

UNCLASSIFIED



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



***TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.***

**Luis Hernandez  
Army Research Laboratory**

**11 December 2008**



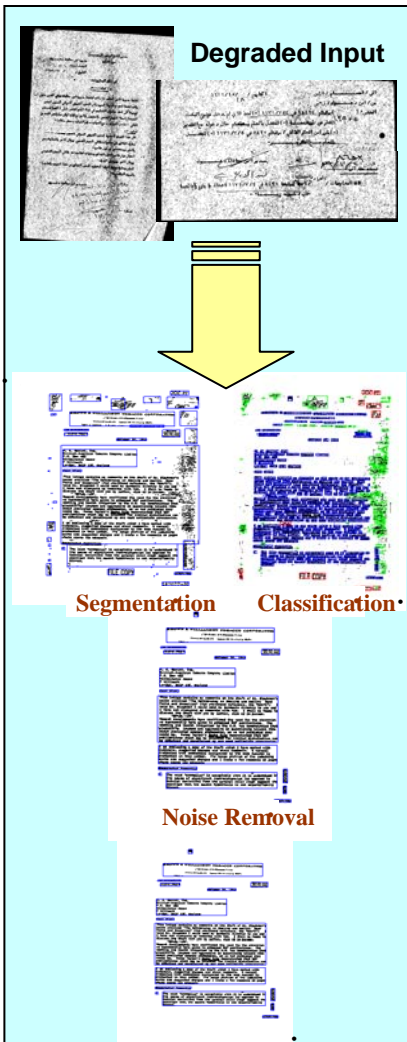
UNCLASSIFIED ***TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.***



# Outline



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



**Workflow and associated process**



**Hand-Written or Degraded Document Images into machine readable documents**



**Military Applications**

## 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



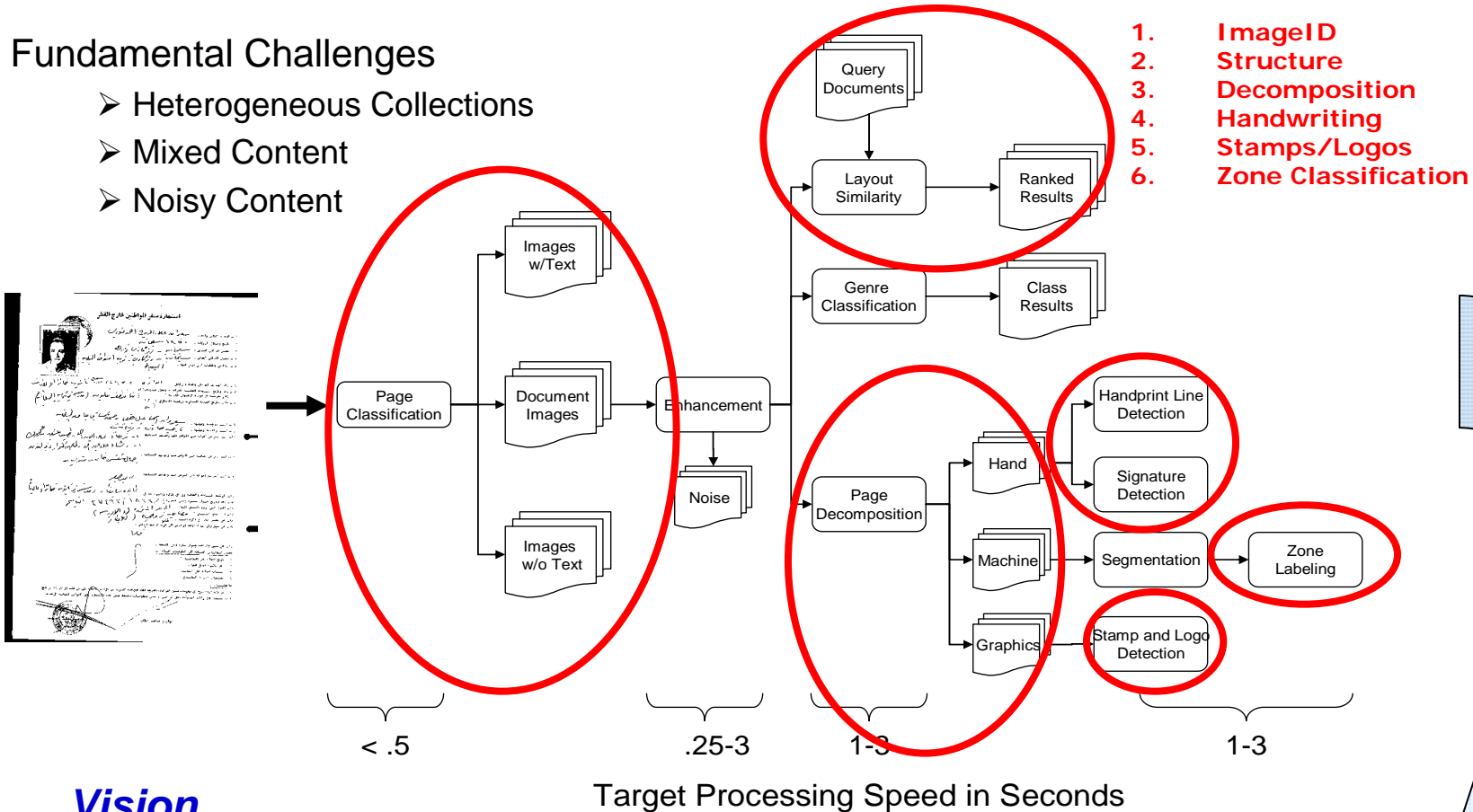
Critical use of MT





## Fundamental Challenges

- Heterogeneous Collections
- Mixed Content
- Noisy Content



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

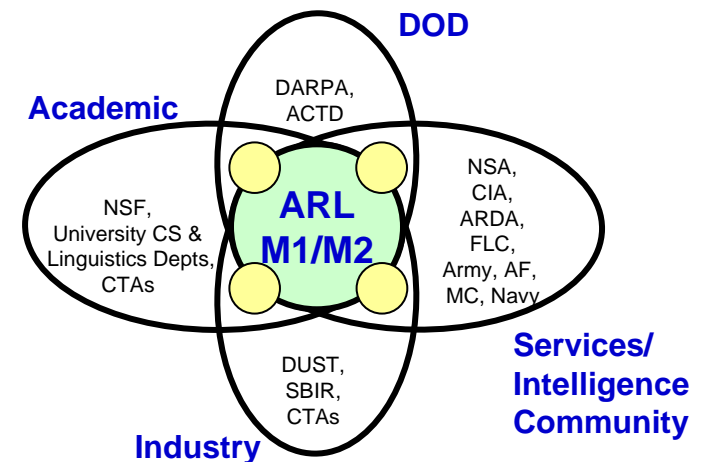
## Vision

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

- 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

- 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 do 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

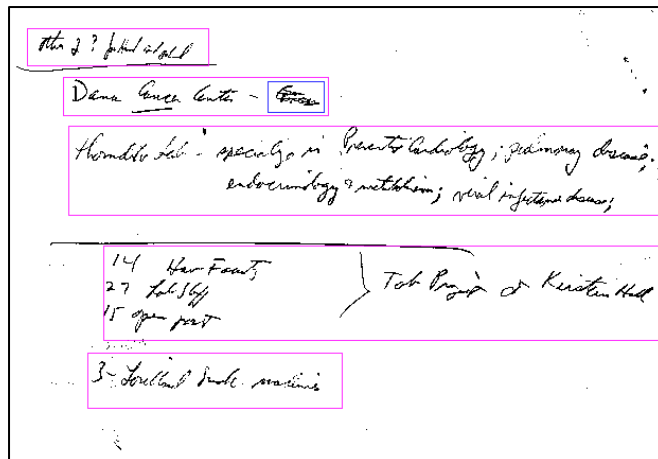


Name	Number of Images	Level of GT	Applications	Comments
AMA-Arabic 1.2	5000	GEDI - Word + PAWs	OCR	Arabic Handwritten
UMD/AMA-Zones	> 25,000	Zones + Type	Segmentation - Classification	Subset of Tobacco
Tobacco Clutter	7.9million	None	Segmentation - Enhancement	English, Some Foreign Language
MADCAT – 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

- 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)



- 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



The screenshot displays the UMD GEDI Tool Interface. The main window shows a document with Arabic text. The text is highlighted in pink and green, indicating different levels of matching or processing. The interface includes a menu bar (File, Edit, Scripts, View, Window, Help), a toolbar with various icons, and a status bar at the bottom. On the left side, there are several panels: a file list, a table with columns for NAME, COLOR, KEY, VISBLE, and COUNT, and a page zoom section.

NAME	COLOR	KEY	VISBLE	COUNT
DL_FASEALDAM	Green	None	✓	5
DL_MATCH	Pink	None	✓	11

Page Zoom

Attribute	Value
gedi_type	DL_PAGE
(col,row)(width,height)	(0,0)x(2552,3249)



# 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>

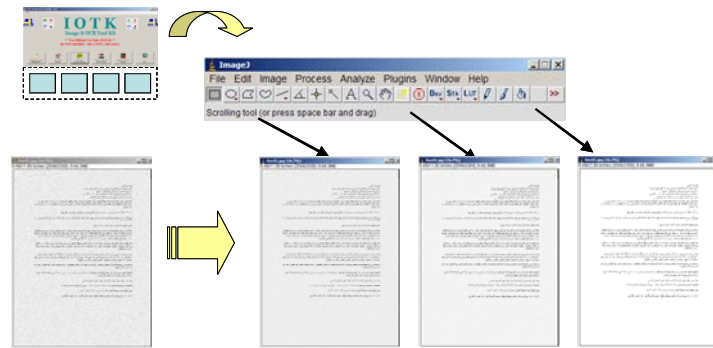
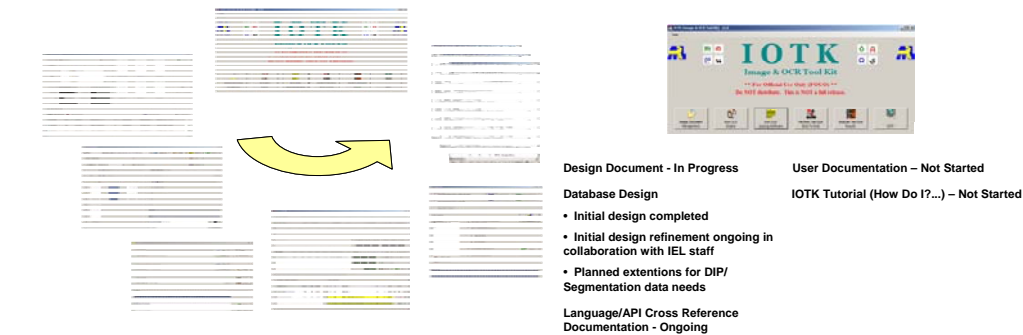
```

## Statistics

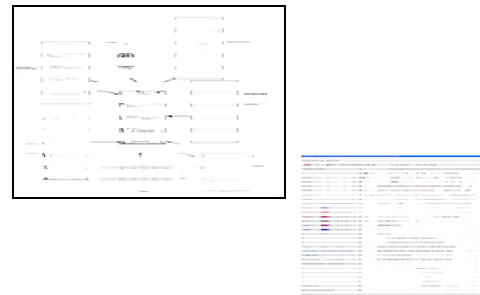
### Distribution of Zone Categories in AMA-Zones Dataset

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	<b>Total</b>	<b>323,015</b>
memo	1,893		
newspaper	615		
periodical	22		
photograph	227		
questionnaire	188		
report	985		
tables	690		
<b>Total Documents</b>	<b>16,943</b>		
<b>Page Count</b>	<b>26,007</b>		

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

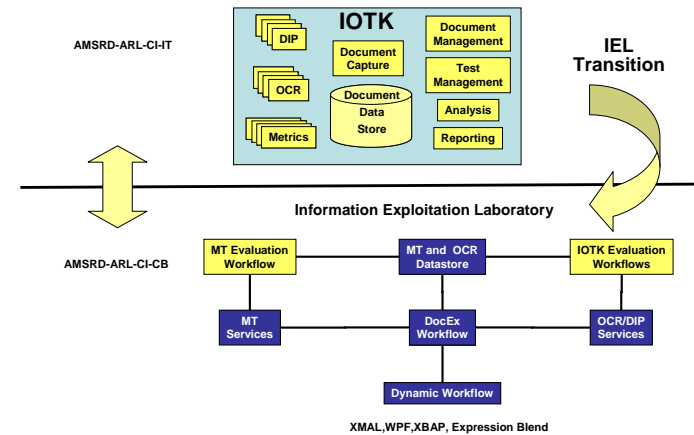


Digital Image Processing Tools for Cleaning/Degrading Images



Segmentation GT Generation, Metrics and Evaluation

## IOTK collaborations to extend capabilities as web services

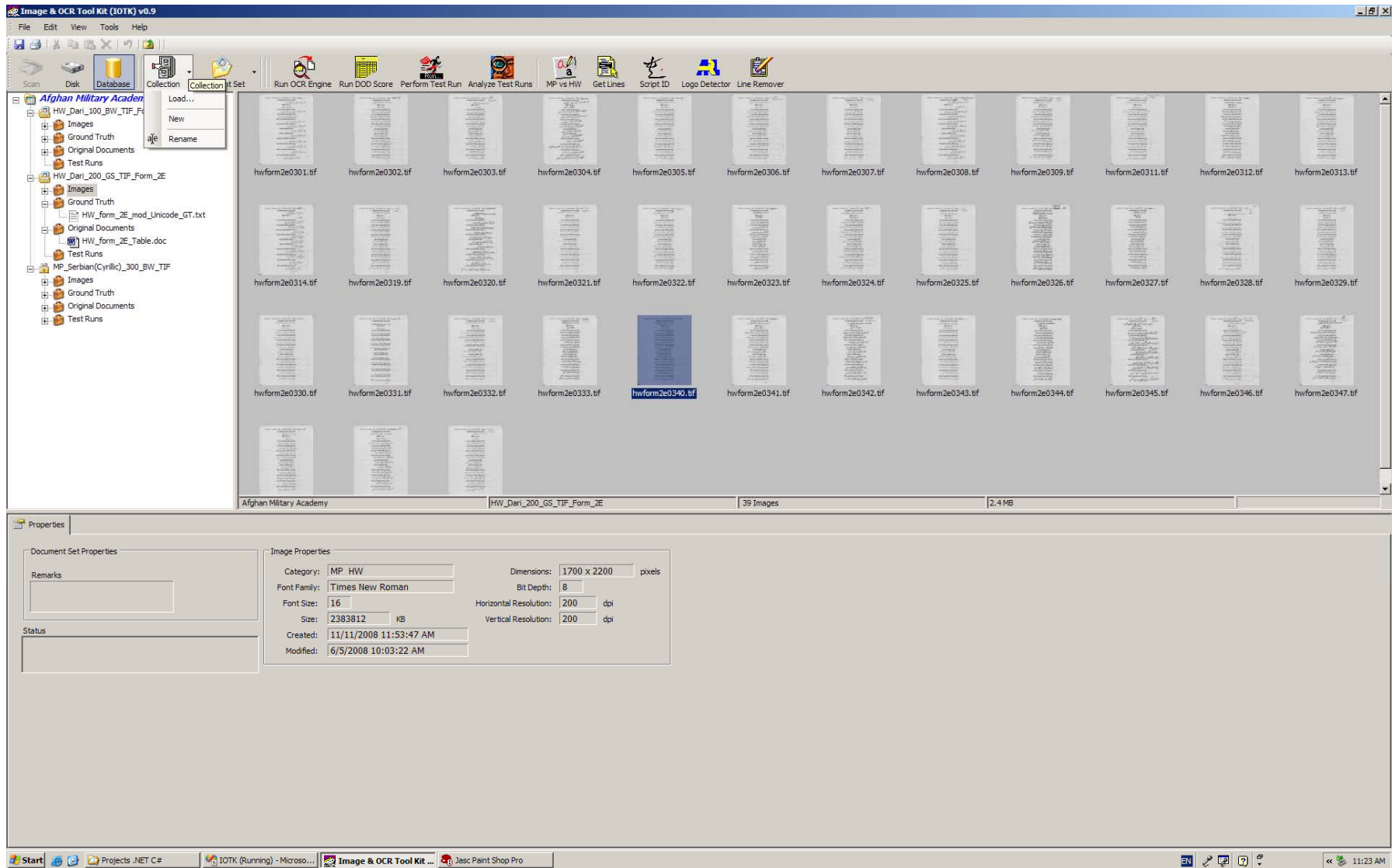


GT Document Image Data Sets





# IOTK User Interface



TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.





# IOTK OCR Module



Image & OCR Tool Kit (IOTK) v0.9

File Edit View Tools Help

Scan Disk Database Collection Document Set Run OCR Engine Run DOD Score Perform Test Run Analyze Test Runs MP vs HW Get Lines Script ID Logo Detector Line Remover

00010028.TIF 00010032.TIF 00010033.TIF 00010038.TIF 00010051.TIF 00010055.TIF

**Verus 2.0.0 Options**

Output File Type: Unicode Text (UTF-16) (\*.txt)

Image Correction: Noise Filter: Automatic

Zone Detection: Automatic

Automatically fix small rotations  
 Automatically fix 90° rotations  
 Automatically crop image

Recognize Diacritics  
 Process Handwritten Arabic Text

OK Cancel Set Default

C:\Documents and Settings\briesh\Desktop 397,754.00 KB

Properties Run OCR Engine

Language: Arabic OCR Engine: NovoDynamics Verus (Ver. 2.0.0, Bld: Unknown) Options

**Input Image List**

File Name	Source
00010028.TIF	(System Disk) C:\Documents and Settings\briesh\Desktop
00010032.TIF	(System Disk) C:\Documents and Settings\briesh\Desktop
00010033.TIF	(System Disk) C:\Documents and Settings\briesh\Desktop
00010038.TIF	(System Disk) C:\Documents and Settings\briesh\Desktop
00010051.TIF	(System Disk) C:\Documents and Settings\briesh\Desktop
00010055.TIF	(System Disk) C:\Documents and Settings\briesh\Desktop

**Output Selection**

Output Folder: C:\Projects.NET C#\IOTK\bin\Debug\My Documents\OCR Output

OCR Output Files

Progress

Status

View Log START Run DOD Score Finish Test Run

Start Projects .NET C# IOTK (Running) - Microso... Jasc Paint Shop Pro Image & OCR Tool Kit (I... Verus 2.0.0 Options 11:32 AM

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.



# IOTK OCR Accuracy Module



The screenshot displays the 'Image & OCR Tool Kit (IOTK) v0.9' application window. The main interface is titled 'Analyze Test Runs' and shows the following data:

**Test Run Archive**

Test Run Name	#Files	OCR Engine	Language	CPU	Document Set	Proc Time	Test Date
anfal_Verus_20	0	NovoDynamics Verus (Ver: 2.0...	Arabic	Intel Core(TM)2 Duo CP...	Anfal Data 1	00:00:08.713	12/11/2008

**Test Run Totals**

Chars	Errors	%Acc	Words	Errors	%Acc
1184	975	0.00	159	154	0.00

**DOD Score File Viewer**

File Name	Chars	Errors	%Acc	Words	Errors	%Acc	Proc Time
00010028score.bt	1184	975	17.65	159	154	3.14	00:00:01.654

**File Statistics**

Chars	Errors	%Acc
1184	975	17.65
Words	Errors	%Acc
159	154	3.14
Proc Time: 00:00:01.654		

**Character Category Breakdown**

Category	Count	Missed	%Right
Arabic Digits	58	57	0.00
Arabic Letters	808	685	0.00
ASCII Spacing Characters	178	95	0.00
ASCII Special Symbols	140	138	0.00

**List of Characters**

Char	Unicode	Description	Count	Missed	%Right
%	0025	Percent Sign	1	1	0.00
(	0028	Left Parenthesis	5	5	0.00
)	0029	Right Parenthesis	5	5	0.00
-	002D	Hyphen-Minus	1	1	0.00
.	002E	Full Stop (Period)	2	0	100.00
/	002F	Solidus (Slash)	18	18	0.00
_	005F	Low Line (Underscore)	108	108	0.00
ء	0621	Arabic Letter Hamza	2	2	0.00
آ	0626	Arabic Letter Yeh w/Hamza above	4	4	0.00
آ	0627	Arabic Letter Alef	176	111	0.00

The interface also includes a 'Properties' pane on the left with 'Document Set Properties' and 'Image Properties' sections, and a taskbar at the bottom showing the Start button and several open applications including 'IOTK (Running) - Micro...' and 'Analyze Test Runs'.

TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.

**Zone Analysis**

Zone ID	Label	Upper Left	Lower Right	Width	Height
Zone 0		(835, 2205)	(848, 2225)	13	20
Zone 1		(1358, 2154)	(1618, 2212)	260	58
Zone 2		(988, 2055)	(993, 2059)	5	4
Zone 3		(130, 2044)	(135, 2049)	5	5
Zone 4		(967, 2011)	(982, 2018)	15	7
Zone 5		(1181, 1963)	(1215, 2004)	34	41
Zone 6		(1045, 1763)	(1268, 1947)	223	184
Zone 7		(267, 1306)	(1630, 1814)	1363	508
Zone 8		(1649, 1370)	(1692, 1412)	43	42
Zone 9		(850, 1248)	(1338, 1259)	488	51

PETS Viewer  
File Name: 00010028.xml  
Segment: Docstrum

Zones (Ground Truth) Image Properties

Coordinate: Pixels

Zone ID	Zone Class	Upper Left	Lower Right	Width	Height
<input checked="" type="checkbox"/> Zone 0	text_sm	(740, 148)	(754, 167)	14	19
<input checked="" type="checkbox"/> Zone 1	text_sm	(616, 185)	(1143, 231)	327	46
<input checked="" type="checkbox"/> Zone 2	text_sm	(216, 252)	(1606, 670)	1390	418
<input checked="" type="checkbox"/> Zone 3	text_sm	(270, 702)	(1616, 908)	1346	206
<input checked="" type="checkbox"/> Zone 4	text_sm	(807, 958)	(1146, 1011)	339	53
<input checked="" type="checkbox"/> Zone 5	text_sm	(1252, 1098)	(1586, 1195)	334	97
<input checked="" type="checkbox"/> Zone 6	text_sm	(260, 1022)	(688, 1219)	428	197
<input checked="" type="checkbox"/> Zone 7	text_sm	(247, 1280)	(261, 1284)	14	4
<input checked="" type="checkbox"/> Zone 8	text_sm	(261, 1248)	(1630, 1838)	1369	590

Zones (Results) Zone Match Breakdown Zone Classification Breakdown

Coordinate: Pixels

View: Results + Ground Truth

Zone Class Filter: All

Zone ID	Zone Class	Upper Left	Lower Right	Width	Height
<input checked="" type="checkbox"/> Zone 0	text_sm	(835, 2205)	(848, 2225)	13	20
<input checked="" type="checkbox"/> Zone 1	text_sm	(1358, 2154)	(1618, 2212)	260	58
<input checked="" type="checkbox"/> Zone 2	text_sm	(988, 2055)	(993, 2059)	5	4
<input checked="" type="checkbox"/> Zone 3	text_sm	(130, 2044)	(135, 2049)	5	5
<input checked="" type="checkbox"/> Zone 4	text_sm	(967, 2011)	(982, 2018)	15	7
<input checked="" type="checkbox"/> Zone 5	text_sm	(1181, 1963)	(1215, 2004)	34	41
<input checked="" type="checkbox"/> Zone 6	text_sm	(1045, 1763)	(1268, 1947)	223	184
<input checked="" type="checkbox"/> Zone 7	text_sm	(267, 1306)	(1630, 1814)	1363	508
<input checked="" type="checkbox"/> Zone 8	text_sm	(1649, 1370)	(1692, 1412)	43	42



## 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



$I$  = 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)  $\geq$  threshold  
L(i) = L(j)

## DETECTED

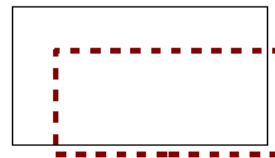
MatchScore(i,j)  $\geq$  threshold  
L(i)  $\neq$  L(j)

## FALSE

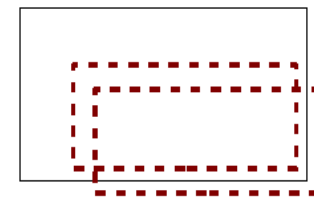
MatchScore(i,all) < threshold

## MISSED

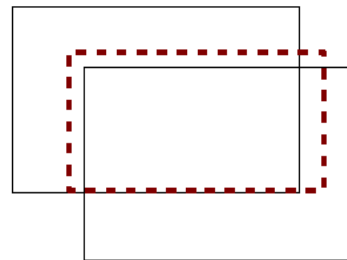
MatchScore(all,j) < threshold



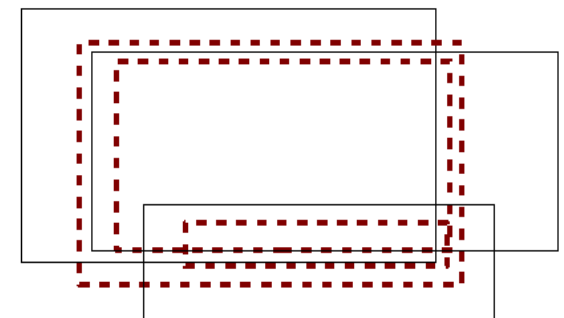
(a) one - one



(b) one - many



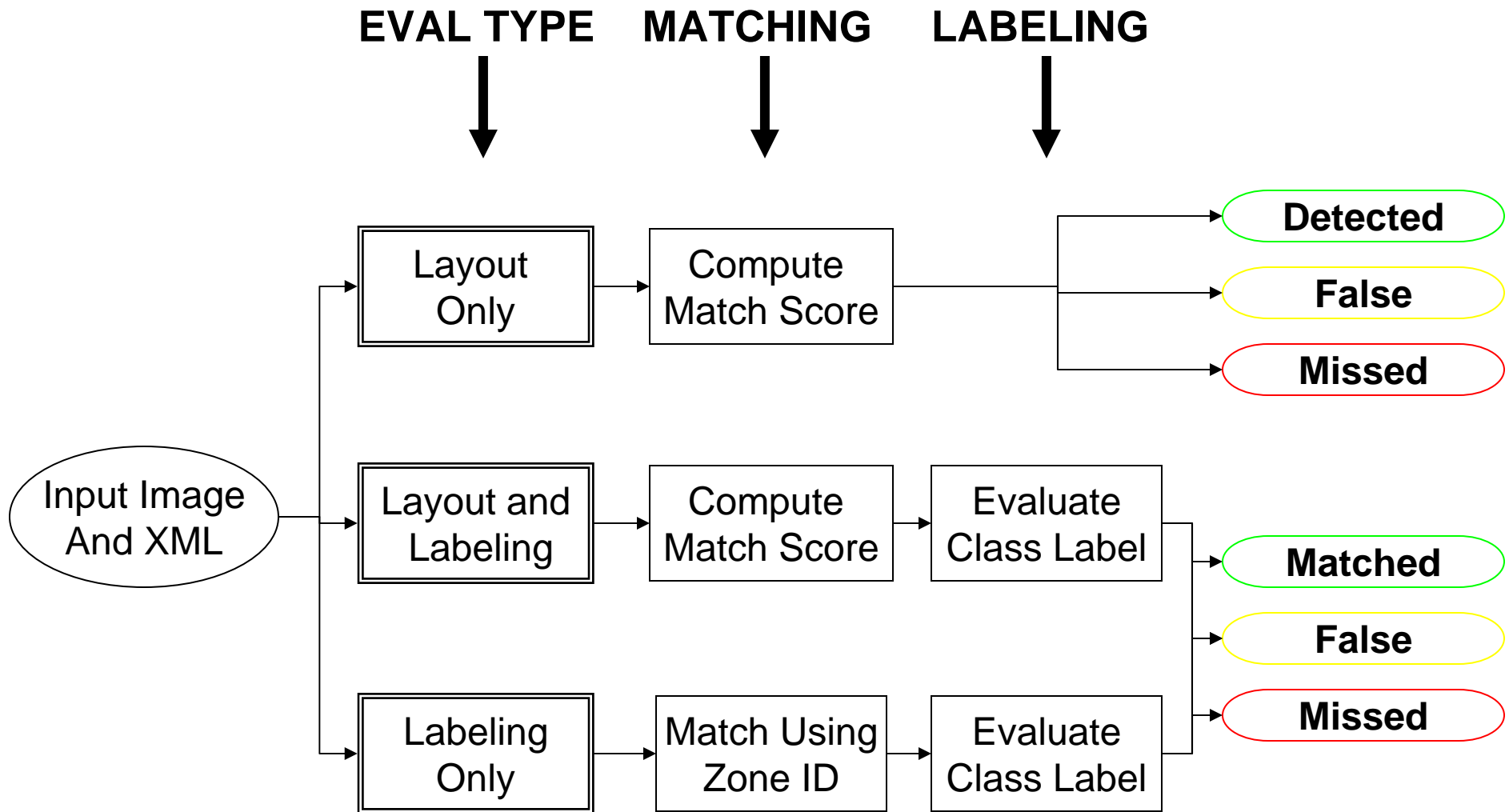
(c) many - one



(d) many - many

— : result

- - - : ground truth





# 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

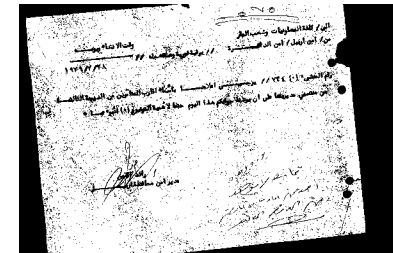
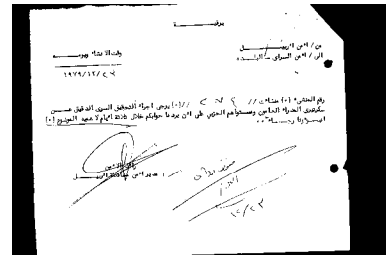
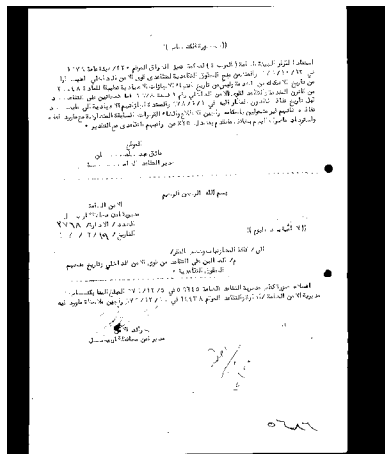
=====

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



## 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 handwriting

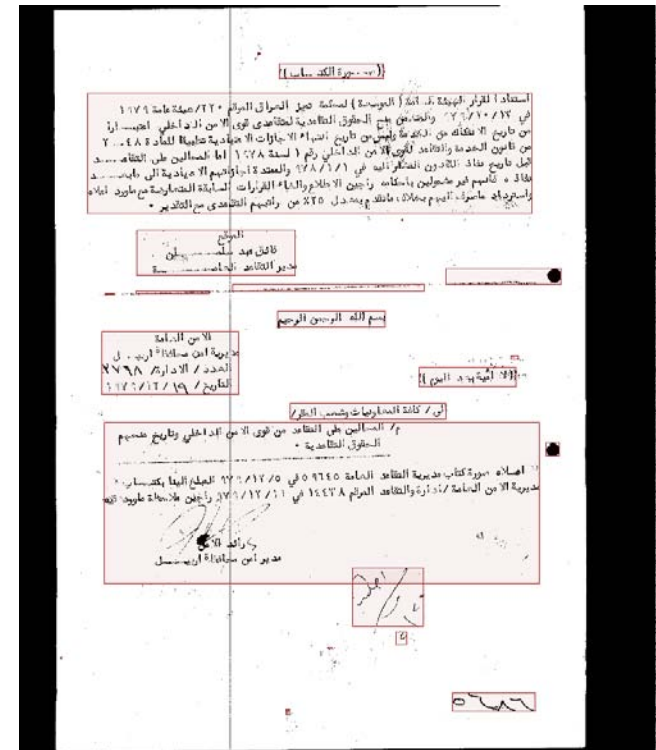


## Geometric

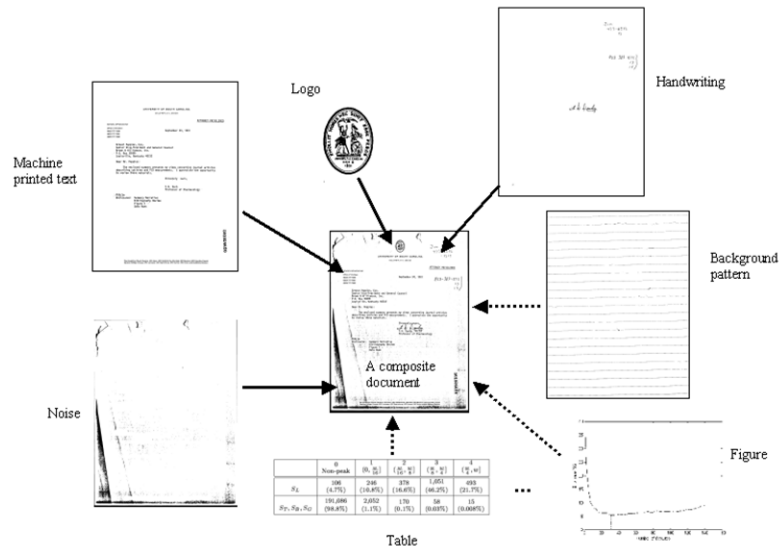
Dividing document into homogenous zones

## Layout

Providing Zone content labeling  
Assigning logical relations based on location



Goal: Decomposition of the page into constituent parts



## 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

## 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

## 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

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 واللغة العالمية الموحدة

في العام 1991 أعلنت الفيدرالية الدولية للمتترجمين المنبثقة  
عن منظمة اليونسكو يوم 30 أيلول/سبتمبر يوماً عالمياً  
للترجمة اعتماداً بأحد جهات الترجمة في تاريخ البشرية وهو  
سانت جبروم الذي قام على ترجمة الكتاب المقدس إلى اللغة  
اللاتينية بتكليف من البابا داما سوس سنة 382 .

لما من الجانب العربي فقد عاشت مصوغة من المترجمين  
الحرصيين على المهنة وعلى رأسهم المترجم عامر محمد  
الحظم في شهر يناير من العام 2006 بتأسيس الجمعية  
الدولية للمتترجمين العرب على شبكة الانترنت وهو عمل  
ريادي في مقدمته أهدافه تعزيز حوار الحضارات عن طريق  
الترجمة .

لان الحركة الحديثة للتقارب الثقافي بعامته، والمعلوماتي على  
وجه الخصوص، كانت العجلة الدافعة لأزد همار الحديثة  
الترجمة بين اللغات أهمية كلفة وبكل الاتجاهات. ووجدت هذه  
الحركة الظاهرة الأكثر ديناميكية في دفع العالم الرقمي على  
شبكة انترنت اليوم وتقريب المسافات بين مشرق ومغرب،  
شمال وجنوب، ما يجعل على إقحام "التغريب" بين  
المجتمعات المتعددة التي تشكل بنية القرية العالمية الواحدة .

يعرف كويتشيرو ما تسورا، مدير هيئة اليه نيمسك، المترجم  
لأنه ذلك الوسيط للمعايد التي يعطي حياها على إيجاد  
الدواجل بين عوالم اتصفت لهوة 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

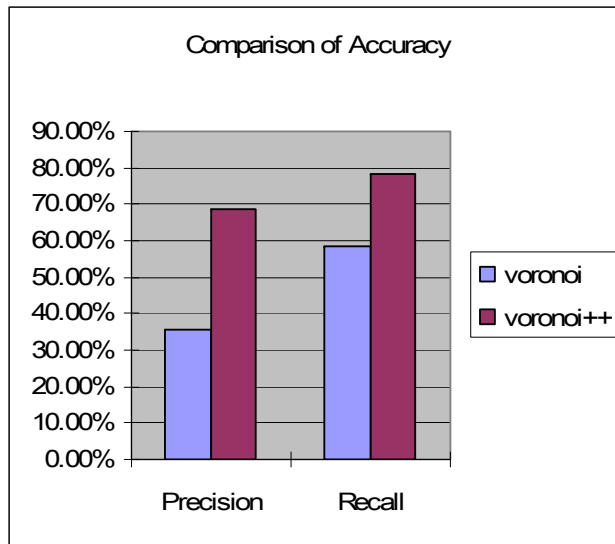


# Evaluations Preliminary Results



- 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!

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

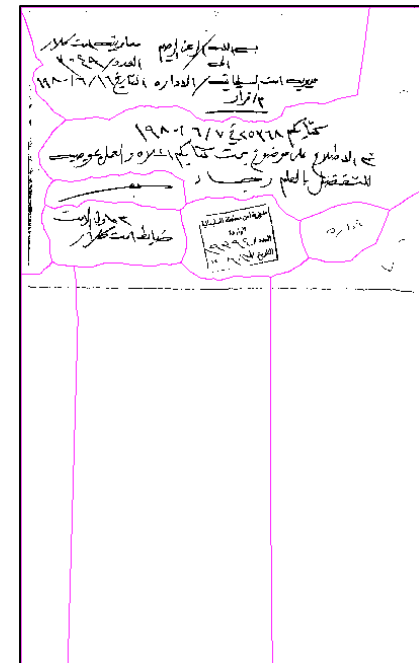


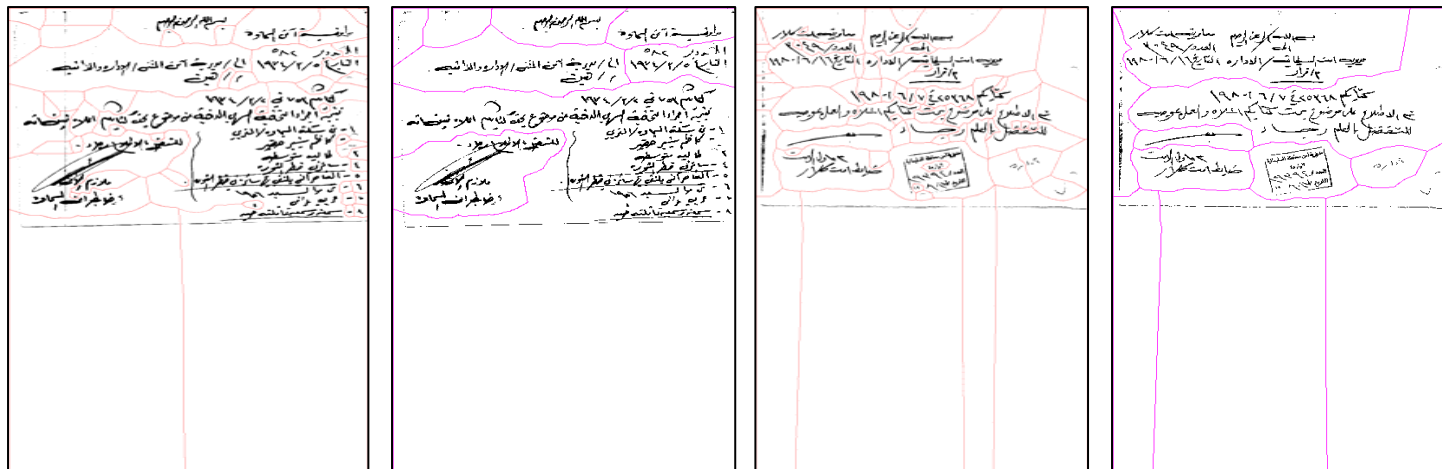
	Precision	Recall
voronoi	35.36%	58.49%
voronoi++	68.78%	78.30%

BEFORE



AFTER





Original Vorinnoi      Versus      UMD Vorinnoi++

## Approach:

- Construct  $C(C-1)/2$  two-against-all classifiers -- indicator classifiers;  $f_{a-b,all}$
- 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

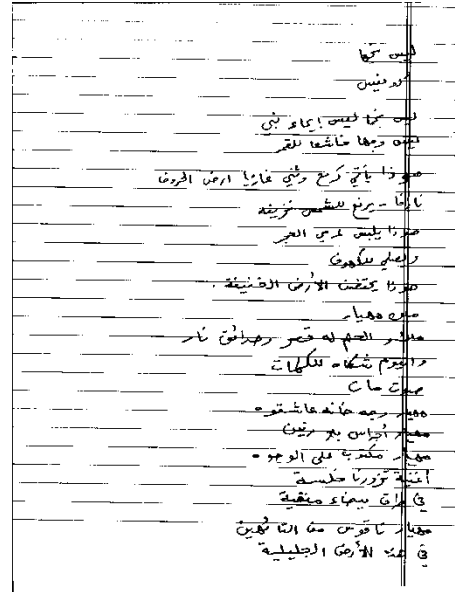
Chemical drawings, small text and symbols, drawing, halftone, logo or seal, map, math, ruling, table and large text

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

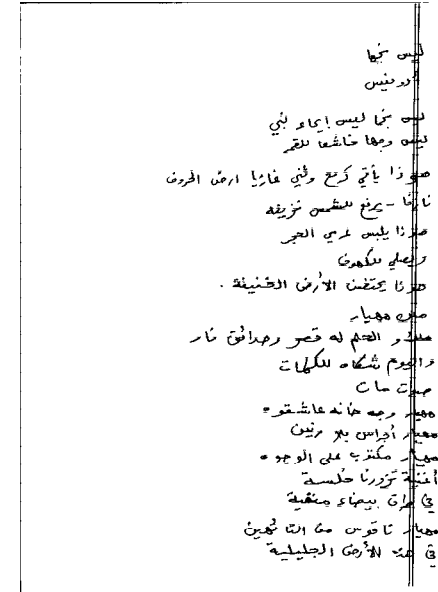
## Results

- 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



## 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

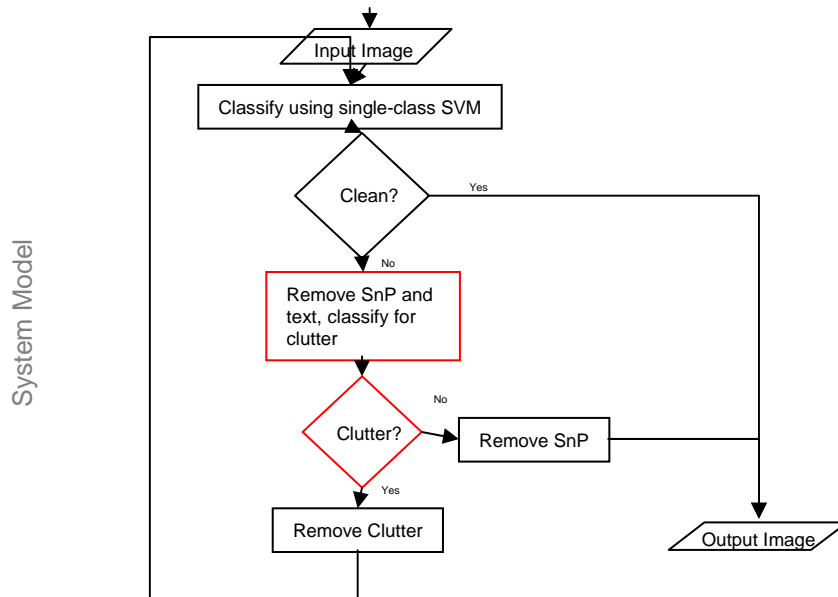
## Results

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

Precision	Recall	F1 score
78.2%	84.15%	81.06%

من فرزند و قدا دار مردم افغانستان با شمولیتم در صفوف  
 من مرد و قدا دار مردم افغانستان با شمولیتم در صفوف  
 قوای مسلح دولت افغانستان بنام خداوند متعال سوگند یاد  
 هوای صحیح دولت افغانستان با شمولیتم در صفوف  
 میکنم که یک افسر "سرپاز" با دسپلین و قدا دار به همه  
 مملکت کرکوت افسر سرپاز با دسپلین و قدا دار باشم  
 اصول، قوانین و مقررات دولت اسلامی افغانستان عزیز  
 اصول قوانین و مقررات دولت اسلامی افغانستان عزیز  
 بوده، تمام اوامر و هدایات امرین و قوماندانان را تحت  
 بوده تمام اوامر و هدایات امرین و قوماندانان را تحت  
 هر گونه شرایط زمان و مکان به مقصد دفاع از تمامیت  
 هر گونه شرایط زمان و مکان به مقصد دفاع از تمامیت  
 ارضی، استقلال ملی و ارزش انقلاب اسلامی نسبت به  
 ارضی، استقلال ملی و ارزش انقلاب اسلامی نسبت به  
 هر نوع ممانع دیگر ترجیح داده حتی از ریختن خون  
 هر نوع ممانع دیگر ترجیح داده حتی از ریختن خون  
 خویش در این راه دریغ نمی نمایم. اگر بر خلاف این  
 خویش در این راه دریغ نمی نمایم. اگر بر خلاف این  
 تکلیف یاد شده عمل نمایم، در صورت جزای سخت  
 تکلیف یاد شده عمل نمایم، در صورت جزای سخت

- Pixel Level Enhancement Method



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

## 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

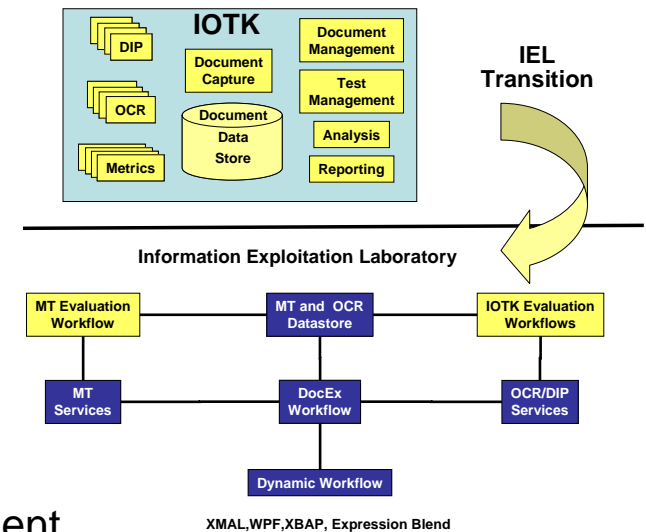
Structured Access to tools – IOTK/GEDI/PETS

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



UNCLASSIFIED



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



***TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.***

**Luis Hernandez**  
**Army Research Laboratory**  
[lhernandez@arl.army.mil](mailto:lhernandez@arl.army.mil)  
**301-394-4301**



**11 December 2008**

UNCLASSIFIED ***TECHNOLOGY DRIVEN. WARFIGHTER FOCUSED.***