

BOBCAT Review

July 10, 2008

David Doermann Wontaek Seo Elena Zotkina Mudit Agrawal





Agenda

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





Overview of 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





Image: Speed	AAW_ARB_20070101.0003_1_LDC0005.fit - DocLib - Graun	dTruthing Editor and Document Interface Progessional Edition (DL-GEDIPro)
Network	le Edit Scripts View Window Help	
Image: State Nume of Files Done: 0 out of 5 Image: State Image:	□ → ₩ ¥ 10 00 □ 00 Zaone Fit T	To Page Wath 💌 🚇 🔍 🔲 🛶 🎦 💷 🚳 🗮 🚦 1 💌 (474, 5611)
Image: Interp:		
Concerter Field 1/56: AMM_ARE_200709304000_1_LDC00001 Anis Page: If Image: Amm_ARE_200709304000_1_LDC000011 Image: Amm_ARE_200709304000_1_LDC000011 Max ARE_200701010000_1_LDC000011 Image: Amm_ARE_20070101000_1_LDC000011 Max ARE_20070101000_1_LDC000011 Image: Amm_ARE_20070101000_1_LDC000011 Max ARE_2007010100_1_LDC000011 Image: Amm_ARE_200701000_1_LDC000011 Max ARE_20070101000_1_LDC000011 Image: Amm_ARE_200701000_1_LDC000011 Max ARE_200701000_1_LDC00011 Image: Amm_ARE_	0.11F 1.Line 2.Merge 3.Wort 4.SpitWood 5.Belle	a latek 7.00 Num of Files Done: 0 out of 5
And Page: if i	Contract Contraction Contraction Contraction Contraction	and Encounted Section
http://www.web.yength 0.002252.3243)	Current File (3/5): ANW_ARB_20070101.0003_1_LDC0005.W	
Home	Mutti-Page: 🔤 🔤 Jump te: 📋 👻 ef 1 🛛 🚺	
Wy JHR. 2007101 NDD. 1. LOC00011 V WY JHR. 2007101 NDD. 1. LOC000111 V WY JHR. 2007111 V WY	Name Image 30ml	
AVJ. #R. 20070101.0001_1_LCC0004# V AVJ. #R. 20070101.0001_1_LCC0007# V Image: State of the sta	AW_ARB_20070101.0003_1_LDC0002.1f 🗸 🖌	a agus als 1 cle ages is
W.J. #B. 20070101 000.1 LCC00011 V W.J. #B. 20070101 000.1 LCC00011 V W.J. #B. 20070101 000.1 LCC00011 V Image: Contract C	AW_ARB_20070101.0003_1_LDC00041# 🖌 🖌	
W. JAB. 200701011.000.1 LDC 000 01 // × × Number 1000.1 LDC 000 01 // × ×	AW_ARB_20070101.0003_1_LDC00051f	Contraction of the second s
Nume COLOR NEV VISELE COUNT Nume COLOR NEV VISELE COUNT Callo Nore 28 Name Color Color Callo Nore 28 Color Color Color Color Callo Nore 28 Color Color <td>W_AHB_20070101.0003_1_LDC0007.0F</td> <td>- Con the New Willie</td>	W_AHB_20070101.0003_1_LDC0007.0F	- Con the New Willie
News Colonted Debygen News Colonted Pobygen Minute Displayer Attrive Displayer	alaurian and the state of the s	
Name COLOR KEY NSELE COLAT Name Nove 2 S MATCH Nove S Name Oriented Polygen Name Convertee Name Convertee Name Oriented Polygen Name Convertee Name Convertee Name Oriented Polygen Name Convertee Name Convertee Name Oriented Polygen Name Convertee Convertee Convertee Polygen Zee Name Convertee Convertee Convertee Convertee Polygen Name Name Convertee Convertee Convertee Convertee Polygen Name Convertee Convertee Convertee Convertee Convertee Convertee Nalpe<		Core of the
Nemal Oriented Pelygen Press Attrific Value attrific Value		
New E COLOR FEY WBELE COUNT ASCAL DAM WOYE 28 MATCH WOYE 28 MATCH WOYE 11 Neamal Orionted Pekygen Name Control ted Pekygen		in the second se
Nermal Orionted Palygen Nermal Orionted Palygen Attribute Contract Contract Attribute Contract Contract <td< td=""><td>NAME COLOR KEY WEBLE COUNT</td><td></td></td<>	NAME COLOR KEY WEBLE COUNT	
MATCH None II Nermal Oriented Pelygen Afriture Value Afriture Value Afriture Value Oriented Pelygen Afriture Value Oriented Pelygen Afriture Value Oriented Value Oriented Pelygen Afriture Value Oriented Oriented Oriented <td>None 28</td> <td>- we where a colored</td>	None 28	- we where a colored
Nermal Oriented Polygen Nermal Oriented Polygen Attrivic X. PAGE all rows/width/reight) (0.0)(2552,3248)	MATCH None V 11	
Nerneal Orionted Polygen Nerneal Orionted Polygen Atrial Atrial Chape Zueno Atrial CL. PAGE DI JOW J(Width) reight) (B. B)(2552, 3249)		
Nermal Oriented Polygen Nermal Oriented Oriented Nermal Oriented Oriented Nermal Oriented Oriented Nermal Oriented Oriented Oriented Oriented Oriented <td></td> <td>the man and the second second</td>		the man and the second
Nermal Oriented Polygen Nermal Oriented Polygen Attritute Oriented Oriented Oriented Oriented Oriented Oriented Oriented Oriented Oriented Oriented Oriented Or		all asked in these Main and and and a
Nermal Orionted Pelygen Nermal Orionted Pelygen Attrake Company (width /reight) Company (width /reight) (0,0)(2352,3249) Orionted Company (width /reight)		at a first of the first
Nermal Oriented Polygen Attribute Attribute Vilue cit. Lype C. PAGE ol/rowl(wkth/height) (0,0)(2552,3249)		Amile sola l'Aus - har a famely the
Nermal Oriented Polygen Attraite Value Attraite Value OL PAGE Value OL PAGE Value OL PAGE Value University Value Value Value </td <td></td> <td>7 low land and and in the court</td>		7 low land and and in the court
Ange Zane Attrive Value cd. type DL. PAGE al rowj(width) reight) (0.0)(2552,3249)	Nermal Oriented Polygen	
Attribute Value Attribute Value id. type DL. PAGE ol.row](width height) (0.0)(2552,3249)		in 1 - the states and there and it is
Attriute Value di <u>type DL PAGE</u> al rowj(width /height) (0.0)(2552,3249)	ago Zuno	
al row (with height) (0,0) 2552,3249)	Atribute Value	- is in the second s
المحملة والمسروجا لعلله المحمد المحملة والمسروجا المحملة المحمد	al/ow)(width/height) (0,0)(2552,3249)	to seleculture as
		the lo - ile the will
		the second s





Data Sets

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

the 2? fatel alold Dana ance lente - From Hometo bet - specially in Praceto Cardoby; pedinoray doeano. endocrimitogy & metablin; while infections doean; Tak Proja or Kinsten Had Jouth Phile main :





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.





Agenda

- > Review of Goals
- > Progress on:
 - Datasets
 - 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
 - $\mathsf{L}(\mathsf{i}) \neq \mathsf{L}(\mathsf{j})$
- 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)≠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 TEXTLINEGT		_	
[X]	17,	DL_TEXTL INEGT			
[OVER]	LL] 9,	/0/8/17, 52.94%			

AAW_ARB_20070101.0003_1_LDC0004.tif





```
_____
Summary of Results
_____
- Total number of R-Zone : 22033
- Accuracy of Zone Detecting : 31.19%
01. Information on Zones
_____
Label Class of Zone Number of Zone Accuracy
     _____
1 DL TEXTLINEGT
                   22033 31.19%
02. Confusion Matrix
_____
Result\GT unmatch 1
   -----
  unmatch 0(0.0%) * 12778(65.0%)
     1 15161(68.8%) 6872(31.2%)*
03. Result Table
_____
Label Total Detected Correct Precsion Recall F-Score Missing FalseAlram
  1 19650 22033 6872 31.19% 34.97% 32.97% 65.03% 68.81%
```





Zone Classification

=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=		=	
S	u	m	m	а	r	У		0	f		R	е	s	u	1	t	2	5	
=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=		=	

- 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%





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	0.6	0.7	0.8	0.9
	106(35,2%)	5(3,5%)	7(53,8%)		0(0,08)
	0(0.08)	0(0.08)	1(7,7%)	0(0.08)	0(0.08)
	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(0.0%)
	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%)*





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







Agenda

- > Review of Goals
- > Progress on:
 - Datasets
 - 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

rectang, team die die pale part wi wie die 10 die best die bester die 135 in 12 wei degte weise besterde 136 in 12 wei degte weise besterde

On some way to day to taken a source the first to bill as held to their bill instant. Remon Lib de alganizer de

(a)

- **Original Image** (a)
- (b) (c) Smearing in Horizontal & Vertical **Directions with** different Thresholds
- Combining using AND operation (d)
- (e) Text regions





(d)





(b)

(c)





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 f_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

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

where

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





- 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
 - > Generate 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 ithedge 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	Ν
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

	Defa	ult para	meters	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





امرا و لاي الليف فيدت بما لغة الان بورك كام سير معد المصلح الموالي المعنية حدكة مترعدة فأتته يدون علم منا م ان اس ت مدرت صافية يعنى من من راب لمن الاته الم است را برمام الم في حرم من عرب المريد ال الأف ت ما ترن (مول بلا) ت إلى مدف ب مست المري الار ، الدر ب بردین بنان ب ~ 1, 1, 1 ~ and best

بتا ريخ ٢ / ١ / ١ / ١ / ١ مام السيد المام بزيارة الى مديرتنا (.) وقد طلب شخصيا المِلاَعُكُم تعياته وتنتياته لجنيع منتسبي د اقرتنا راجياً ليم الموفقيه والمداد في افْ البم (،) تغن تو توجيهات السيد العام ما يلي -

سى /كسافة المعاونيسسات، مستقلم سن /أمسن السليدانيد / السياسيد .

// < حرم (// مسمن

Α

Ν

F

Α

D

Α

Т

Α

- ٢ العروك والسلامة في التعلمل مع المواطئين مع ضرورة الضرب بشده على الخارجين عن الثاني (والمغلين بالا من ه
 - ٣ ضرورة ا^ويجاد الأضل العلاقق معتمدين الا^عماليب الحديثه بين منتسبي الجهاز ورقابة العارون وساع شكواهم ومعالجتها بشكل مهاشير معالتقيد بالضيط العسكرى -
 - ع . التمرك وفق سياق الغط المياسي العام للقياده المياسيه في النطقة الشناليه .
- ج م تعبيق الثقام بين الجناهير والجهاز من غلال التعامل الا "نساني مع الجناهير ، هذا وقد إحك السيد العام رهاية العامليين في منطقة الحكم الذاتي رهاية غاصه ولرفع ستواها من كافة الوجوم .

للعمل بموجوم توجيبهات السيد العام بكل دائه (،) مكور ا^ومن منطقة الحكم الذا تي للتفضل بالعلم رجسيسيسيا^م.

واعد الامن

ن رامسن العليم انيه / الاميامية 11 Care 120 يتَأَثَرُنُ ٢ / ١ / ١ / ١ / ١ ، قام المند المام يؤيارة إلى بديوتنا (.) رفد طَّلب شمه يأ أَيْلاَنُكُ حميات وتناجاته ليعين ماشمين بالاوليا والجاليين الموققية والساداد. في ألكانهم (-) كامن 🕌 ورجيهات السيد العام بالشيء و . صرورة قيم جهاز الاخراقي ملطقة الملم الذاتي لمبناته المالية وكماية اقتدا مل ب الموافقين وغورة الطها الجهاز على كافة المنتهات بنا يغمن هذا الطيم وسالجة النواتف بلقس أأبورنا م و العربية، والعلامة في الشمليان مع المواطنين مع ضرورة الضرب يشده على الغارجين هن الثانين (والمغلين بالاحن و م يا ضرورة اليجاب الأصل الحلائق بمقتدين الا البالهب الحديثة بين شكسي الحيار ورنابة المارون وسباو غكواهم وبعالجتها يشكل جاغوا مج الثقيد بالغيط الاستكرىء ي . التمرك وفق سياق الحكَّ السياسي العام للقياده السياسية في البلطاة الشنالية ه هو ، تعليق اللقة بين الجاهير والجهاز بن علال التعالي الاحماني مع الجاهير ، هذا وك الأك المبدد العام رفاية الماطيين في متطلة الحكم اللذاني رهاية عامه ولران منتوقعا من كانة الرجزة م للعمل بهومه أوبيهات السبد العام يكل وك (،) سكور المن منطط العكم اللاحي للطفل بالعلم رجسيسيا ال 1 ÷. واعد الاحن

بتا من ٢ / ٤ / ١٩٧٢ . قام السيد المام بزيارة الى مديرتنا (.) وقد طلب شخصيا أم لانكرا تعياته وتنتياته لجبيع منتسبي د اقرتنا راجياً ليم الموفقيه والمداد في افْــَالهم (،) تَغْنَ يُؤْ توجيهات السيد العام ما يلي م

_____/كـــافة المعاونيــــات.

- ٢ ضرورة فهم جهاز الا •من في منطقة الحكم الذاتي لمهماته الحاليه وكيفية التعامل مع الموالينين وضورة تتقيف الجهاز على كافة المستويات بما يضمن هذا التفهم وممالجة المواقف بنفس أبن م • المرونه والسلام في التعلمل مع المواطنين مع ضرورة الضرب بشده على الخارجيين عن القامن (
 - • المروده والملاحة في التعلمل مع المواطلين مع صرورة أنصرب يشده عن العدر بيدرا عن المار. والمخلين باللا°من ه
 - ٣ ضرورة الإجاب الأضل الملاقق معتندين الا اساليب الحديثة بين متنسبي الجباز ورقاية. المارون وساع شكراهم ومعالجتها بشكل مهاشرا مع التقيد بالضيط العسكرى .
 - ع ، التحرك وفق سياق الخط المهاسي العام للقياده السياسيه في المنطقة الشماليه ،
 - و ه تعبيق الثقة بين الجناهير والجهاز من غلال التعامل الا^عنساني مع الجناهيره هذا وقد إكد السيد العام رواية العامليين في منطقة الحكم الذاتي رواية غامه ولرفع ستواها من كافة الوجود ه

للممل بموجهة توجيبات السيد العام بكل دقه (٠) مكور اعن منطقة الحكم الذا تي للتفضل بالعلم رجسيسيسياء".



ليسوا ركسانة المعاوميسات سيبين والسبق السايد أنها والأمواحية و 11 Sang Ko متاريخ ۲۰۷۷ (۲۰۱ م. قام السبد الحام بهترة التي تدبيقا الإس) وها طلب تسعيها اللائكام فميات ومنبائه لجبع متعمي كوافوننا واجهاكهم المواقيه والمعادد في الأساليم (-) عامر، وإ توجههات كالنباد كالحا والاعان 🚣 و ر شرورة فهم جنها إ الا "مراقور متعادَّة الملكم كانا أي لمستانه الحاليد وكلية الدمامل مع المولِّقينُ وشورة شقف الجواز على كافة المستهات ينا يشمن هذا التلبيم وحالجة النواف يتفى البور ال ج . البروك والسلامة في التمليق بوالبوا فانون بوخيروة الغرب يقدد على الغارجمور من العادي ﴿ والمظلون والآمن م اج . حرورة الايمان الاغلاق مقتدين الاختاليب العديقة بن متنبي الجبار ورماية البارون وساع فكواهم وحالمتها بشكل جاشوا معالتقها باكتيط المبكري و ج يُرَ الشيرار: يَقِي سَبَاق الحَطَّ البِيَاسي العام للقامة السيامية في الشِعْدة الشالية ، • معيق الثان بين الجناهير والجياز من غلال التعامل إلا الساني مع المناهير. عذا وقا إلاك البنيد العام رفاية العاطيين في منطقا الحكم الذائي ومانة غامه ولرفع منتزاها س کابة الرجو ب للمعل بموجود توجيبات السيد العام بكل دقدار واراكور احن سطلا العكم الغااتي للتعمل بالعلم وجسمان ريان بالاعن

ورامسن المطيم انيه والاسهاجية 11 Care 110 يتَأَثَرُنُ ٢ / ١٩٢٤ / ١٩٩٩ ، قام السند المام يزيارة إلى تدييرتنا (.) رفد طَّلب (شمه) الإلائل حمياته وتناباته لجنيع ملتسبي بدافراننا واجها آمهم الموققم والسدان في أأكنهم (-) قامن الأرا ورجيهات السيد العام بالشيء و . صرورة قيم جهاز الاخراقي ملطلة الحكم الذاتي لميماته العالفة وكولوة التعامل بع الموليَّقِينُ وغورة الطهب الجهاز على كافة المستنهات بنا يغمن هذا التقيم وسعالجة النواتف بلغس أنوره ج د التريد والعلامة في الشمليان بع التواطئين بو ضرورة الغرب يشده على الشارجين هن التأون (والمغلين بالاحن و م يا ضرورة اليجاب الأمال الحلائق بمقتدين الا البالهب الحديثة بين شقسي الحياز ورنابة المارون وسباو غكواهم وبعالجتها يشكل جاغوا مج الثقيد بالغيط الاستكرىء ي . القمران وفق سياق الحاد المواسي العام لللباده السياسيه في البلطانة الشناليه م هو ، تعليق اللقة بين الجاهير والجهاز بن علال التعالي الاحماني مع الجاهير ، هذا وك الأك الميد العام رفاية الماطيين في متطلا العكم اللذاني رهايةً عامه ولران منتزقما من كانة الرجور م

Α

Ν

F

Α

D

Α

Т

Α

للمعل بموجهة توجيبات السيد العام يكل دائد (،) تكور الآن مطط المكم اللا تي للتقفل رالمام رجسية مسياناً.





aah1aa00.tif - DocLib - GroundTru File Edit Scripts View Window He	thing Editor and Docume In	ent Interface Professional Edition (DL-GEDIPro)
	Zoom: Fit	To Window 💌 🗨 🔍 🖑 →→ 🛅 🔛 🔤
0. TIF 1. Line 2. Merge 3. Wo	rd 4. SplitWord 5. Ref	ine 6. Lock 7. QC Num. of Files Done: 0 out of 4
Current File: aah1aa00.tif Multi-Page: Jump to: 1	of 1	
Name	Image Xml	
aah1aa00.tif	 	1183024520
aah27e00.tif	< <	KENTER DET
aah28e00.tif	✓✓	Alling at the restorements for the
aah33e00.tif	✓✓	all at the maximum over 1,000,000 peptintion are forth on
aah42d00.tif	✓	Ga straded adabta was angled and ba agreew reveals angle at a set hview of Region and Talor for such at the years 198-60. So this was not
aah45c00.tif	✓×	1 fort in New 7, Wilse a Add all Prinet of Hergins, When
aah45f00.tif	✓	Group the script charited, the fails ware addents to sweeth contribute to the Informational State of Carway of Samb. It was not possible to minimum
aah5aa00.tif	✓	The Net Anti, pl spiriture that a mathematical star state, but the state of the measure measurement in the state of the measurement of the state of
aah61f00 tif		The short was particularly how in the case of branchik's and when
NAME COLOR KEY	VISIBLE COUNT	sequences of and some
DL_FASEALRAM None	✓ 14	Ver ficht navarlicht witer erch ist site erstall mie Far file innigen sticht re-
DL_MISS None		between the president results and the present of the
		initial States, realization and rank or states or source provides a state of the post-
		The accountry of the daths as recorded for the approximations described by
		write way the angetel by dividing the subre of detin out we
		Territolica by the master analyzed on the assess pairs the spectral structure. The angustylability many, on its bad als the man of Static encycles by the
		W. D. Doversment, is a measure of the out effect of alrayer in chardflaubles.
Normal Oriented	Polygon	the second second
Page Zone		
Attribute Multiple Selections	Value 19	

Zone Classification

=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=
S	u	m	m	а	r	У		0	f		R	е	s	u	1	t	5	3
=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=	=

- 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%





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(0.0응)	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(0.0%)	0(0.0읭)

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(0.0응)	9(11.3%)	4(100%)
0(0.0%)	0(0.0%)	0(0.0응)	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(0.0응)	30(37.5%)	0(0.0%)
0(0.0%)	120(83.3%)*	0(0.0응)	0(0.0%)	0(0.0%)
0(0.0%)	0(0.0%)	0(0.0응)*	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%)*





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



