for the evaluation of page segmentation and classification,
2. provided modifications to the GEDI toolkit to include annotation of pixel level data, handwriting and irregular shaped regions
3. annotated and delivered a wide variety of datasets to test the evaluation framework, and
4. performed a set of trial evaluations on other algorithms developed under related efforts at the University.

A series of software deliverables and reports were products of this effort.

As part of the Phase I work, a number of trends emerged that while previously known, have not been well documented. In particular, it was observed that many document analysis programs such as line detection, page segmentation and classification work well on clean documents, when documents are complex or noisy, these algorithms perform sub-optimally.

Proposed Work

Phase II work we will continue efforts in document analysis related to the processing of low quality and complex documents which specifically address the problems of document triage. The primary goal will be to enhance the current infrastructure and test these capabilities using PETS. We have identified a number of capabilities that are essential to the governments current workflow and seek to fully implement existing prototype algorithms for integration into the existing ARL evaluation architecture. In particular we will implement and delivery code for:

Noise and Clutter Detection and Removal:

> *Existing Matlab code will reimplemented in Doclib and enhanced to add capabilities for salt and paper and background noise. The resulting programs will be evaluated using the previously developed PETS software and reports will be generated.*

Rule Line Removal:

Software being developed as part of the MADCAT project will be tested using the PETS framework and delivered as part of DocLib

Enhanced Page Segmentation:

The UMD Voronoi++ algorithm is currently being enhanced and will be delivered as part of DocLib and tested using PETs.

Zone Classification:

The current UMD zone classification software will be integrated into DocLib and extended to be trainable on Government data. The software will work in conjunction with the Enhanced page Segmentation defined above.

All of these modules will be delivered along with source code.