



MadCat, BOBCAT & Doclib

July 23, 2008

David Doermann Wontaek Seo Elena Zotkina Mudit Agrawal Guangyu Zhu Wael Abd-Almageed Levon Panel Orri Ganel





Agenda

- Review of Current Handwriting Efforts
 - What is present
 - What is Missing
 - LAMP Focus in each Program
- Bobcat Progress on:
 - Datasets
 - Evaluation Methodology
 - Segmentation Survey and Tools
- Open Discussion of Additional Plans





MadCat Phase I

- Machine Translation Evaluation in Gale Style
- Documents Transcribed from existing Gale data
- Ground Truthed to the Word level
- OCR at the Line Level





What's Missing?

- Any type of page level analysis
- Ground Truth of Complex documents – Coming in Phase II?
- Document Analysis evaluation tools
 - Likely to be absent for all phases, hence Bobcat and NSA
- Structured Environment such as DocLib for Development
- Pushing Doclib to the Public





LAMP Role

- MADCAT: Funded as part of BBN Team
 - Line Segmentation
 - Page Segmentation
 - Enhancement page level and content level
- BobCat: Funded incrementally for evaluation tools (Through Sept 2008)
- NSA: Doclib Support (Through Dec 2008)





LAMP Future

- Looking for additional funding for 2008-2009
- Focus on Enhancement, Page Segmentation and Page Normalization
- Interested in continuing to develop evaluation and GT tools.
- 2 new students starting Fall 2008





Overview of BobCat Goals

- Transition the test methods, metrics, and procedures ... as part of the assessment infrastructure,
- Provide tools ... to extend groundtruthed datasets to include Arabic Anfal images.
- Provide test designs, data analysis procedures, and interpretation guidelines for evaluating COTS and GOTS OCR systems and other DIP tools





- Provide a basis for Phase II of MadCat
 - Groundtruthing Guidelines
 - Evaluation Metrics
 - Data Representations
- Issues:
 - How do we extend representations to Handwriting
 - How do we represent uncertainty
 - How do we provide a dataset useful for various tasks
 - segmentation, OCR, content labeling, etc





Specific Tasks

- Data
 - Zone Classification and Segmentation GT
- Tools
 - Update GEDI to allow handwritten data rep
- Evaluation
 - Zone Classification Tools



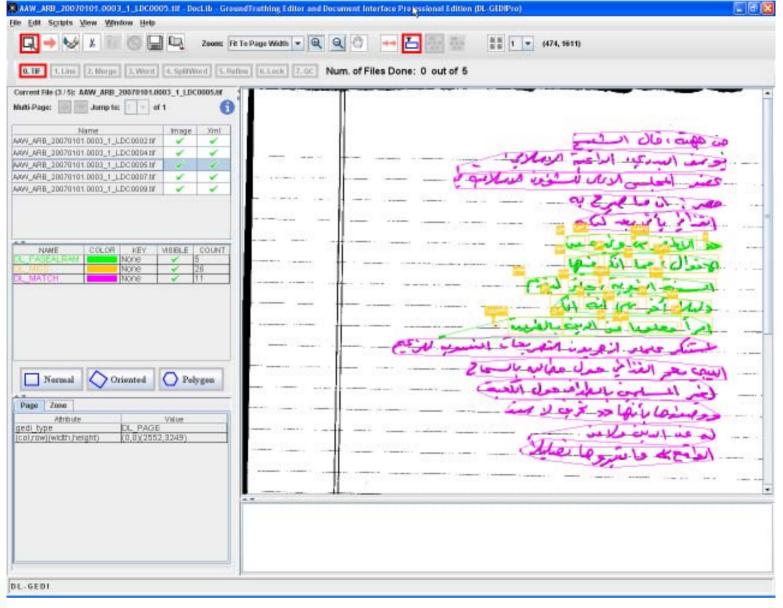


GEDI Tool

- Overview
 - Generic Tool for Representing Regions and Attributes on images
- Project Specific Extensions
 - Reading Order
 - Representation of Run Length Encoded Data for Line Segmentation
 - Direct Integration of Evaluation Capabilities











Data Sets

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

the 2? Juli alall Dana ance lente - From Hometo bet - special o in Pranto landwity; pedinoray drand. endocrimology ? methicin; wind infectione down; far Four Tot Prais or Kinstin Ha 3 Louth Phile maline





Statistics

| 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 | MACHINEPRIN | IT 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 | | |





Anfal Data

- Line of text GT with polygons
- Lines Split by
 - Physical Location
 - Change in Attribute hand/machine, size
- Reading Order used to link segments of a line





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.





Remaining Tasks

- Evaluation of Existing Data
- Sponsor testing of software
- Integration of OCR evaluation
- Feedback from MADCAT Participants





Recent Deliverables

- GEDI Toolkit
- 26,000 page Tobacco Litigation Corpus
- Full Presentation of July 20th
- Software for Classification and Segmentation Evaluation





Agenda

- > Review of Goals
- > Progress on:
 - Dalaseis
 - Evaluation Methodology
 - Segmentation Survey and Tools
- Open Discussion of Additional Plans







Evaluation Methodology and Software

Wontaek Seo David Doermann





Evaluation Modules

- Zone Classification
- Segmentation
 - Line Segmentation
 - Zone Segmentation





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"
- Zone types "can" be used for matching
- Software is integrated into DocLib
- Full match matrix is built to store the score of each pair of zones.





Matching score

- 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

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





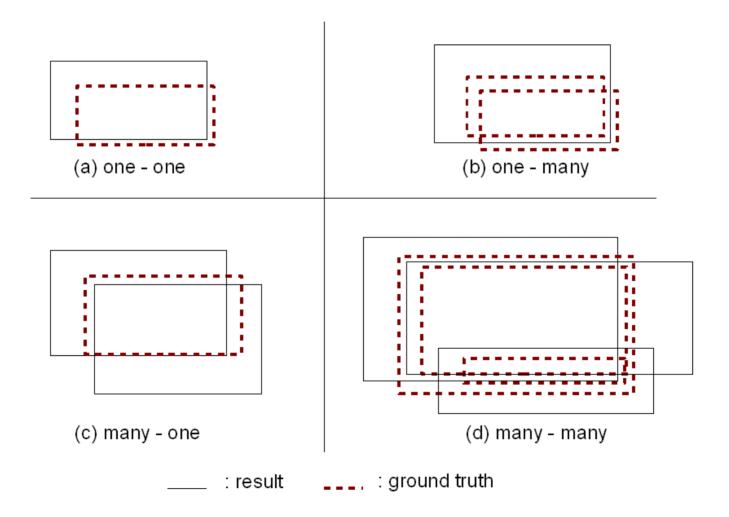
Types of result

- MATCHED
 - MatchScore(i,j) \geq threshold
 - -L(i) = L(j)
- DETECTED
 - MatchScore(i,j) ≥ threshold
 - $-L(i) \neq L(j)$
- FALSE
 - MatchScore(i,all) < threshold</p>
- MISSED
 - MatchScore(all,j) < threshold</p>





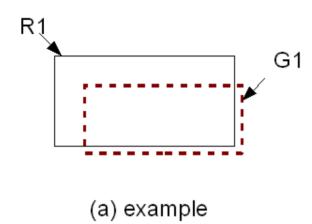
Matching examples







one-one



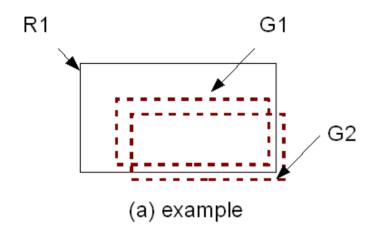
| result\GT | G1 |
|-----------|--------|
| R1 | 85.00% |

(b) matching score

- Representation
 - L(A) : Label of A
- L(R1) = L(G1)
 R1 is matching to G1
- $L(R1) \neq L(G1)$
 - R1 is detecting G1 w/ the different label



one-many



| result\GT | G1 | G2 | |
|-----------|--------|--------|--|
| R1 | 90.00% | 85.00% | |

(b) matching score

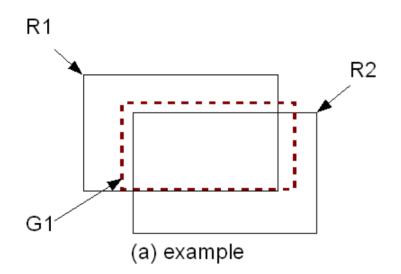
• L(R1) = L(G1) = L(G2)

- compare the matching scores
 - R1 is matching to G1
 - G2 is missing
- $L(R1) = L(G2) \neq L(G1)$
 - R1 is matching to G2
 - G1 is missing
- $L(R1) \neq L(G1) \neq L(G2)$
 - compare the matching scores
 - R1 is detecting G1 w/ the different label
 - G2 is missing





many-one



| result\GT | G1 |
|-----------|--------|
| R1 | 95.00% |
| R2 | 90.00% |

(b) matching score

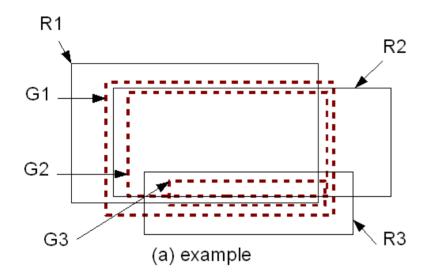
• L(R1)=L(R2)=L(G1)

- compare the matching scores
 - R1 is matching to G1
 - R2 is false alarm
- $L(R1) \neq L(R2) = L(G1)$
 - R1 is false alarm
 - R2 is matching to G1
- L(R1),L(R2)≠L(G1)
 - compare the matching scores
 - R1 is detecting G1 w/ the different label
 - R2 is false alarm





many-many



| result\GT | G1 | G2 | G3 |
|-----------|--------|--------|--------|
| R1 | 90.00% | 85.00% | 50.00% |
| R2 | 80.00% | 82.00% | 45.00% |
| R3 | 30.00% | 0.00% | 85.00% |

(b) matching score

1st step

 find the set of matched zone which is not matched to same ground truth zone

• 2nd step

- find the set of detected zone which is not matched in the 1st step
- The R which is not set at any steps is false alarm
- The G which is not set by any R is missing





Software

- PEZS : Performance Evaluation tool of Zone Segmentation
- Usage
 PEZS -r { FILE | DIR } -g { FILE | DIR } -img { FILE | DIR }
 [-o FILE -v DIR -m FILE -t NUM -detail -lid -rle -seg]

Note: Currently zone labeling eval is in Java... All will be in DocLib for final release.





Options

- r { FILE | DIR } : path to the result file or directory
- g { FILE | DIR } : path to the ground truth file or directory
- img { FILE | DIR } : path to the image file or directory
- **o FILE** : set file name of file to be saved
- v DIR : set directory where the GEDI type xml output for visualization will be saved



t NUM : set the threshold of matching score



Options

- m FILE : result zones which is in a ground truth zone will be merged if it's type is in the FILE
- detail : result of each zone will be added to the output when it is set
- rle : run-length code will be added to the visualization output
- seg : label matching will not be performed when it is set





Software Output

Zone Segmentation Evaluation Result. Generated on Sat Jul 5 11:10:57 2008

```
Result of Individual File
```

[O] : Detected, [-] : Detected w/ Different Type, [X] : Undet

AAW_ARB_20070101.0003_1_LDC0002.tif

Page ID : 1

| [0] | 1, | DL_TEXTLINEGT, | z10, | DL_TEXTLINEGT, 85.32% |
|-------|---------|-----------------|------|-----------------------|
| [0] | 2, | DL_TEXTLINEGT, | z11, | DL_TEXTLINEGT, 86.36% |
| [0] | з, | DL_TEXTLINEGT, | z2, | DL_TEXTLINEGT, 85.90% |
| [0] | 4, | DL_TEXTLINEGT, | z12, | DL_TEXTLINEGT, 80.03% |
| [0] | 5, | DL_TEXTLINEGT, | z1, | DL_TEXTLINEGT, 85.36% |
| [X] | 6, | DL_TEXTL INEGT | | — |
| [0] | 7, | DL_TEXTLINEGT, | z13, | DL_TEXTLINEGT, 85.38% |
| [X] | 8, | DL_TEXTL INEGT | | _ |
| [X] | 9, | DL_TEXTL INEGT | | |
| [X] | 10, | DL_TEXTL INEGT | | |
| [X] | 11, | DL_TEXTL INEGT | | |
| [X] | 12, | DL_TEXTL INEGT | | |
| [0] | 13, | DL_TEXTLINEGT, | z4, | DL_TEXTLINEGT, 86.00% |
| [0] | 14, | DL TEXTLINEGT, | zO, | DL TEXTLINEGT, 84.70% |
| [0] | 15, | DL_TEXTLINEGT, | z14, | DL_TEXTLINEGT, 85.99% |
| [X] | 16, | DL_TEXTL INEGT | | _ |
| [X] | 17, | dl_textlinegt | | |
| [OVER | ALL] 9/ | /0/8/17, 52.94% | | |

AAW_ARB_20070101.0003_1_LDC0004.tif





| Summary | y of Res | ults | | | | |
|---------|----------|--------------------------|-----------|-----|--|------------|
| | | of R-Zone Zone Detect | 19% | | | |
| | | n on Zones | | | | |
| | | | r of Zone | | | |
| | | TLINEGT | 22033 | 31. | | |
| | nfusion | | | | | |
| Result | t\GT | unmatch | 1 | | | |
| unma | | 0(0.0%)* .61(68.8%) | | | | |
| | sult Tab | | | | | |
| Label | Total | Detected | Precsion | | | FalseAlram |
| 1 | 19650 | 22033 | | | | 68.81% |





Zone Classification

Summary of Results

Total Number of Sample : 21786Overall 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% |





02. Confusion Matrix

| Out\GT 00 01 02 03 04 05 06 07 08 09 | $\begin{array}{c} 00\\ 20068(97.3\%)*\\ 69(0.3\%)\\ 93(0.5\%)\\ 46(0.2\%)\\ 19(0.1\%)\\ 284(1.4\%)\\ 38(0.2\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\end{array}$ | $\begin{array}{c} 01\\ 70(34.8\%)\\ 124(61.7\%)*\\ 1(0.5\%)\\ 0(0.0\%)\\ 1(0.5\%)\\ 2(1.0\%)\\ 3(1.5\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\end{array}$ | $\begin{array}{c} 02\\ 11(3.7\%)\\ 0(0.0\%)\\ 264(88.3\%)*\\ 5(1.7\%)\\ 0(0.0\%)\\ 8(2.7\%)\\ 6(2.0\%)\\ 0(0.0\%)\\ 5(1.7\%)\\ 0(0.0\%)\\ 5(1.7\%)\\ 0(0.0\%)\\ \end{array}$ | $\begin{array}{c} 03\\ 14(18.4\%)\\ 1(1.3\%)\\ 23(30.3\%)\\ 35(46.1\%)*\\ 0(0.0\%)\\ 2(2.6\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 1(1.3\%)\\ 0(0.0\%)\\ 1(0.0\%)\end{array}$ | $\begin{array}{c} 04\\ 12(23.5\%)\\ 1(2.0\%)\\ 4(7.8\%)\\ 0(0.0\%)\\ 33(64.7\%)*\\ 1(2.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\end{array}$ |
|--|---|--|--|--|--|
| | $\begin{array}{c} 05\\ 106(35.2\%)\\ 0(0.0\%)\\ 9(3.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 182(60.5\%)*\\ 0(0.0\%)\\ 182(60.5\%)\\ 182(60.5\%)\\ 182(60.5\%)\\ 182(60.5\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\end{array}$ | $\begin{array}{c} 06\\ 5(3.5\%)\\ 0(0.0\%)\\ 18(12.5\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 0(0.0\%)\\ 120(83.3\%)*\\ 0(0.0\%)\\ 1(0.7\%)\\ 1(0.7\%)\\ 0(0.0\%)\end{array}$ | $\begin{array}{c} 07\\7(53.8\%)\\1(&7.7\%)\\0(&0.0\%)\\0(&0.0\%)\\4(30.8\%)\\0(&0.0\%)\\0(&0.0\%)\\0(&0.0\%)\\0(&0.0\%)\\0(&0.0\%)\\1(&7.7\%)\\0(&0.0\%)\end{array}$ | $\begin{array}{c} 08\\ 0(\ 0.0\%)\\ 0(\ 0.0\%)\\ 9(11.3\%)\\ 0(\ 0.0\%)\\ 0(\ 0.0\%)\\ 30(37.5\%)\\ 0(\ 0.0\%)\\ 0(\ 0.0\%)\\ 41(51.2\%)*\\ 0(\ 0.0\%)\end{array}$ | $\begin{array}{c} 09\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 4(& 100\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%)\\ 0(& 0.0\%) * \end{array}$ |





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 |





GEDI Integration and Enhancements

• Demo of Version 2.0.2

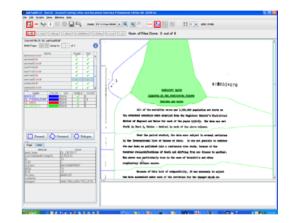
| 💥 SOO3BIN.TIF - DocLib - Gra | undTruthing Editor and Doc | cument Interface Professional Edition (DL-GEDIF | ro) |
|--|---------------------------------------|--|--|
| <u>File Edit Sc</u> ripts <u>V</u> iew <u>W</u> in | dow <u>H</u> elp | | |
| | Zoom | t: Fit To Window ▼ 🔍 🔍 🕅 | |
| 0. TIF 1. Line 2. Merge | 3. Word 4. SplitWord 5 | 5. Refine 6. Lock 7. QC Num. of Files D | one: 0 out |
| Current File: S003BIN.TIF | | | n and zeros defense of a |
| Multi-Page: Name Name S007BIN.TIF | i f v of 1 | Briteria de la constante | Proc. 1. The set of web of the Control of the set of the Set of the neutral sectors in sectors in the set of the intervention of the operation of the sectors in the set of the sectors intervention of the sector intervention of the sector in the sectors in the sectors. |
| S003BIN.TIF | · · · · | 📜 🕴 🧹 🦳 bid a legan génerit a inferen | e provid membrane de alla Clegit d'art realla |
| V001BIN.TIF | | anarte la of Linne roll the Periodor | In the theory of the second sec |
| NAME COLOR DL_INCORRECT DL_CORRECT | KEY VISIBLE COUN None 4 None 16 | Active of the second sec | The star provided in the star provided in the star star provided in the star star star star star star star star star |
| Attribute | Value | di Balah wanti Andreado da | adilia - e et actor, petra travecta a metilizzaria a |
| gedi_type | DL_INCORRECT | For the second secon | o dava te taolot e ele: nem o la jeter Vantano ella trefa Vantano ella trefa |
| (row,col)(width,height) | (157,101)(677,921) | polizzačeg doble musica in the periodic of database of the dat | Contained (Charles Holds) |
| id GT type | 000 halftone | - | |
| RS type | drawing | | |
| contents | | | |
| zoneType | non-text | | |
| | | | |





Agenda

- > Review of Goals
- > Prograss on:
 - Daiaseis
 - Evaluation Methodology



- Segmentation Survey and Tools
- Open Discussion of Additional Plans





Survey of Page Segmentation and Evaluation Algorithms

Mudit Agrawal David Doermann





Page Segmentation Algorithms

• Geometric

Dividing document into homogenous zones

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





Focus

- Identify the primary segmentation Algorithms
 Quick overview of each
- Identify likely candidates for Segmentation of Anfal Data
- NOTE:

Anfal type line finding is supported by MadCat....





Geometric Page Segmentation

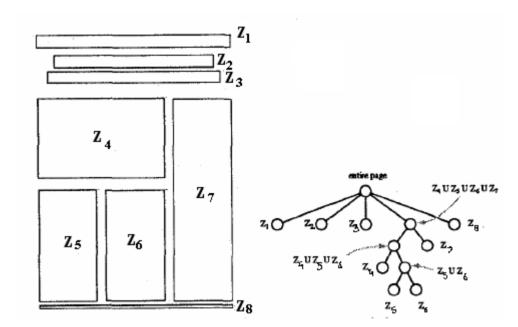
- X-Y cuts
- Smearing
- Whitespace Analysis
- Constrained Text-Line Detection
- Docstrum
- Voronoi based





Recursive X-Y cuts

- At each step, the pixel projection profiles are calculated in both horizontal and vertical directions
- Zone division is performed at most prominent valley in either projection profile
- Process is repeated recursively until no sufficient wide valleys are left in either profile







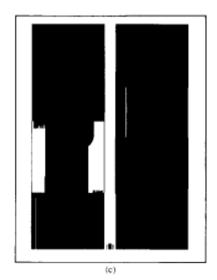
Smearing

- (a) Original Image
- (b) (c) Smearing in Horizontal & Vertical Directions with different Thresholds



(a)





- (d) Combining using AND operation
- (e) Text regions









Whitespace Analysis

- Find a set of maximal white rectangles (covers)
- Covers are sorted by

 $K(c) = \sqrt{\operatorname{area}(c) * W(|log_2(\operatorname{height}(c)/\operatorname{width}(c))|)}$

- Weighing function assigns higher weights to tall and long rectangles
- Covers are combined one by one (as per their weights)
- A segmentation is the uncovered area left by the union of the covers combined so far





Constrained Text-Line Detection

- Only needs to find a list of obstacles that lines of text do not cross
- Obstacles = gutters, e.g. figures or thin vertical lines
- Tall whitespace rectangles, column separators are candidates for gutters
- Using a robust least square method, contribution of each character to the overall match score of a text-line is penalized by the square of the distance of the alignment point from the base line





Docstrum

- Connected components are separated into two groups (using size ratio factor $\rm f_{\rm d})$
 - Dominant characters
 - Characters in titles and section headings
- For each connected component, K nearest neighbors are found
- Text-lines are computed using transitive closure on within-line nearest neighbor pairings (threshold f_t)
- Text-lines are merged using parallel and perpendicular distance thresholds to form blocks





Voronoi Based Segmentation

- Based on iterative removal of partitions
- Can be trained
- Can be extended to consider context
- Can be made robust to noise





Options for Arabic?

- X-Y cuts
- Smearing
- Whitespace Analysis
- Constrained Text-Line
- Docstrum
- Voronoi based

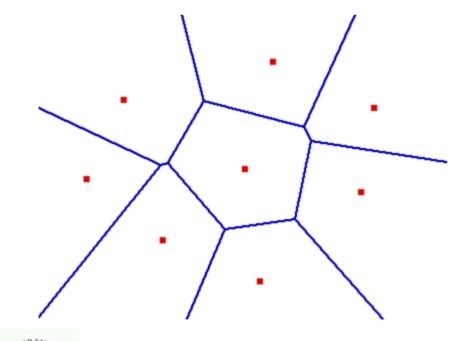
Layout too Complex Layout too Complex Noisy More Types of Zones Zone Overlap Maybe





Step 1

Point Voronoi Diagram



Voronoi Region of point p_i

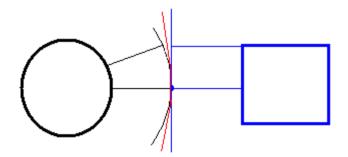
 $V(p_i) = \{ p \mid d(p, p_i) \le d(p, p_j), \forall j \neq i \}$





Step 2

Area Voronoi Diagram



Voronoi Region of area g_i

 $V(g_i) = \{ p \mid d(p, g_i) \le d(p, g_j), \forall j \ne i \}$ where $d(p, g_i) = \min d(p, g_i)$

 $d(p,g_i) = \min_{q \in g_i} d(p,q)$





- Area Voronoi approximation using Point Voronoi diagram:
 - P_i = {p_{i1},... p_{im}} be a set of points lying on the boundary of a figure g_i
 - Senerate point voronoi from generators $P = P_1 U P_2$... U P_n
 - For all *i,j,k* delete voronoi edges from points of same figure, i.e. p_{ij} and p_{ik}





Procedure

- Labeling
- Border Following
- Sampling rate [sr]
- Create area voronoi diagram using sampled points
- Select appropriate Voronoi edges
 - Min distance
 - Area ratio





Features for selection

• Min Distance

 $d(E) = \min_{1 \le i \le m} d(p_i, q_i)$ where $p_i \& q_i$ are pair of points constituting ith edge between CCs

• Area Ratio

 $a_r(E) = \frac{\text{max of areas of } 2 \text{ CCs}}{\text{min of areas of } 2 \text{ CCs}}$





• Delete an edge if

$$-d(E)/T_{d1} < 1$$

$$-d(E)/T_{d2} + a_r(E)/T_a < 1$$

where $T_{d1} < T_{d2}$





Parameters

| Parameter | Description | Sensitive (Y/N)? |
|-----------|-----------------------|------------------|
| sr | Sampling rate | Y |
| nm | Size Th on noise CC | Y |
| Ch | CC height Th | N |
| Cw | CC width Th | Ν |
| Cr | CC aspect ratio Th | N |
| Az | Min area Th of a zone | Ν |
| Br | Max aspect ratio Th | N |
| SW | Smoothing window | N |
| Td1 | Inter char Th1 | Y |
| Td2 | Inter char Th2 | Y |
| Та | Area ratio Th | Y |











Error Measurements & Metric Definitions

- Ground-truth data had only text-line blocks
- Three types of textline based error metrics
 - Ground-truth textlines that are missed
 - GT textlines whose bounding box is *split*
 - GT textlines that are horizontally merged

$$\rho(I,G,R) = \frac{\#\mathcal{L} - \#\{C_L \cup S_L \cup M_L\}}{\#\mathcal{L}}.$$

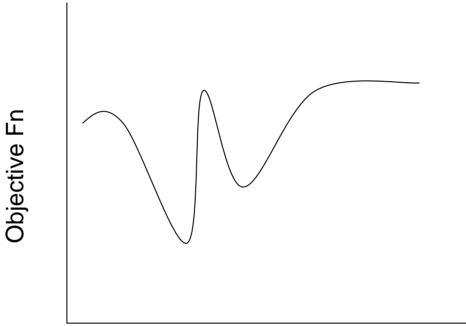
where

 C_L missed S_L split M_L merged





Training of Page Segmentation Algorithms



Algo parameters





Objective Function

Minimizing the objective function:

$$f(\mathbf{p}^A; \mathcal{T}, A, \rho) = \frac{1}{\#\mathcal{T}} \left[\sum_{(I,G)\in\mathcal{T}} 1 - \rho(G, Seg_A(I, \mathbf{p}^A)) \right]$$

where

p^A is parameter vector for A

A is segmentation algorithm

 τ is a training dataset

 ρ is performance metric as textline accuracy

I is document image

G is ground - truth





Percentage of different types of errors made by each algorithm

| | Default parameters | | | Optimized parameters | | |
|------------|--------------------|-------|--------|----------------------|-------|--------|
| Algorithm | Split | Merge | Missed | Split | Merge | Missed |
| Dummy | 0.0 | 65.5 | 0.0 | 0.0 | 65.5 | 0.0 |
| X-Y cut | 5.6 | 7.8 | 0.4 | 5.6 | 7.8 | 0.4 |
| Smearing | 3.8 | 1.0 | 5.7 | 3.8 | 1.0 | 5.7 |
| Whitespace | 6.6 | 1.3 | 0.0 | 5.0 | 2.6 | 0.0 |
| Text-line | 5.1 | 1.3 | 0.2 | 5.1 | 1.3 | 0.2 |
| Docstrum | 4.5 | 9.0 | 0.0 | 2.5 | 3.6 | 0.01 |
| Voronoi | 4.9 | 0.8 | 0.02 | 2.9 | 1.3 | 0.02 |





Challenges in Handwriting Documents

- 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
- Regions not rectangular in nature, hence bounding box may not be the best representation





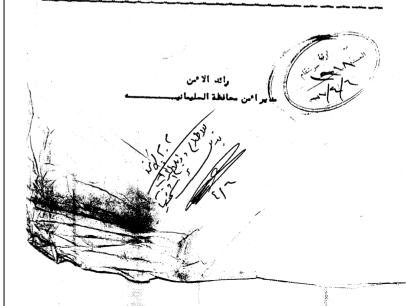
سى /كسافة النعاونيسسات. سن /أمسن السليدانية / السياسية .

11 < 9 11 0 -----

بتا^مريخ ٢/٢/٤/٢ . قام السيد المام بزيارة الى عديرتنا (.) وقد طلب شخصياً *الإلكار* تحياته وتننياته لجبيع منتسبي د اقرتنا واجياً لهم العوفقيه والسداد في الألهم (.) تغن تو توجيهات السيد العام ما يلي -

- ٢ التروته والسلامة في التعلمل مع المواطنين مع ضرورة الضرب يشده على الغارجيمن عن القالون (والمغلين بالا^من •
 - ٣ ، ضرورة ا^ويجاد. ا^وقضل الملاقق معتمدين الا^وساليب الحديثه بين منتسبي الجهاز ورعاية. المارون وساع شكواهم ومعالجتها بشكل مباش<mark>ر،</mark> مع التقيد بالفيط العسكرى ،
 - ع ، التحرك وفق سياق الخط المياسي العام للقياده السياسيه في المنطقة الشماليه ه
- و ه تعبيق الثقام بين الجناهير والجهاز من غلال التعامل الاحساني مع الجناهير ، هذا وقد إحك السيد العام رهاية العامليين في منطقة الحكم الذاتي رهاية غاصه ولرفع ستواها من كافة الوجوه .

للعمل بموجوبة توجيبهات السيد العام بكل دقة (م) مُكور الأمن منطقة الحكم الذا تي للتفضل . بالعلم رجسسيسسان".



Α Ν F Α

D A T A

الليف ليدت تما لغة الان بورك كاتم سير معد المعديل وسيمه المعنية حدكة مترعد فتحاجه ورون علم منا مع ان اس ت عديت صافية يعنى مسي فارت لمن الرته الم است رأ لاطام الم في 29 من ج ويدلاله ال الى ف ت ف ترن رمول بدى ت بال معلم مرم مرد مرد بر مرد مرم مرم المع مهرف صفران مست - بد ، سور، الدر) - ب بردی باز بر ب ار' رانس and best

امرا و لاي

سى /كندافة المعاونيسسات.//+ بن رامين الملهد اليه / الامهامية 11 care 110 يتَأَثَرُنُ * / 6 / ١٩٢٤ ، قام السند السام يزيارة إلى تديرتنا (.) رتد طَّلَب (. شمراً الإلائُمُ المعيانة وتنابياته لجنهم ملتسبي فاافرقنا واجها آلهم الموققم والسفاف في ألكهم (-) كامن وال ورجيها درائسيد العام بارش م و . صرورة قيم جهاز الاخراقي ملطقة الحكم الذاتي لميماته العالفة وكولوة اقتداعل بن الموليقيني وغورة الظهد الجهاز طى كافة المنتهات بنا يغمن هذا الطيم وسنالجة النواتف بلغس أنجر والأ ج د التريد والعلامة في الثمليان بن التواطئين بن ضرورة الغرب يشده على الشارجين هن التأون (والمغلين بالاحن و م يا ضرورة الهجاب الأصل الحلائق معقد بن الا اسالهب الحديثة بين شكسي الحربار ورناية . المارون وسباو غكواهم وبعالجتها يشكل جاغوا مج الظهد بالغيط الاستكرىء ي . التمرك وفق سياق الحاد السياسي العام لللباده السياسية في النقطة الشنائية ه هو ، تعنيق الله بين الساهير والجهاز بن علال اللعاس الاحساني مع المناهير ، هذا وك الأك المبدد العام رفاية الماطبين في متطلا العكم اللذاتي رهاية عامه ولران منتزقما من كانة الرجور م للعمل بموميه فوببينات السيد أقعام بكل والا (،) سكور المن مطط العكم اللا تي للطفل بالعلم رجسيسا ال راي الاحن

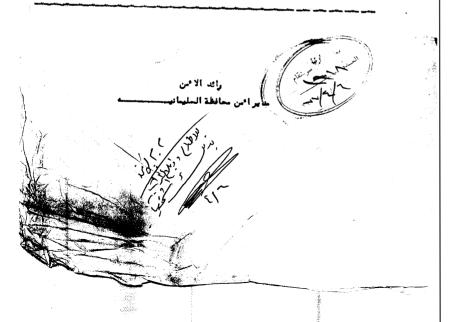
لسبى /كسبافة المعاونيسساته. سببين /أمسين السليم اليه / السياميه .

11 < a vs 11

بتا"ريخ ٢ /٤ / ١٩٢٢ ، قام السيد المام بزيارة الى مديرتنا (،) وقد طلب شخصياً "بلائكم تعياته وتنتياته لجبيع منتسبي د اقرتنا راجياً ليم الموفقيه والمداد في الألبم (،) تغن يُ توجيبات السيد العام ما يلي م

- ٢ ضرورة فهم جهاز الا "من في منطقة الحكم الذاتي لمهماته الحاليه وكيفية التعامل مع المواطنيني" وضورة تتقيف الجهاز على كافة المستويات بما يضمن هذا التفهم ومعالجة المواقف بنفساً في • ٥ • المروته والسلامه في التعلمل مع المواطنين مع ضرورة الضرب بشده على الغارجيمن عن القاون (
 - والمخلين بالا°من ه
 - ٣ ضرورة اليجاد الأضل الملاقق معتندين الا "ساليب الحديثه بين منتسبي الجهاز ورعاية المارون وساع شكواهم ومعالجتها بشكل مباشور معالتقيد بالفيط العسكرى -
 - ع ، التحرك وفق سياق الخط المياسي العام للقياده السياسيه في المنطقة الشماليه ،
- ه م تعميق الثقة بين الجناهير والجهاز من غلال التعامل الا "نسائي مع الجناهير، هذا وقد اكد السيد المام رواية المامليين في منطقة الحكم الذاتي رواية غامه ولرفع ستواها من كافة الوجود .

للممل بموجهة توجيبيات السيد العام بكل دقه (،) مكرر أ^ومن منطقة الحكم الذا تي للتفضل بالعلم رجسيسيسسا^و.



بى ركسانة (لمعاونيسنات) بيبين والسبن السابيدانية والأمواعية و 11 CONS! متا من ۲۰۷۲، أو قام السبد الحام بهارة التي تدبيرها الإان وها طلب تسميم البلائكم همياته وصبائه لجبع منتسبي أواقرتنا واجزاكهم الموقور والسداد في الأسالهم و ، و علين فُرّ . توجيهات كالنياد، كالحام با يلن 🔨 ر ر شريرة فيم جنها إ الا "تراقي التلاقة التلكم الله أي المينات الحاليد وكلفية التمامل مع الترقيقيني) وشورة تثقيف الجمار على كافة الممتهات ينا يشمن هذا التلبيم وحالجة النواف يتغنى الرواني ال ج . البرود والسلامه في التمليل بر البرا فون بر ضرورة الغرب بقده على الغار همر، من العاون ﴿ والملين بالاان د اج . جزيرة الإيباد الانتال العلاقق معتندين الاخباليب الندباريد بنن منشيق الجينار وربابة البارون وساع فكراهم وحالمتها بشكل جاشوا معالتقها بالنيط المبكري. ج يُ التحراد وفق سباق الترط السيامي الحام للقامة السرامود في التفقة الشناك م اً تسبق الثان بين الجناهير والجياز بن غلال التعامل إلا الساني مع الساهير . عدًا وف إلايا البنيد العام رفاية العاطيين في منطقا الحكم الذائي ومانة غاهه ولرفع منتزاها س كابة ألوجوه للمحل بموجود توجيبات السيد العام بكل دته وروار عورسطانا العكم الفاتي للتعمل بالعلم رجسما



بن رامسن العليد اليه / الامهامية

11 Care 1/ -----

Α

Ν

F

Α

D

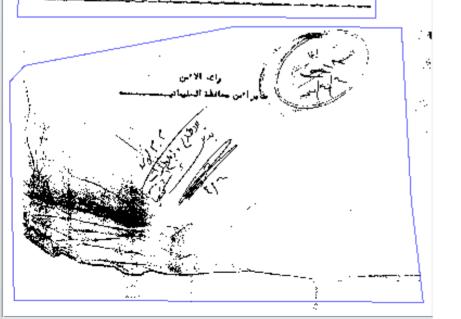
Α

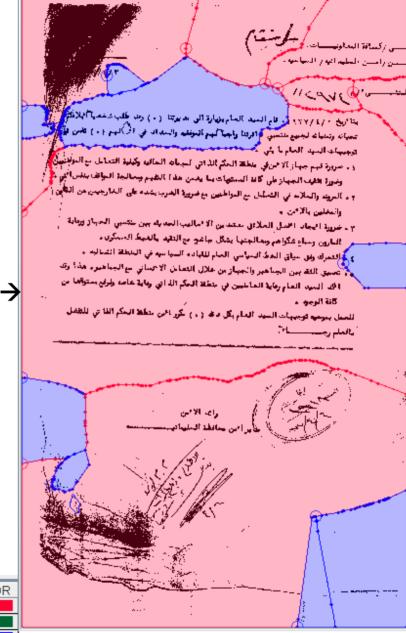
Т

Α

- يتاكريغ / ١٩٢٧/٢٤، قام المعد المام يزيارة إلى مديرتنا. (،) وند طلب شنعها الإنكام تحيانه وتبلياته لجنيع ملتمين د اقرائنا واجها آميم الموققية والمداد. في أن أليم { ،) كانيم { ،) كانين قو توجيها تد المياد المام با شي ،
- ي . صرورة قيم جياز الاحراقي علطتة الحكم الذاتي لميماته الماله، وكبلية اقتماعان بالتولينيين وضرورة والهان الجياز على كافة المستنهات بما يخمن هذا التقيم وسمالجة المواتف بلغان أخي ذ الحق
- جاد البريند والملاب في التمليل بع المراطنين بع ضرورة الضرب يشده على الشارجيني من الطَّيَّن (والمغلين بالا^{رم}ن ب
 - ٣ خزرية الهجاب الأصل الحلائل معقد بن الالماليب الحديلة بين عقسي الحياز ورداية. المارون وساره شكراهم وبعالجتها يشكل جاغوا مع التقد بالغيط المعكري ه
 - ي . التمرأة وفق سياق الحطَّ السياسي الحام لللباده السياسية في البلطاة الشنالية -
- يو ، تصبق الله بين المناهير والجنباز بن علال التعالن الاحساني مع المناهير ، هذا وقد الآك البيد العام رفاية الماطيين في منطقة المكر الله ابن وعاية عامه ولراي ستوقعا من كانة الرجود ،

للعمل بموجهة فرميهات السيد العام يكل دائد (،) أكرر الاين مطط المكم الله في للقضل رالعلم رجيب مسياحة







← GT

Evaluation \rightarrow

| NAME | COLOR |
|--------------|-------|
| MATCHED | |
| DL_FASEALRAM | |
| FALSEALRAM | |
| DL_DETECT | |
| DL_MISS | |
| MARKUP | |
| MISSED | |





| 0. TIF 1. Line 2. Merge 3. Wor | d 4. Split | | n: Fit To Window 👻 🗨 🍳 🖤 🕂 🛅 🎰 5. Refine 6. Lock 7. QC Num. of Files Done: 0 out |
|--|--|--|--|
| Current File: aah1aa00.tif Multi-Page: A Jump to: 1 | of 1 | | |
| Name | Image | Xml | |
| aah1aa00.tif | Image: A set of the set of the | Image: A set of the set of the | A 183024070 |
| aah27e00.tif | Image: A second s | Image: A second s | Koratre bata |
| aah28e00.tif | Image: A set of the set of the | Image: A second s | - PULSES & TU REPORTS LATER |
| aah33e00.tif | Image: A set of the set of the | Image: A start of the start of | All of the metallity many per 1,000,000 peptieties are forth on |
| aah42d00.tif | Image: A second s | Image: A start of the start of | Can stracted actability wave completed area for any new of the system in any action in being of Ingliant and Tables for such of the system 100-50. So date was are: |
| aah45c00.tif | | X | 🗮 ford in Derr 7, fullet a self al l'in east of the global states - |
| aah45f00.tif | Image: A second s | ~ | Could have not to 2 obtailts, the data ware subject to a sevenil contribution to the Totarenational High of Gauges of Santh. The man one possible to problem |
| aah5aa00.tif | ✓ | Image: A start of the start of | The West Area, and payorithe of the in a matching on him windly. Despine of the manyour semilurativities of death and dividing from one demonstra vanishes. |
| aah61f00 tif | | ~ | In stort manufactures in our of an actual process of starts to our of the stort of the stor |
| NAME COLOR KEY | VISIBLE | COU | And |
| | | 14 | We have an excellent water rank of the restations for the simple which re- |
| DL_MISS None | ✓ | 0 | minimi force and, minimpane realized. With empiryous or the strategy of the |
| DL_DETECT None | ✓ | 5 | latina Suntes, realamété asu ne vasi nam ar Santa era vas portires |
| DL_MATCH None | | 0 | the hands (he want operation) offer each new periodical Benefitikae, a staty was accessing of the fastile as recorded the the sport capitable part unlike |
| | | | "Web Proclassifiawides for look its own and oil verticing. Supershilling whet the source of darks the same and new and the source of darks the same of darks |
| | | | arminica by the matter easilyted at: that each takes the system projectory. |
| | | | The surparyields of Weak, of the bas in the cost of their character for the T. D. Downrowski, is a measure of the cot affant of damps in characteristics, |
| | | | |
| Normal Oriented | О Р | olygon | - ad |
| | | | |

R.

ny Research Laboratory

NERSIN

Zone Classification

Summary of Results

Total Number of Sample : 21786Overall 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% |





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(0.0응) | 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(0.0응) | 0(0.0%) |
| 07 | 0(0.0%) | 0(0.0%) | 0(0.0%) | 0(0.0응) | 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(0.0응) | 0(0.0%) |
| | | | | | |
| 06 07 08 | 38(0.2%) 0(0.0%) 0(0.0%) | 3(1.5%) 0(0.0%) 0(0.0%) | 6(2.0%) 0(0.0%) 5(1.7%) | 0(0.0%) 0(0.0%) 1(1.3%) | 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(0.0응) | 0(0.0%) | 0(0.0응) |
| 0(0.0%) | 0(0.0%) | 0(0.0응)* | 0(0.0%) | 0(0.0응) |
| 4(1.3%) | 1(0.7%) | 1(7.7%) | 41(51.2%)* | 0(0.0%) |
| 0(0.0%) | 0(0.0%) | 0(0.0응) | 0(0.0응) | 0(0.0응)* |



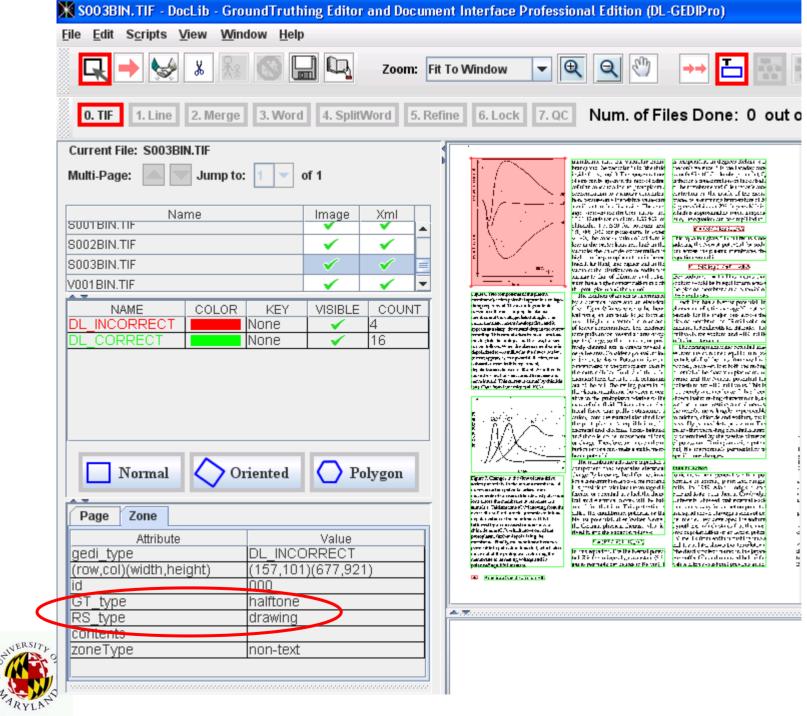


03. Precision and Recall

| 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 |









Remaining Tasks

- Evaluation of Existing Data
- Sponsor testing of software
- Integration of OCR evaluation
- Feedback from MADCAT Participants





Recent Deliverables

- 26,000 page Tobacco Litigation Corpus
- Full Presentation of July 20th
- Software for Classification and Segmentation Evaluation



