Best of Breed Configurable Active Testbed – Document Image (BOBCAT – DI) Project

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

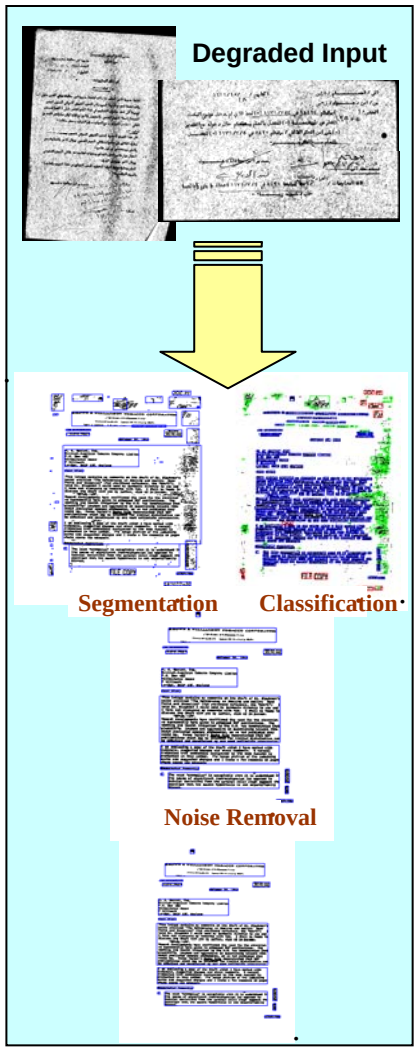
Luis Hernandez
Army Research Laboratory

11 December 2008

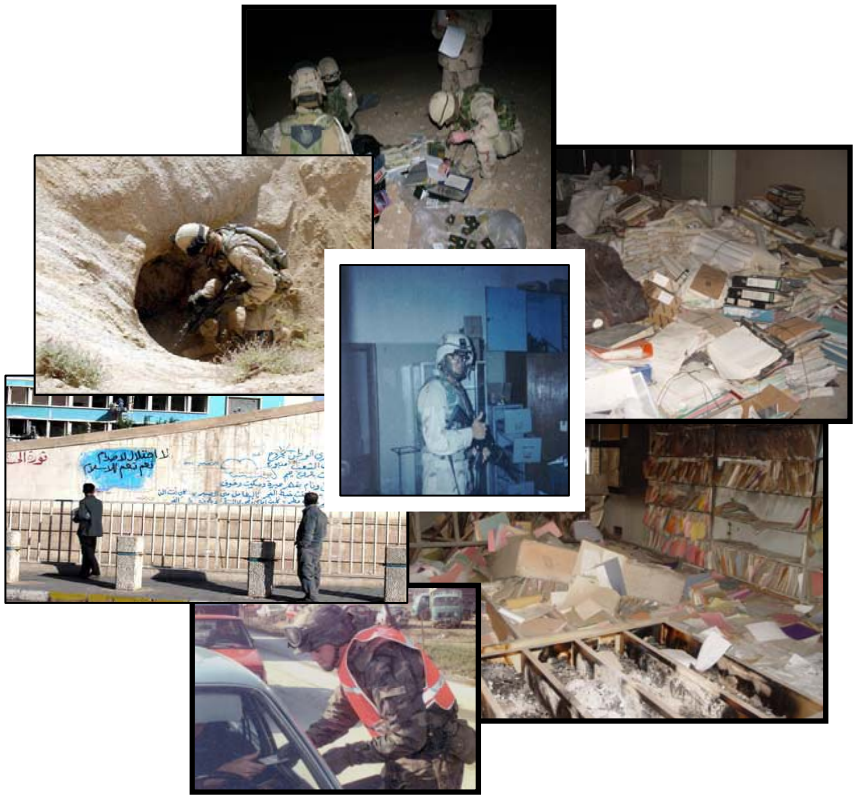
Outline

- Background
- Goals
- Objectives
- Project Accomplishments
  - Datasets
  - Tools
  - Evaluations
- Future Plans

The Challenge



Workflow and associated process



Hand-Written or Degraded Document Images into machine readable documents

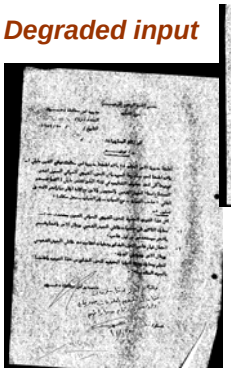
Military Applications

Document Image Processing, Analysis and Optical Character/Handwriting Recognition

Research and Experimentation

- Image Noise Removal and Enhancement
- Document Image Analysis
- Machine printed and hand-written character recognition
- Image to text services for Machine Translation and retrieval processes

Degraded input



Military Applications

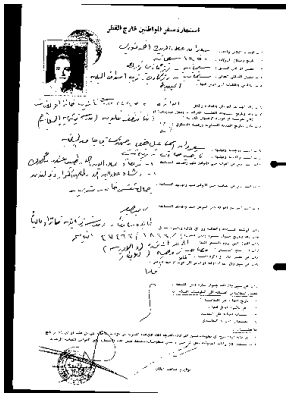
Critical use of MT

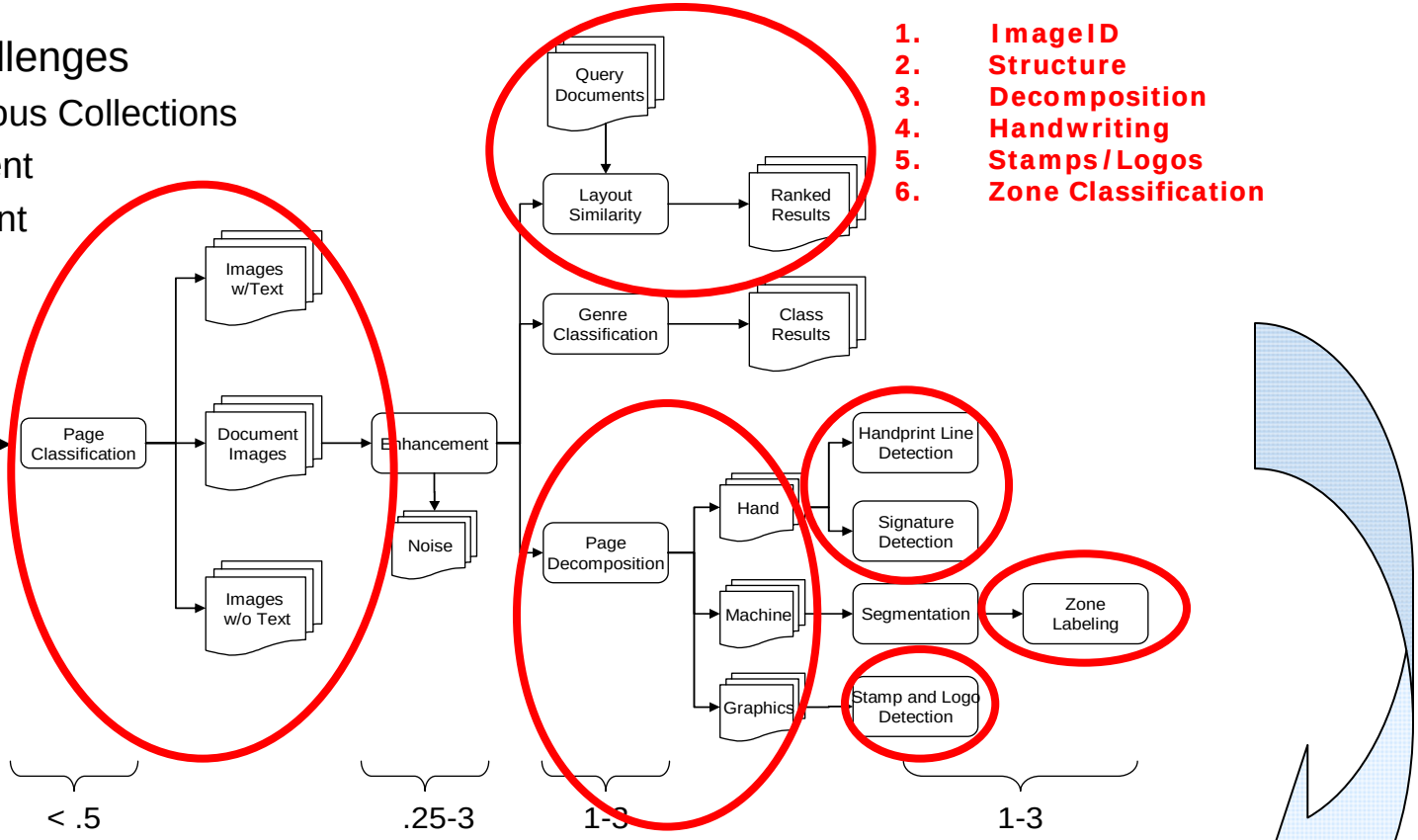
Ultimate Goal of Systems

Fundamental Challenges

- Heterogeneous Collections
- Mixed Content
- Noisy Content

1. ImageID
2. Structure
3. Decomposition
4. Handwriting
5. Stamps/Logos
6. Zone Classification





Target Processing Speed in Seconds

Vision

Pursue intelligent digital capture, segmentation, recognition and processing of documents to enable fused, timely information from all relevant document sources to the Warfighter.

Evaluation

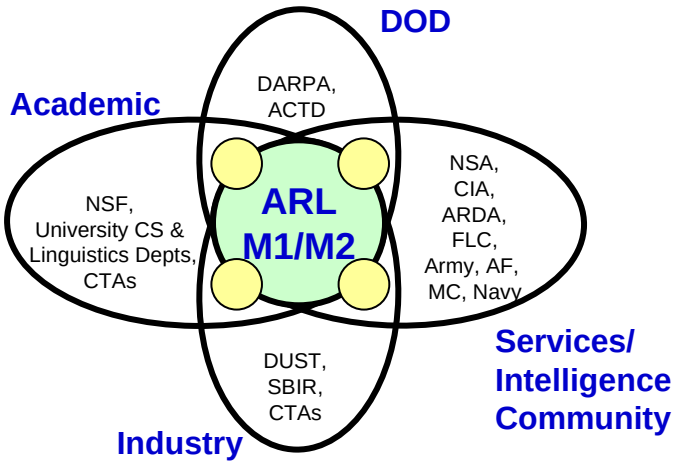
- How to evaluate such systems in various phases, in order to:
  - Evaluate the state of the art (comparison of systems)
  - Measure progress during development of new algorithms
  - Measure progress against operational needs
- Each is important
  - Technological progress in the field
  - Technology insertion for initial capabilities

Bobcat-DI: Project Goals

- Provide for a Web Enabled Government environment for:
  - evaluating COTS and GOTS OCR, Hand Writing Recognition systems and associated digital image processing technologies.
  - investigating the effects of character recognition accuracy on end-to-end accuracy of embedded MT systems
  - examination of downstream effects of image categorization and routing on MT output and analyst's extraction of metadata

- Use proven metric techniques
  - Character Error Rate
  - Character Accuracy Rates
  - Word Error Rate
  - Precision and Recall
  - Others

Collaborations



ARL collaborates heavily w/many agencies in IC and across services

BOBCAT - DI: Project Objectives

- Layout the framework for a document image testing environment
- Provide tools … to extend ground-truth data collections to include Arabic Anfal images
- Develop and transition the test methods, metrics, and procedures … as part of an assessment infrastructure
- Provide test designs, data analysis procedures, and interpretation guidelines for evaluating COTS and GOTS OCR systems and other DIP tools

BOBCAT - DI: Tasks

- Data Sets
  - Zone Classification and Segmentation GT
  - Character/Word level GT
- Tools
  - Modify UMD's Groundtruthing Editor Document Interface (GEDI) to allow handwritten data representation
  - Develop DocLib Extensions/add-on routines
  - Extend ARL Image and OCR Toolkit (IOTK) UI
- Evaluation
  - Conduct Pilot Segmentation evaluations
  - Conduct Pilot Zone Classification evaluations

BOBCAT – DI: Datasets

Overview

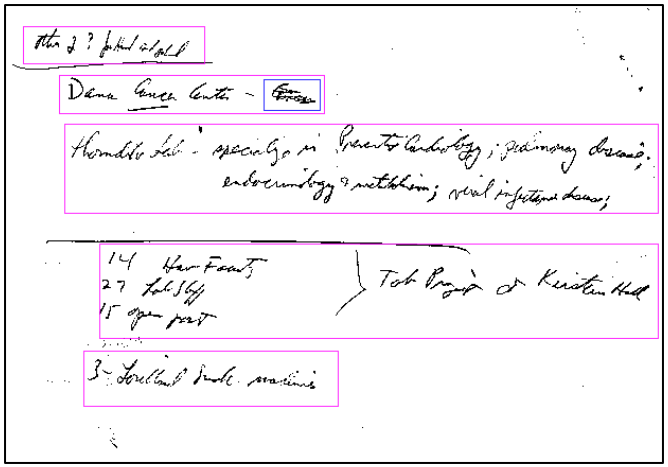
- Many datasets exist as simple collections of images
- Most to not accurately reflect the challenges faced by government organizations
- There is a significant need for datasets that can easily be evaluated across application – OCR, Page Segmentation, Classification, Indexing and Retrieval, etc.
- UMD/ARL is establishing a common format usable in GEDI and IOTK for many tools and for exchange between algorithms.

BOBCAT – DI: Data Sets

Name	Number of Images	Level of GT	Applications	Comments
AMA-Arabic 1.2	5000	GEDI - Word + PAWs	OCR	Arabic Handwritten
UMD/AMA-Zones	> 25,000	Zones + Type	Segmentation - Classification	Subset of Tobacco
Tobacco Clutter	7.9million	None	Segmentation - Enhancement	English, Some Foreign Language
MADCAT – Dev, Train, Eval Set	~10000 Total	MADCAT - Word/Line	OCR	Arabic Handwritten
Anfal Box 50-53 Subset	> 7000	None	Segmentation- Enhancement	Arabic Handwritten and Machine Printed
LAMP - Syn Lines	50	Implicit in base images	Line Removal Line Detection	Arabic Handwritten and Machine Printed
UWASH	1600	Zones + Type	Segmentation - Classification	English Only

Sample GT from Data Sets

- Segmentation/Classification
  - 26,007 pages of Tobacco Litigation Corpus
  - 320,000 + zones
  - Useful for Large Evaluations



Anfal GT set of
- Handwritten
- Machine Printed
- Graphics

BOBCAT –DI: Tools

GEDI – Ground Truth Annotation

IOTK – UI, Workflow and Evaluation process management

DOD Scoring Software – Accuracy evaluation of OCR text

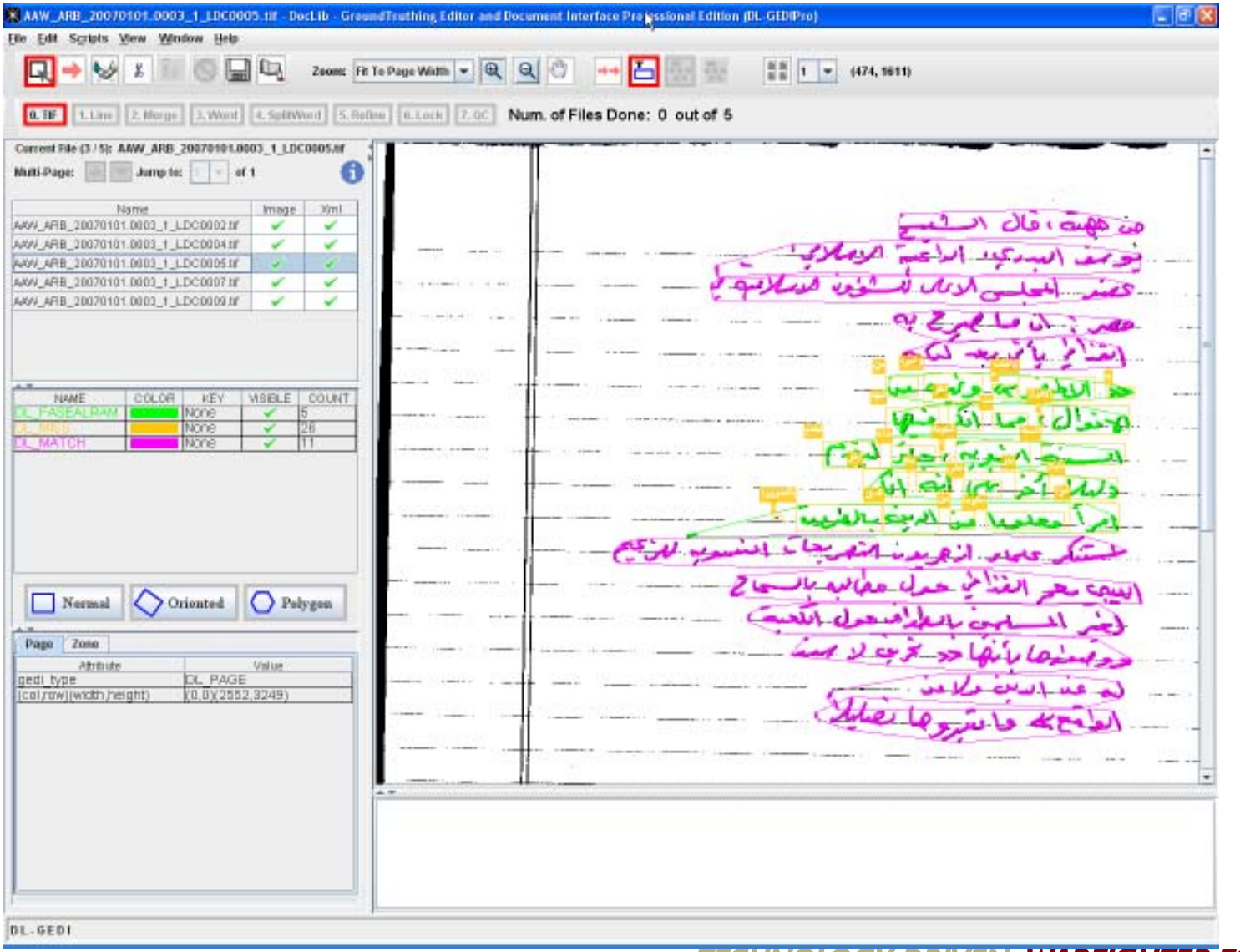
PETS – Performance Evaluation of Layout based Algorithms (Page Segmentation, Zone Classification)

UMD GEDI Tool

- Generic Tool for representing regions and attributes on images

- Bobcat -DI Project Extensions
  - Polygons for complex layouts
  - Reading order
  - Representation of run length encoded data for line segmentation
  - Direct integration of evaluation capabilities via scripts
  - Results visualization

UMD GEDI Tool Interface



Example XML Format Output

```
<?xml version="1.0" encoding="UTF-8"?>
<!--GEDI was developed at Language and Media Processing
Laboratory, University of Maryland.-->
<GEDI xmlns="http://lamp.cfar.umd.edu/GEDI" version="1.0">
   <USER name="Elena" date="5/23/2008 17:24"
dateFormat="mm/dd/yyyy hh:mm">    </USER>
   <USER name="Orri" date="6/11/2008 12:52"
dateFormat="mm/dd/yyyy hh:mm">    </USER>
   <DL_DOCUMENT src="sample.tif" docTag="xml"
NrOfPages="3">
      <DL_PAGE gedi_type="DL_PAGE" src="sample.tif"
pageID="1" width="1728" height="2292">
         <DL_ZONE gedi_type="DL_TEXTLINEGT" id="2"
col="1285" row="269" width="166" height="335"
                        orientationD="16.169" contents=""
offsets="" segmentation="word">
         </DL_ZONE>
      </DL_PAGE>
      <DL_PAGE gedi_type="DL_PAGE" src="sample.tif"
pageID="2" width="2592" height="3300">
      </DL_PAGE>
      <DL_PAGE gedi_type="DL_PAGE" src="sample.tif"
pageID="3" width="2592" height="3300">
      </DL_PAGE>
   </DL_DOCUMENT>
</GEDI>
```

Statistics

Distribution of Zone Categories in AMA-Zones Dataset

Category	Documents		Zone Type	Count
advertisement	451		FORM	3,679
bibliography	158		GRAPHICS	3,430
calendar	44		HANDPRINT	50,138
drawings	597		Image	1,484
email	962		LOGO	4,070
fax	815		MACHINEPRINT	210,696
foreign	761		MARKUP	27,533
form	1,407		SIGNATURE	5,552
graphic	518		STAMP	5,074
handwritten	2,766		TABLE	5,559
letter	2,561		TITLE	5,800
list	395			
marginalia	888		**Total**	**323,015**
memo	1,893			
newspaper	615			
periodical	22			
photograph	227			
questionnaire	188			
report	985			
tables	690			
Total Documents	**16,943**			
Page Count	**26,007**			

Image and OCR Tool Kit (IOTK)

Image and OCR UI, workflow and process framework, and data store in a toolkit paradigm

Design Document - In Progress

Database Design

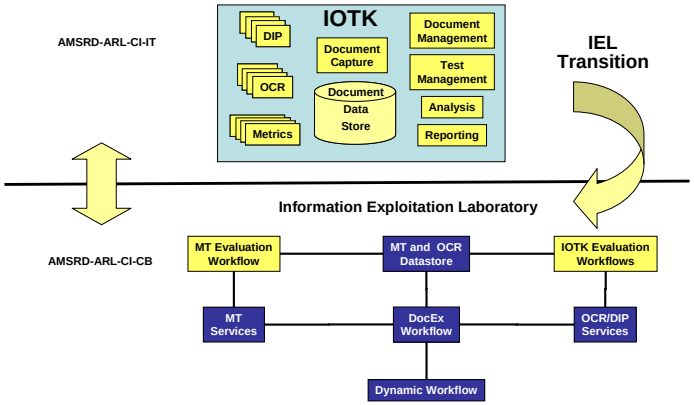
- Initial design completed
- Initial design refinement ongoing in collaboration with IEL staff
- Planned extentions for DIP/ Segmentation data needs

Language/API Cross Reference Documentation - Ongoing

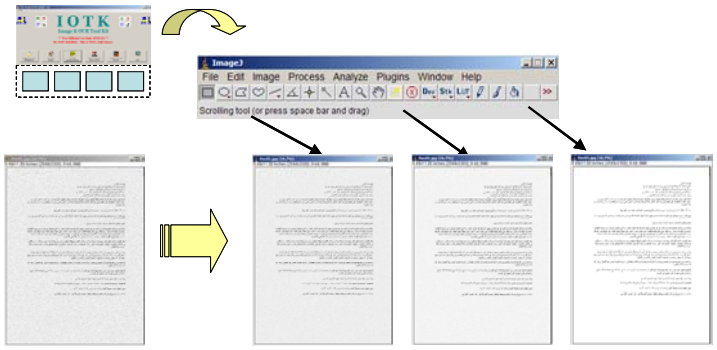
User Documentation – Not Started

IOTK Tutorial (How Do I?...) – Not Started

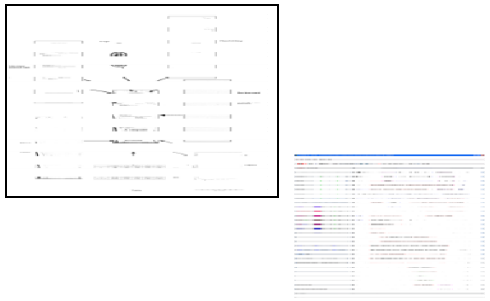
IOTK collaborations to extend capabilities as web services



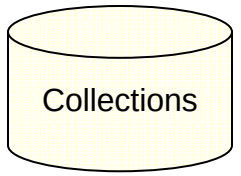
XMAL,WPF,XBAP, Expression Blend



Digital Image Processing Tools for Cleaning/Degrading Images

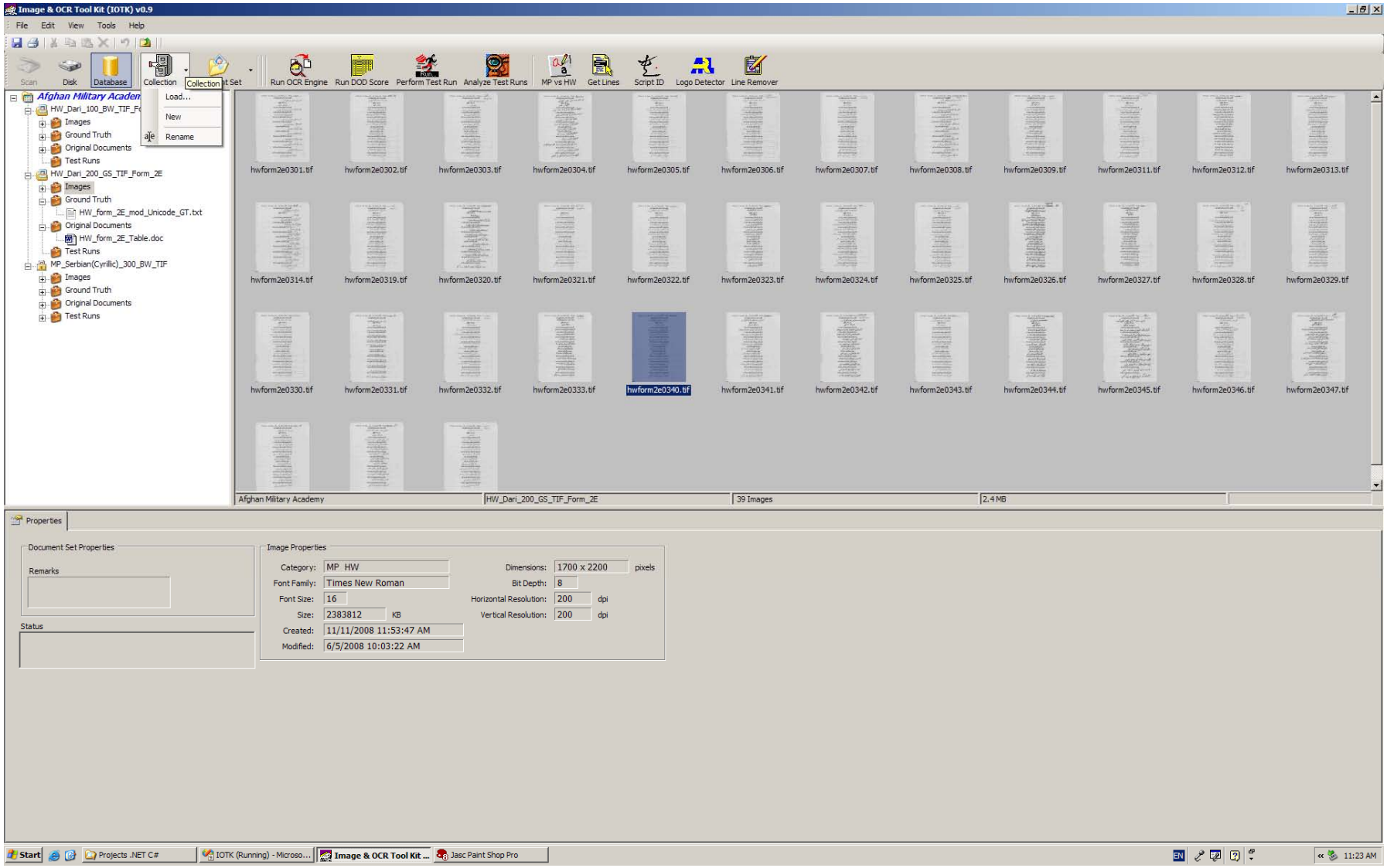


Segmentation GT Generation, Metrics and Evaluation

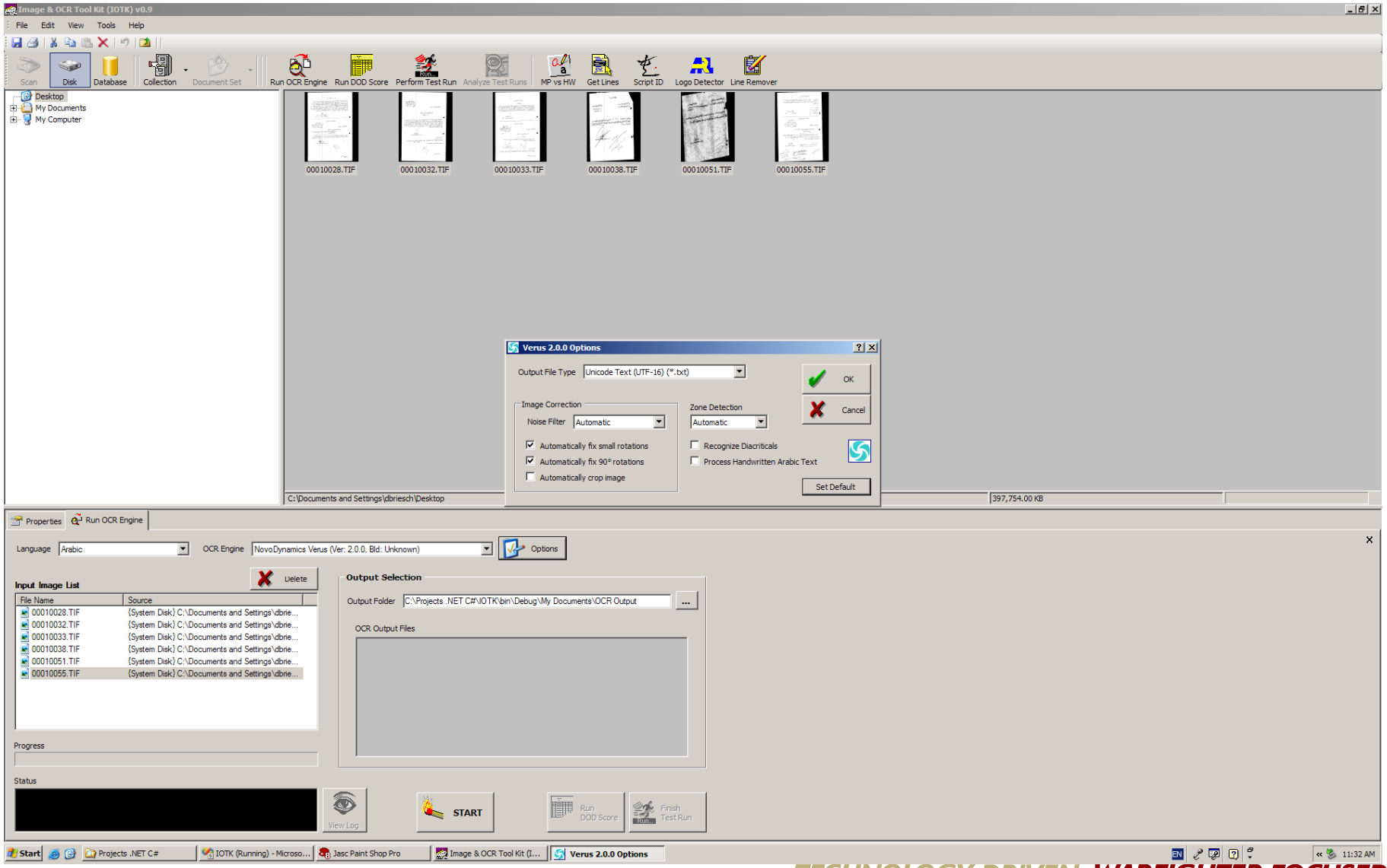


GT Document Image Data Sets

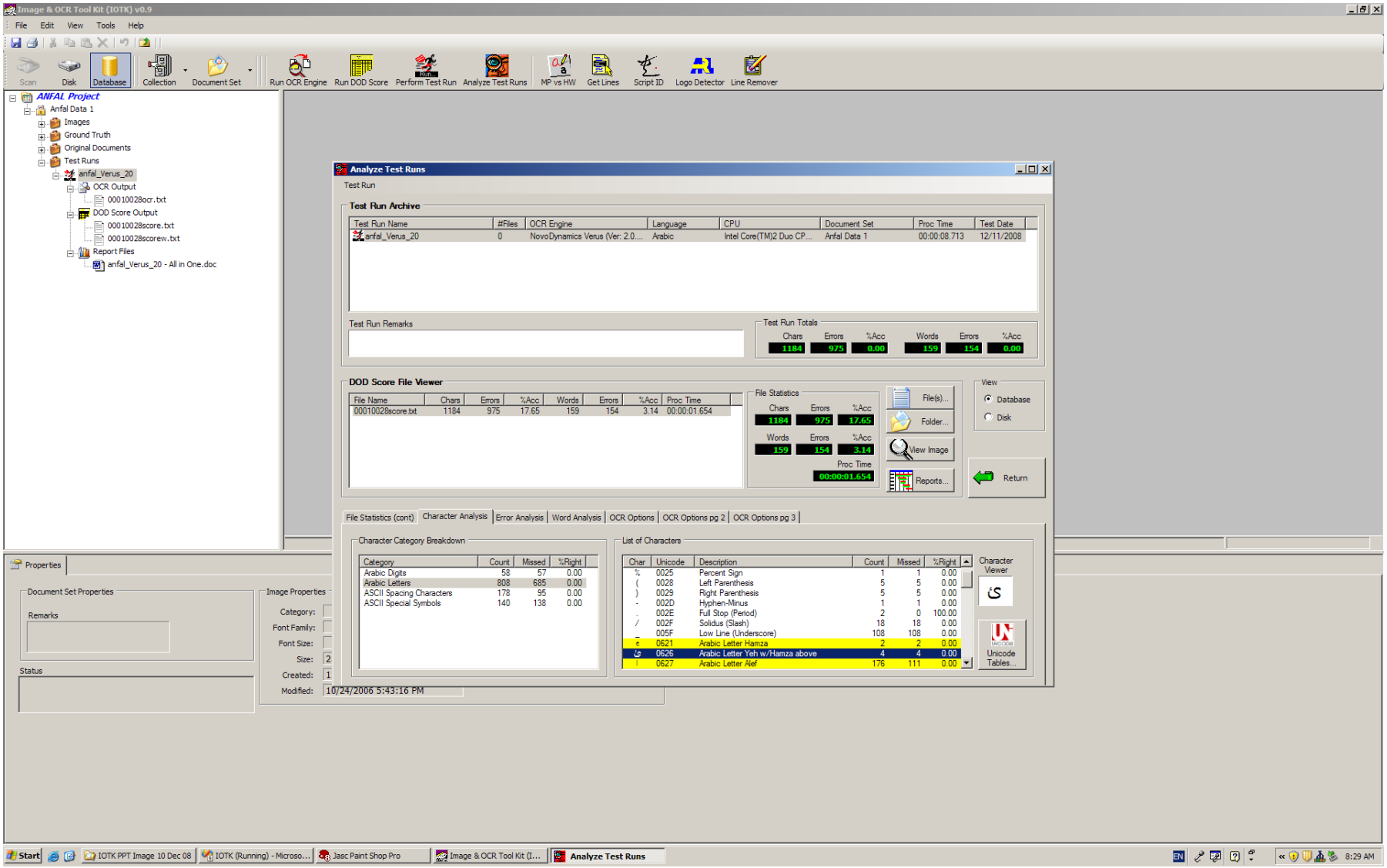
IOTK User Interface



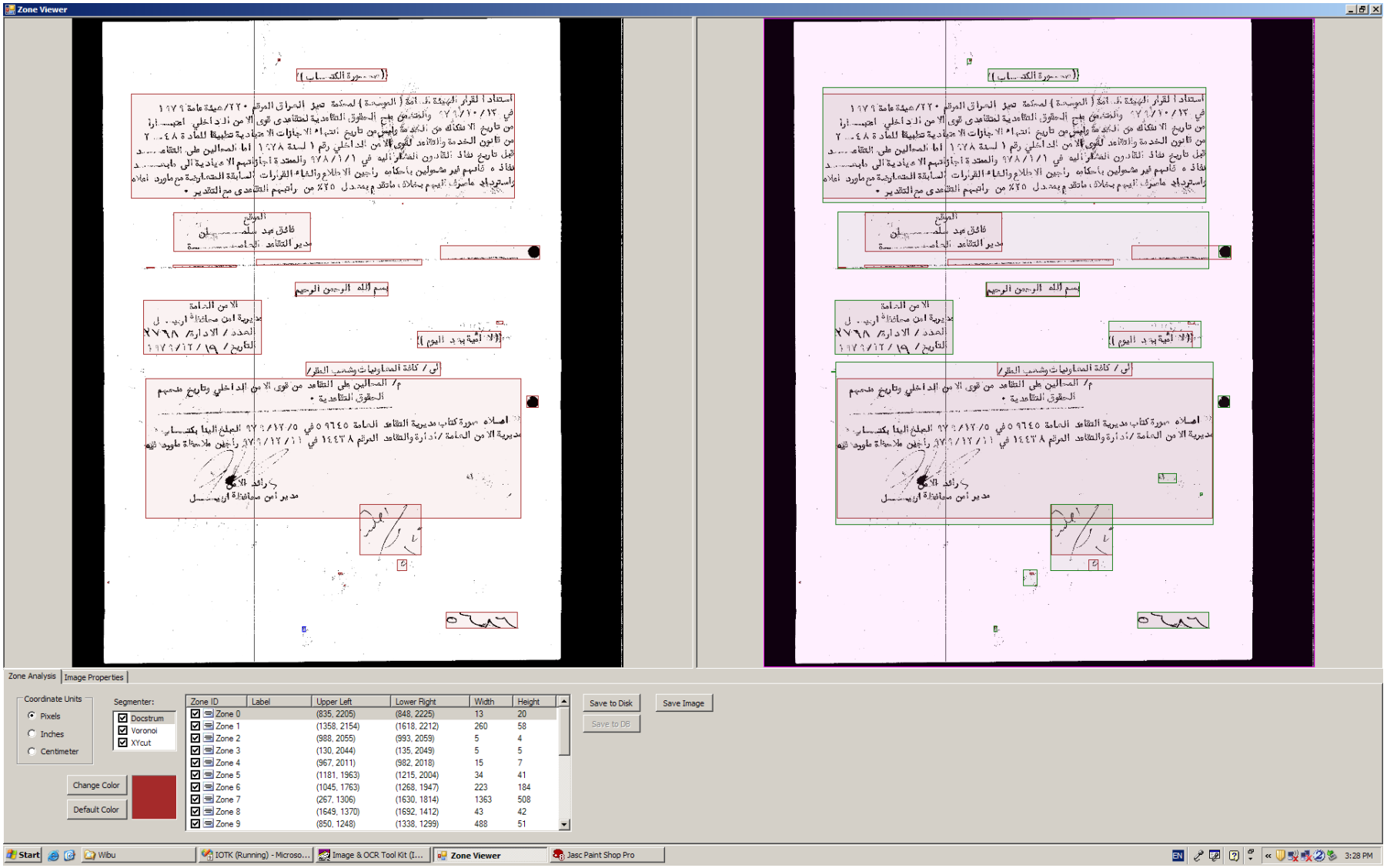
IOTK OCR Module



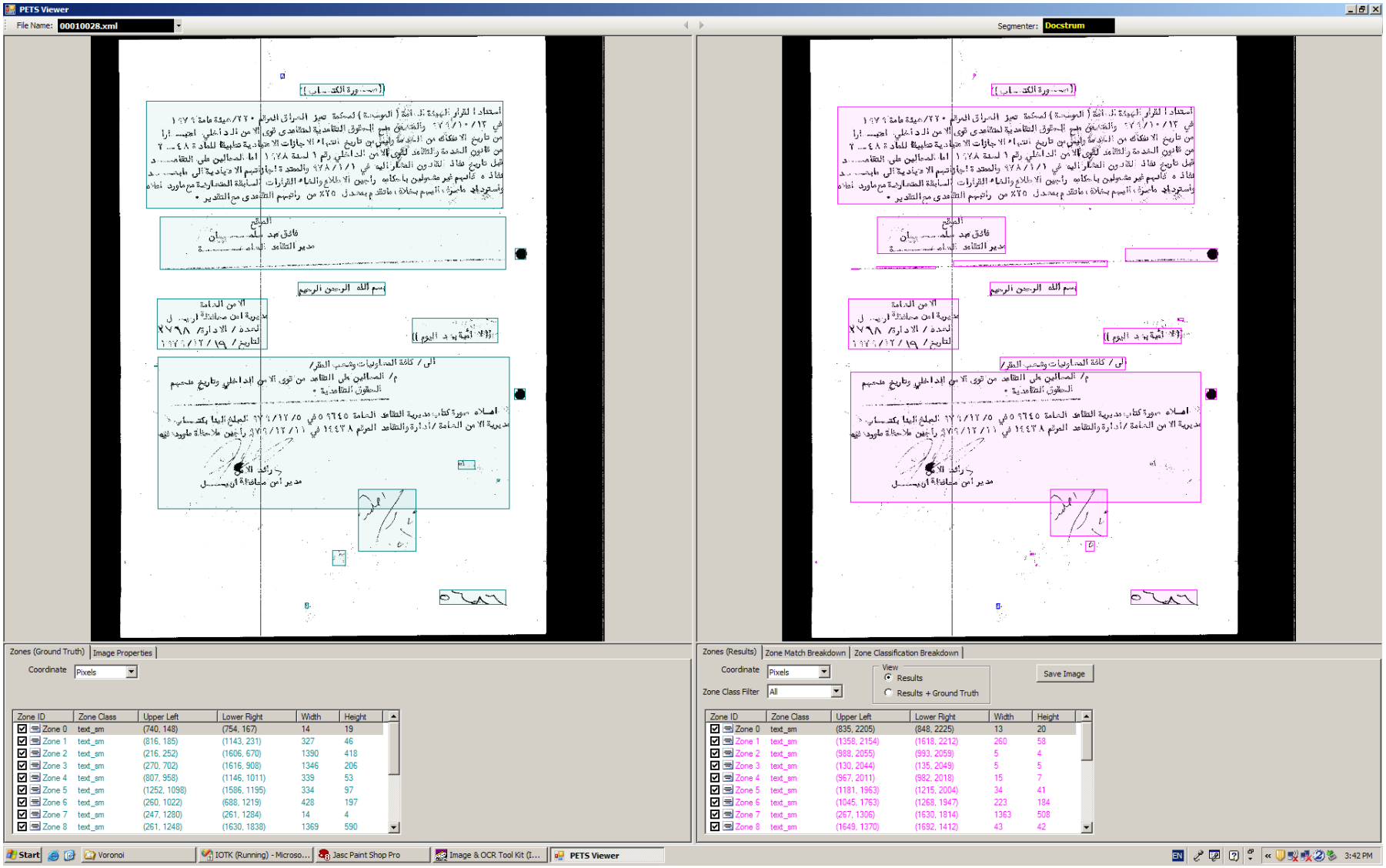
IOTK OCR Accuracy Module



IOTK Segmentation Module



IOTK PETS Module



Performance Evaluation Tool for Segmentation (PETS)

General Concept:

- Given two zones to be compared, calculate the matching score if there is at least one shared ON pixel
- Four types of result
  - MATCHED: location and zone type
  - DETECTED: location but not zone type
  - FALSE: Extra zone in Results
  - MISSED: Zone not matched from GT
- Threshold is set to determine which zones are matched for “detection”
- Full match matrix is built to store the score of each pair of zones.
- Software follows DocLib design paradigm.
- Will be provided to DocLib as an Add-On component

Matching Score and Result Types

I = set of all ON pixel in Image

R_i = set of all ON pixel in the result zone

G_j = set of all ON pixel in the ground truth zone

T(s) = function that count the elements of set s

$$MatchScore(i,j) = \frac{T(G_j \cap R_i \cap I)}{T((G_j \cap R_i) \cap I)} \times 100$$

MATCHED

MatchScore(i,j) ≥ threshold
L(i) = L(j)

DETECTED

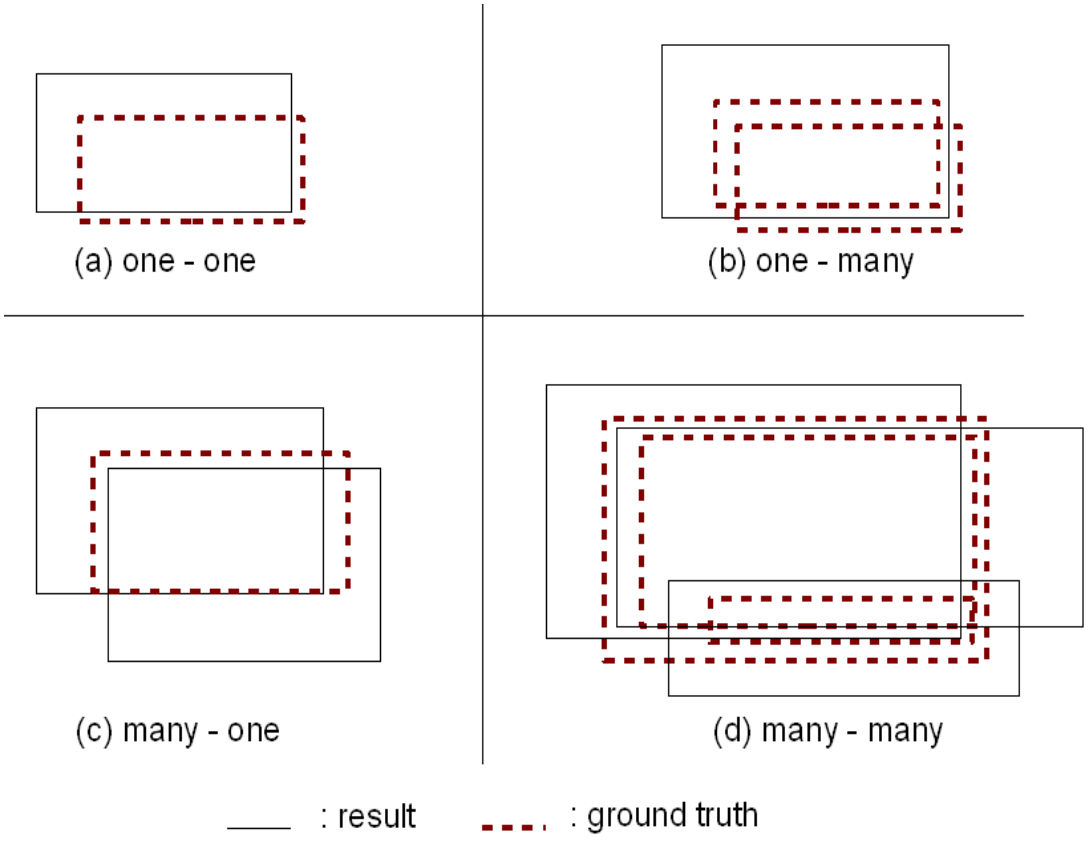
MatchScore(i,j) ≥ threshold
L(i) ≠ L(j)

FALSE

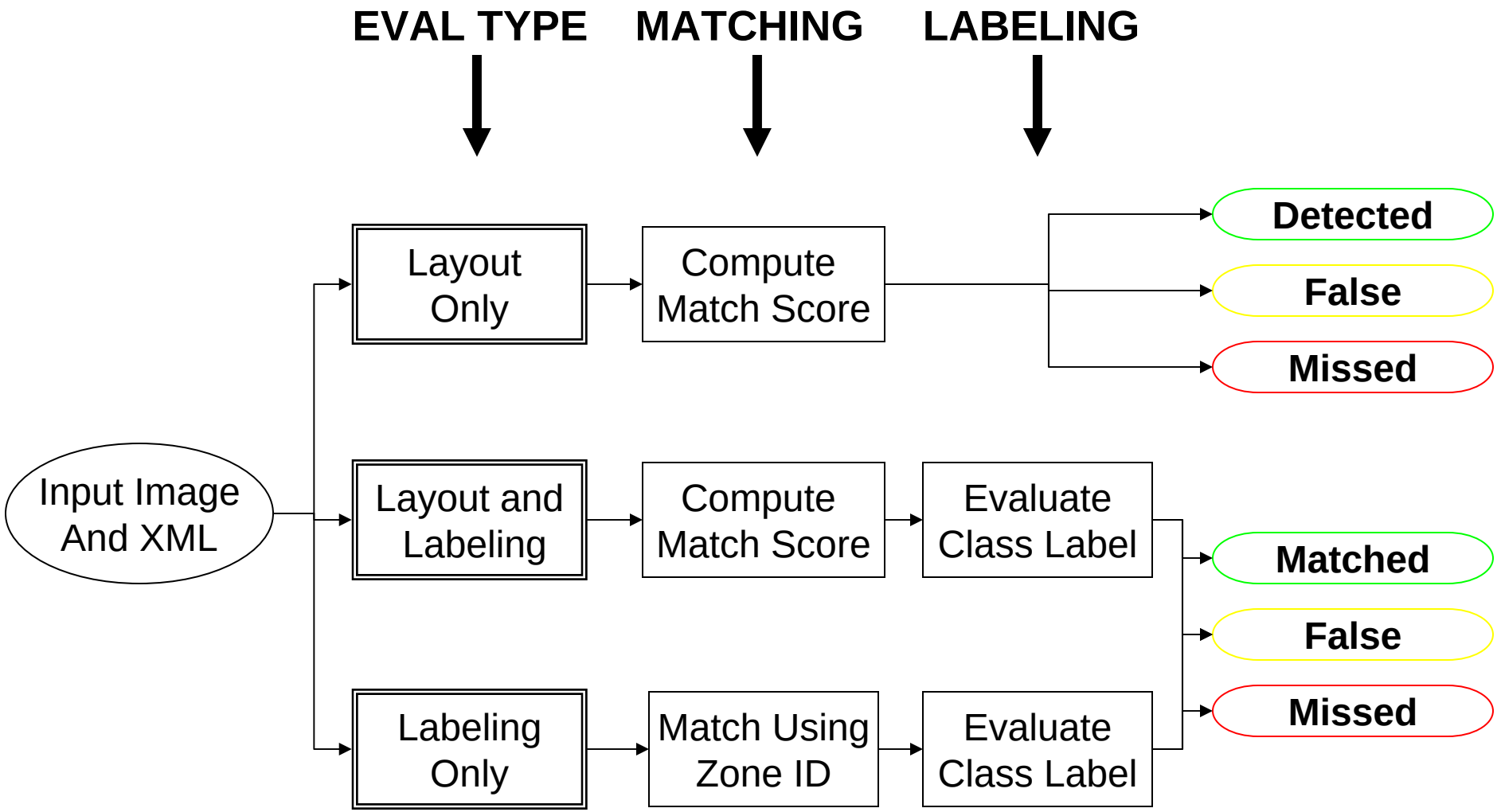
MatchScore(i,all) < threshold

MISSED

MatchScore(all,j) < threshold



Evaluation Framework



Segmentation and Classification

==========
Summary of Results
==========

- Total Number of Sample : 21786
- Overall Accuracy : 95.78%
- Average of Each Class Accuracy : 55.31%

01. Information on Classes
==========

Label	Name of Class	Number of Sample	Accuracy
00	text_sm	20617	97.34%
01	ruling	201	61.69%
02	drawing	299	88.29%
03	table	76	46.05%
04	text_lg	51	64.71%
05	math	301	60.47%
06	halftone	144	83.33%
07	logo	13	0.00%
08	chm_drawing	80	51.25%
09	map	4	0.00%

Segmentation and Classification

02. Confusion Matrix
==========

Out\GT	00	01	02	03	04
00	20068(97.3%)*	70(34.8%)	11(3.7%)	14(18.4%)	12(23.5%)
01	69(0.3%)	124(61.7%)*	0(0.0%)	1(1.3%)	1(2.0%)
02	93(0.5%)	1(0.5%)	264(88.3%)*	23(30.3%)	4(7.8%)
03	46(0.2%)	0(0.0%)	5(1.7%)	35(46.1%)*	0(0.0%)
04	19(0.1%)	1(0.5%)	0(0.0%)	0(0.0%)	33(64.7%)*
05	284(1.4%)	2(1.0%)	8(2.7%)	2(2.6%)	1(2.0%)
06	38(0.2%)	3(1.5%)	6(2.0%)	0(0.0%)	0(0.0%)
07	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)
08	0(0.0%)	0(0.0%)	5(1.7%)	1(1.3%)	0(0.0%)
09	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)

05	06	07	08	09
106(35.2%)	5(3.5%)	7(53.8%)	0(0.0%)	0(0.0%)
0(0.0%)	0(0.0%)	1(7.7%)	0(0.0%)	0(0.0%)
9(3.0%)	18(12.5%)	0(0.0%)	9(11.3%)	4(100%)
0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)
0(0.0%)	0(0.0%)	4(30.8%)	0(0.0%)	0(0.0%)
182(60.5%)*	0(0.0%)	0(0.0%)	30(37.5%)	0(0.0%)
0(0.0%)	120(83.3%)*	0(0.0%)	0(0.0%)	0(0.0%)
0(0.0%)	0(0.0%)	0(0.0%)*	0(0.0%)	0(0.0%)
4(1.3%)	1(0.7%)	1(7.7%)	41(51.2%)*	0(0.0%)
0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)*

Segmentation and Classification

03. Precision and Recall
==========

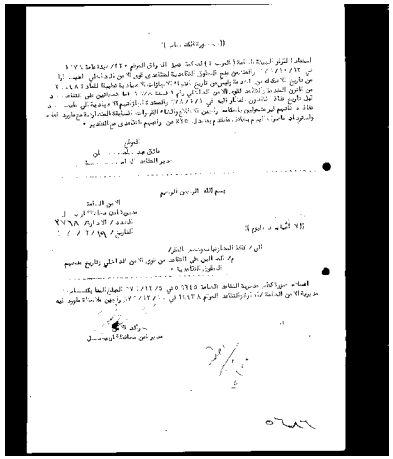
Class\Eval	precision	recall	detected	correct	total
00	98.89%	97.34%	20293	20068	20617
01	63.27%	61.69%	196	124	201
02	62.12%	88.29%	425	264	299
03	40.70%	46.05%	86	35	76
04	57.89%	64.71%	57	33	51
05	35.76%	60.47%	509	182	301
06	71.86%	83.33%	167	120	144
07	0.00%	0.00%	0	0	13
08	77.36%	51.25%	53	41	80
09	0.00%	0.00%	0	0	4

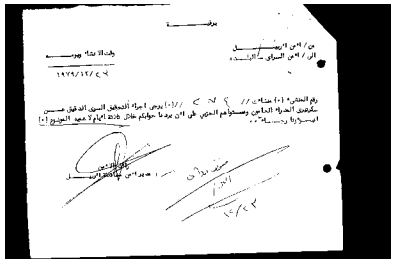
Bobcat – DI: Pilot Evaluations

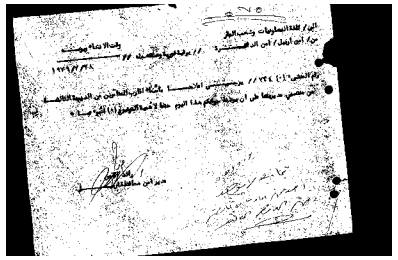
State of the Art:

- **Stable tools are available for OCR evaluation**

- **Few tools are general enough to handle complex layouts and the layout of noisy handwritting**







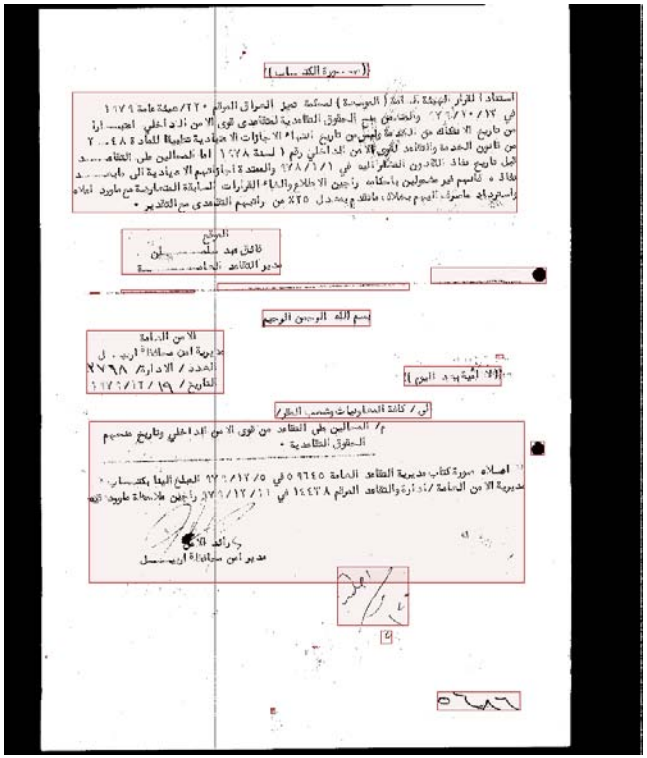
Survey: Page Segmentation Algorithms

Geometric

- Dividing document into homogenous zones

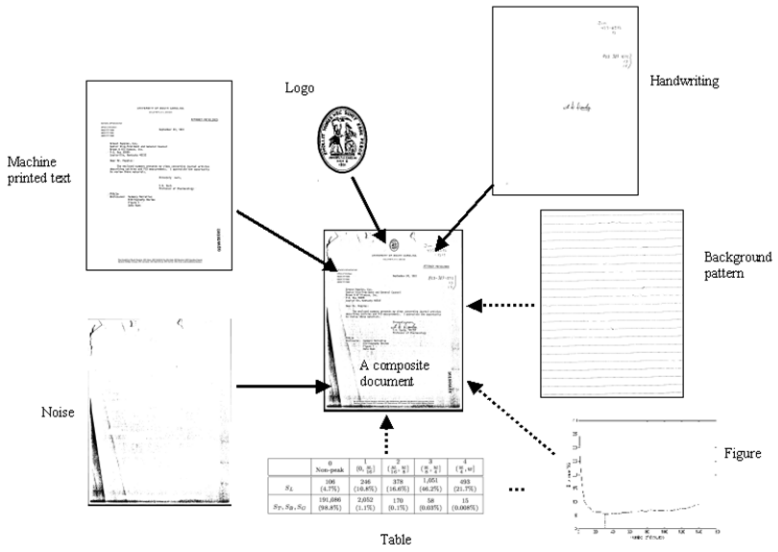
Layout

- Providing Zone content labeling
- Assigning logical relations based on location



Page Segmentation and Analysis

Goal: Decomposition of the page into constituent parts



Problems

- Page Layout Analysis
- Page Segmentation
- Handwritten Line Detection
- Text Enhancement

Localization for the purpose of “internal routing” of content

State of the Art:

- Whitespace or texture based segmentation
- Assumes objects are physically segmentable

Handwriting Challenges:

- Curvilinear text lines and small or missing linear inter-line gaps
- Stray marks which make rectangular white space analysis difficult
- Local skew
- No well-defined baselines, no Manhattan layout
- Regions not rectangular in nature - bounding box may not be the best representation

Geometric Page Segmentation

Options for Arabic?

- X-Y cuts
  - Layout to complex
- Smearing
  - Layout to Complex
- Whitespace Analysis
  - Noisy
- Constrained Text-Line Detection
  - More types of zones
- Docstrum
  - Zone Overlap
- Voronoi based
  - Maybe

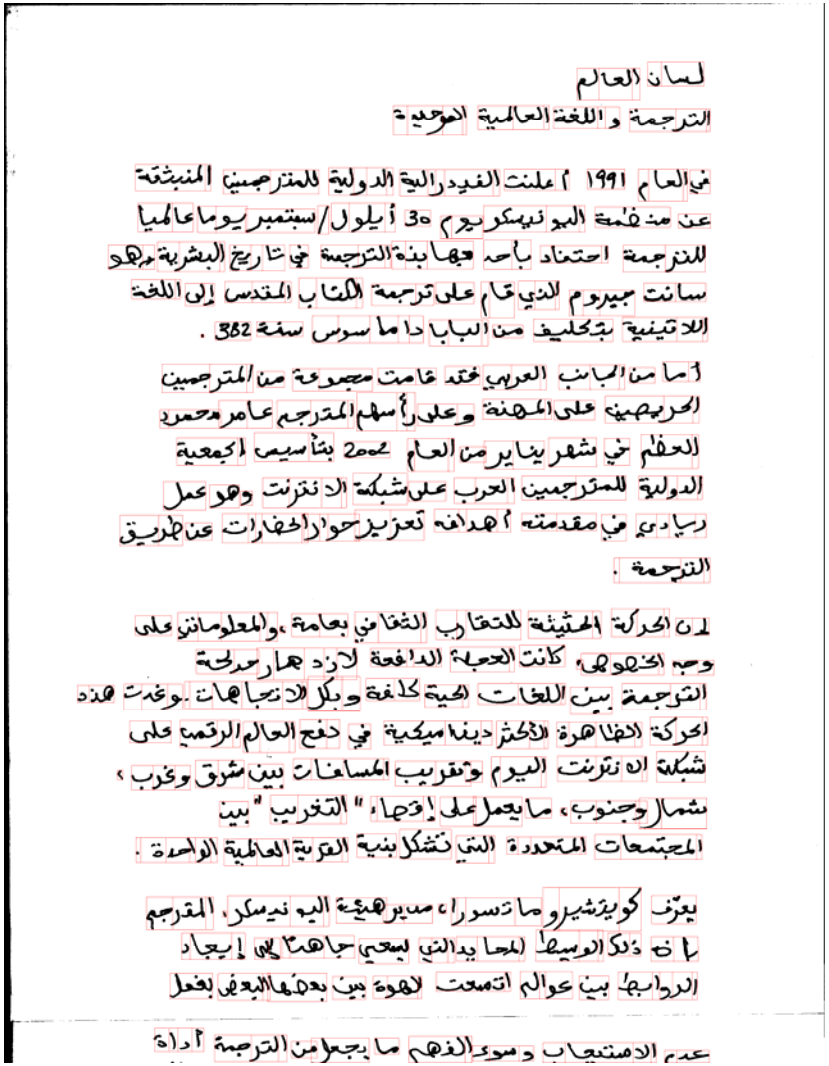
Sample Results

Collaborations: Efforts Leveraged

- ✓ Doclib
  - An integrated Development Framework
  - Core algorithms so we don't have to reinvent
  - Unified representations
  - Industry level testing and maintenance
- ✓ ARL IOTK and Information Exploitation Laboratory (IEL) Efforts
- ✓ UMD - LAMP developmental technologies
- ✓ Multilingual Automated Document Classification and Translation Program (MATCAT)
  - Cutting Edge Research
  - Arabic Handwritten Expertise

Supporting MADCAT

- Set of Word Boxes Mapped to Lines
- Run Length Encoded Data in each zone
- Algorithms return Polygons which are matched at the line level.
- All Annotation done with GEDI



لسان العالم
الترجمة و اللغة العالمية الموحدة

في العام 1991 أعلنت الفيدرالية الدولية للمترجمين المنبثقة
عن منظمة اليونيسكو يوم 30 أيلول/سبتمبر يوما عالميا
للترجمة احتفاء بأحد مجالات الترجمة في تاريخ البشرية وهو
سانت جيروم الذي قام على ترجمة الكتاب المقدس إلى اللغة
اللاتينية بتكليف من البابا داماسوس سنة 382 .

أما من الجانب العربي فقد قامت مجموعة من المترجمين
العربيين على المهنة وعلى رأسهم المترجم عامر محمود
الحكيم في شهر يناير من العام 2006 بتأسيس الجمعية
الدولية للمترجمين العرب على شبكة الانترنت وهو عمل
ريادي في مقدمته أهدافه تعزيز حوار الحضارات عن طريق
الترجمة .

لأن الحركة الحثيثة للتقارب الثقافي بعامة ،والمعلوماتي على
وجه الخصوص، كانت الدعامة الدافعة لازدهار حركة
الترجمة بين اللغات الحية كافة وبكل الاتجاهات .وغدت هذه
الحركة الظاهرة الأكثر ديناميكية في دفع العالم الرقمي على
شبكة الانترنت اليوم وتقريب المسافات بين شرق وغرب ،
شمال وجنوب، ما يعمل على إذكاء " التغريب " بين
المجتمعات المتعددة التي تشكل بنية القرية العالمية الواحدة .

يعرّف كويتشيرو ماتسوراء مدير هيئة اليونيسكو، المترجم
بأنه ذلك الوسيط المحايد الذي يسعى جاهدا إلى إيجاد
الروابط بين عوالم اتسعت الهوة بين بعضها البعض بفعل

عدم الاستيعاب وسوء الفهم ما يجعل من الترجمة أداة

Bobcat – DI: Reports

- Evaluation and Technology Survey
  - Algorithms, Datasets and Evaluation Metrics
  - Page Segmentation, OCR, Zone classification and Image processing

- PETS Software Specification and Manual
  - Evaluation Algorithms
  - Software Usage

- Experimental Laboratory Environments: Image and OCR Tool Kit (IOTK) Utility Exploration, ARL Technical Memorandum 2 -2008

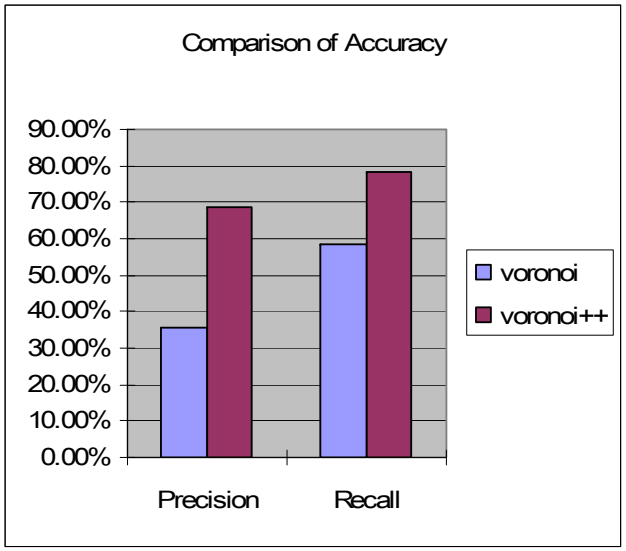
- PETS based Pilot Evaluations
  - Identification and evaluation of five technologies
    - ❑ Page Segmentation
    - ❑ Zone Classification
    - ❑ Rule Line Detection and Removal
    - ❑ Text Line Detection
    - ❑ Clutter Detection and Removal

Evaluations Preliminary Results

- Evaluations were performed on UMD-LAMP Software Developed for the DOD/MADCAT
  - Page Segmentation : Voronoi++
  - Zone Classification
  - Rule Line Detection and Removal
  - Text Line Detection
  - Clutter Detection and Removal
- All software being integrated into DocLib
- Pixels Accurate Evaluations used evolving “ImgDIFF” enhancements to PETS
  - Line Removal
  - Line Detection
- Publications in progress!

Page Segmentation

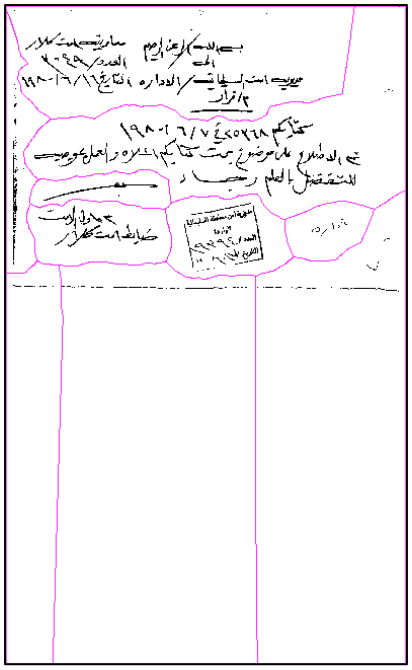
- Extension of State of the Art Approach to Deal with Handwriting – Variable tolerance for partitions, Adaptive Parameter Tuning



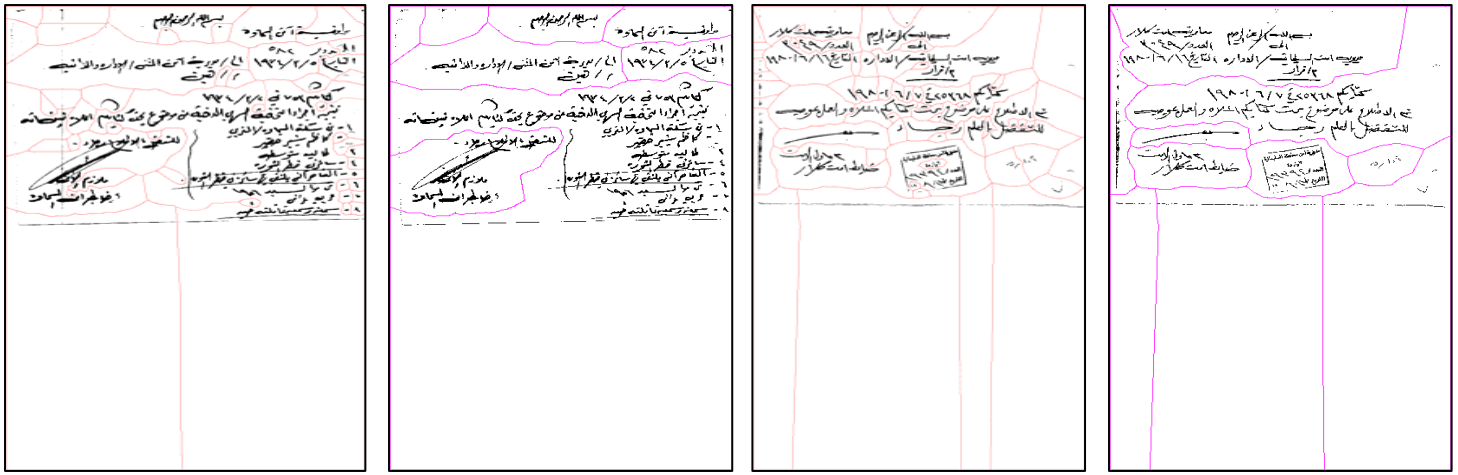
	Precision	Recall
voronoi	35.36%	58.49%
voronoi++	68.78%	78.30%

BEFORE

AFTER



Need Object Measures



Original Vorinoi **Versus** **UMD Vorinoi++**

Zone Classification

Approach:

- Construct *C(C-1)/2* two-against-all classifiers -- indicator classifiers; $f_{a\text{-}b,all}$
- Construct *C(C-1)/2* one-against-one classifiers; $f_{a,b}$
- Use indicator classifiers to determine which binary classifier to use

Results

Test on UW dataset - 1690 documents – 24531 zones

10 zone classes

Chemical drawings, small text and symbols, drawing, halftone, logo or seal, map, math, ruling, table and large text

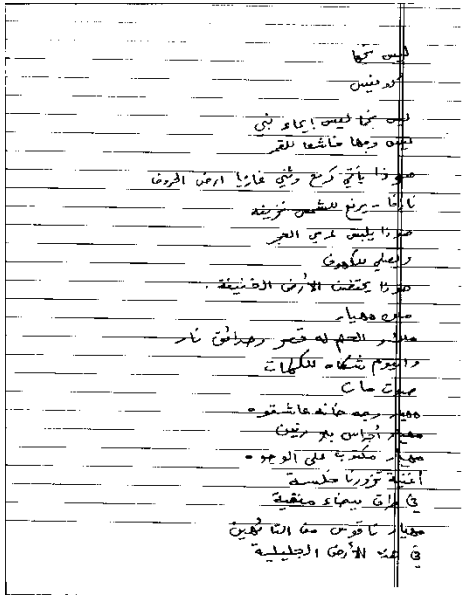
	1-vs-1	Wang et al.	Hybrid
Unbalanced	93.1%	98.45%	97.3%
Balanced	88.2%	N/A	96.6%

Rule Line Detection and Removal

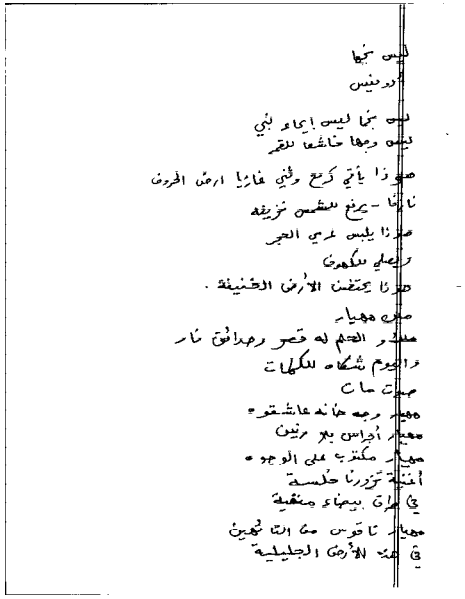
Results

- Trained 12 one-class SVMs
- Test on synthetic dataset
- Compute average and median F1

BEFORE



AFTER



	F1 Measure					
AVERAGE	0.742	0.653	0.742	0.662	0.746	0.624
MEDIAN	0.741	0.659	0.738	0.697	0.748	0.653

Text Line Detection

Affinity Propagation – Unsupervised Cluster Technique

- Operates in similarity space, rather than the feature space
- Number of clusters need not be *a priori* specified
- Can cluster models on non-Euclidean manifolds

Results

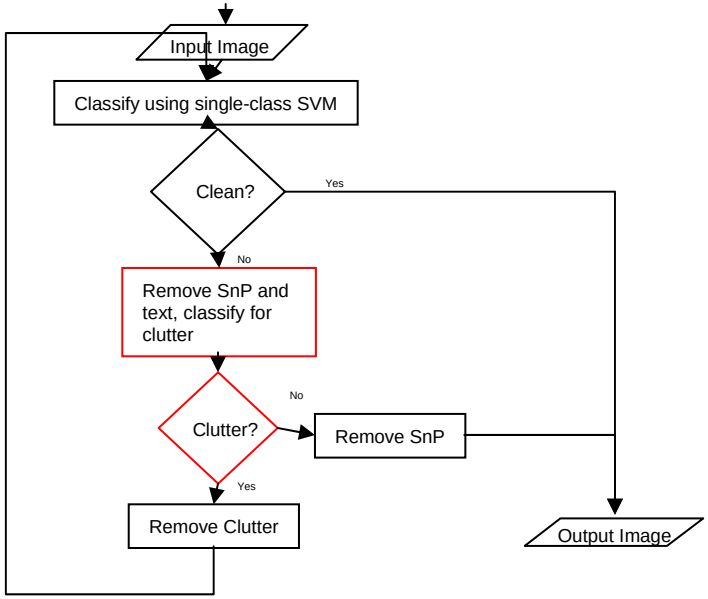
- Algorithm evaluated on the LDC dataset
- 1250 document images
- 21145 text lines

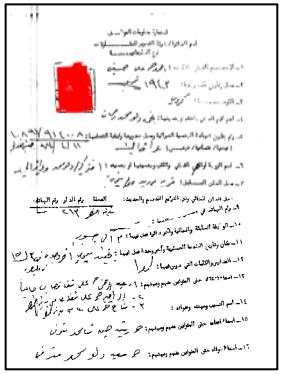
Precision	Recall	F1 score
78.2%	84.15%	81.06%

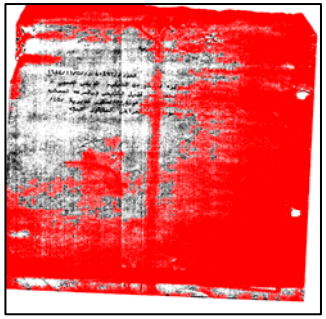
Clutter Removal

- Pixel Level Enhancement Method

System Model







Dataset	**100 images** **Handwritten, printed** **English, Arabic** **50 each for noisy n clean**
Train vs Test	**30:20 from each class**
# of Features Linear SVM	**7**
Accuracy	**97.5% Pixel Accuracy**

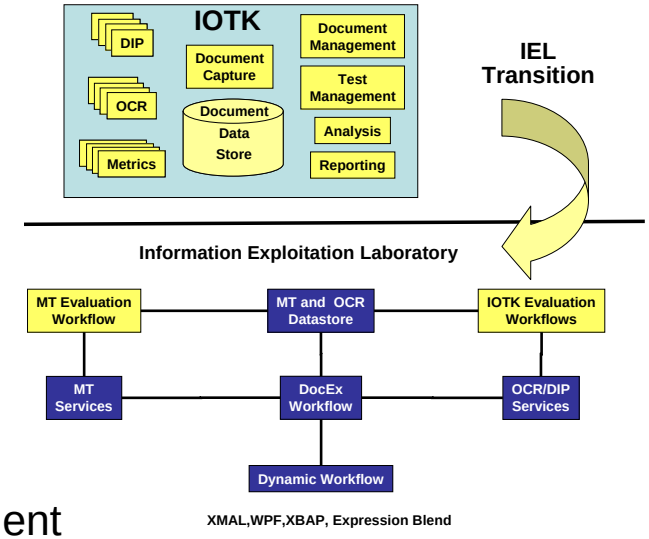
Future Plans

Additional Work:

- Creation of Additional Datasets
- Extensions to GEDI
  - Semi-automated Annotation
- Extensions to IOTK

- Extensions of PETS for Image Based Evaluation

- Collaborations to Extend capabilities as web services



Value Added

Community would benefit significantly from a common environment

DocLib

Structured Access to tools – IOTK/GEDI/PETS

Datasets and results of comparable research

Data and Evaluation

Resource where researchers could share data

- Data would be partitioned into test and training sets
- Individuals could run their data on state of the art algorithms
- Evaluation results could be tracked across time

Proposed Framework

Claims:

- Community would benefit significantly from a common environment
  - DocLib
  - Structured Access to tools – IOTK/GEDI/PETS
  - Datasets and results of comparable research

Data and Evaluation

- Resource where Researchers could Share data
  - Data would be partitioned into test and training sets
  - Individuals could run their data on state of the art algorithms
  - Evaluation results could be tracked across time

Summary: Project Contributions

- Development of Standard Practices framework
- Ability to generate and provide organized access to datasets
- Access to enhanced Annotation and interface tools (GEDI, IOTK)
- Collaborative Annotation of Existing sets
- New datasets
- Embedded Development (DocLib) and Evaluation (PETS) Tools
- Technical Documents
- Support collaborations with ongoing efforts

Summary

- Background
- BOBCAT – DI Goals
- BOBCAT – DI Objectives
- Project Accomplishments
  - Datasets
  - Tools
  - Evaluations
- Future Plans - Environment Layout and Framework

Best of Breed Configurable Active Testbed – Document Image (BOBCAT – DI) Project

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

Luis Hernandez
Army Research Laboratory
lhernandez@arl.army.mil
301-394-4301

11 December 2008