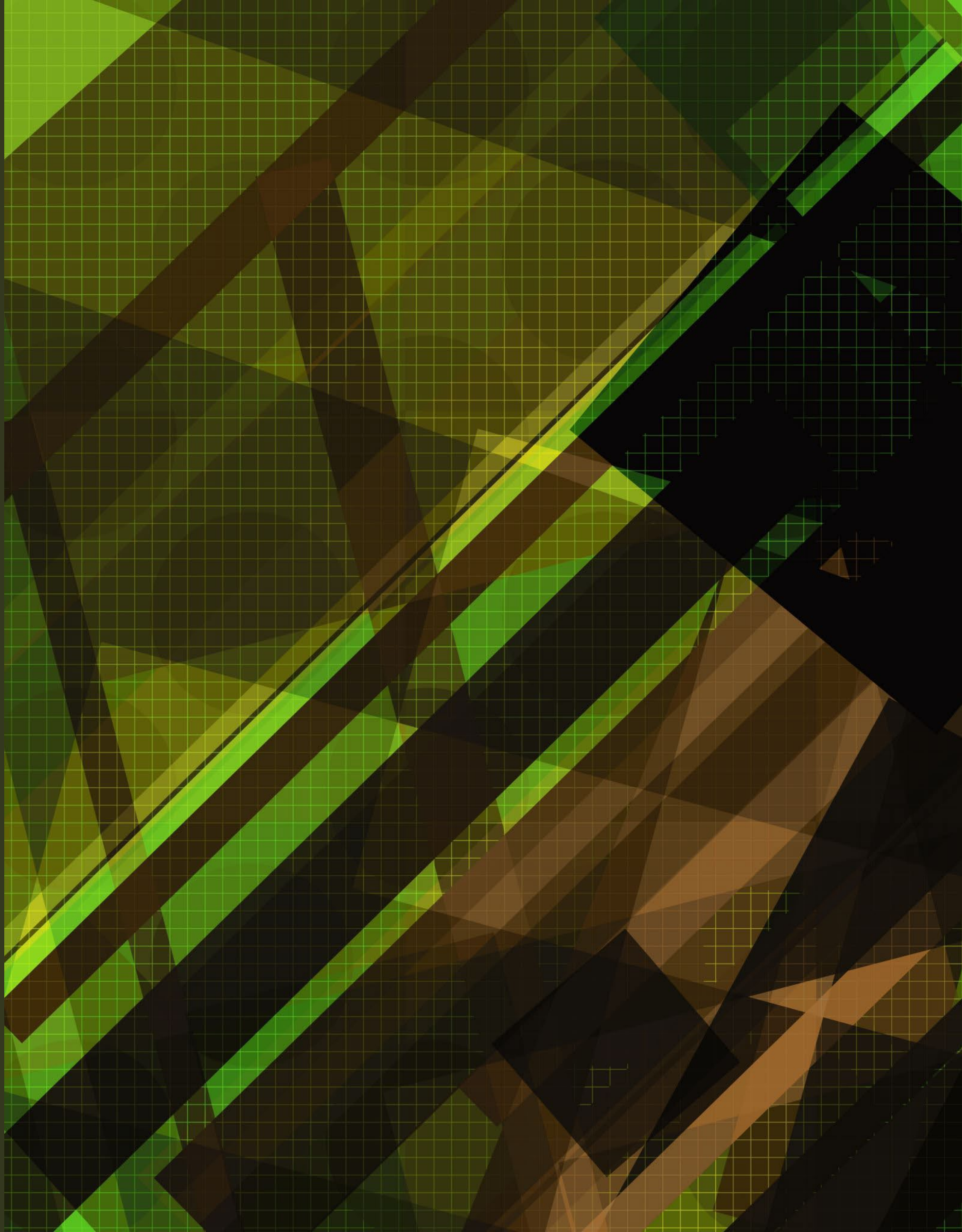
USING DEEP LEARNING TECHINQUES TO IMPORVE OPTICAL CHARACTER RECOGNITION FOR HISTORICAL DOCUMENTS

by Mengyue Zhao

beczhaozmy@gmail.com



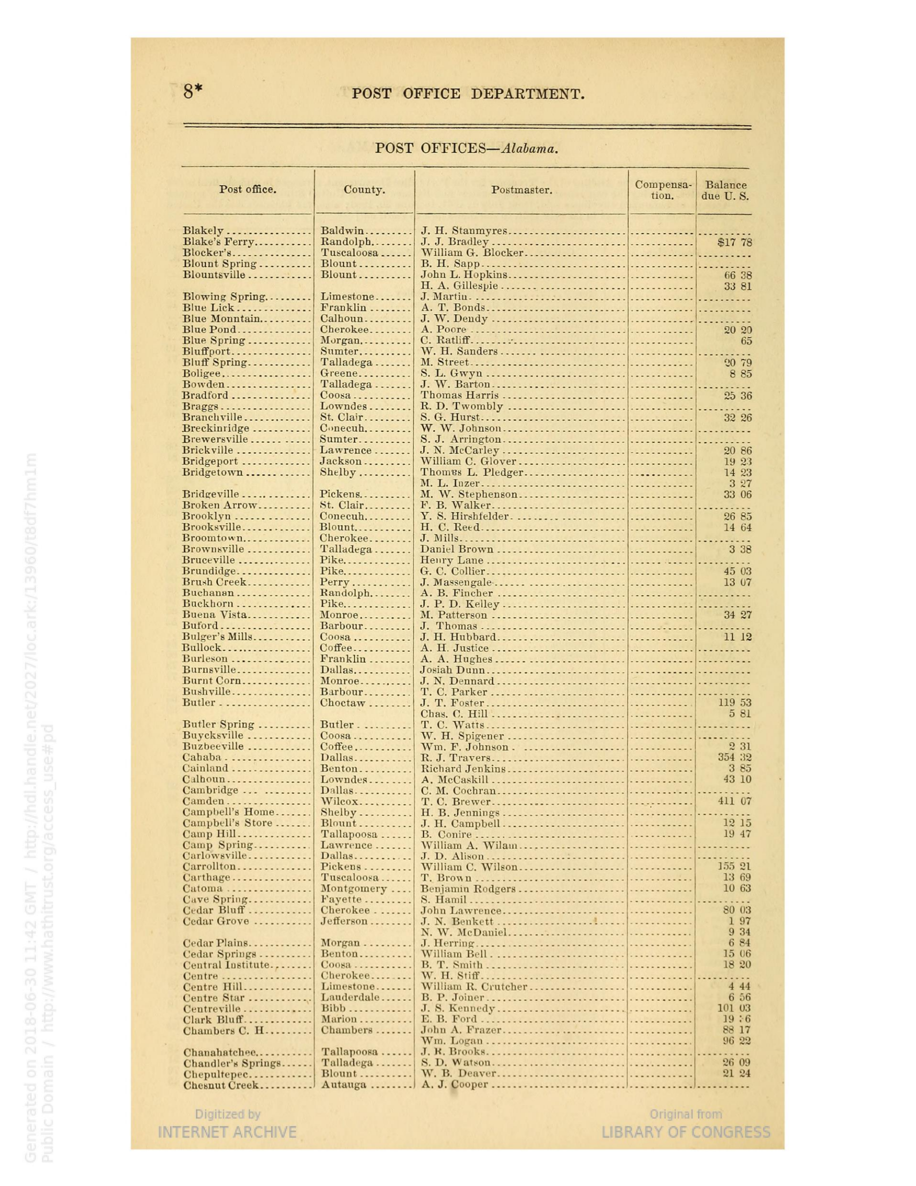
Problem Introduction

Historians have a hard time extracting large amounts of data from raw scans. Especially from historical tables which commercial for which commercial software are not designed.

Linking different historical sources requires very high accuracy of picture-to-data-frame recognition.

I found a very interesting large dataset that requires an improvement on such software. This project is designed to handle tens of thousands of raw scans as shown in the picture on the right side.

The alternative to this solution is manual input by typists, which is costly, inaccurate and often-time yields unknown or stringent mistakes that are hard to correct.



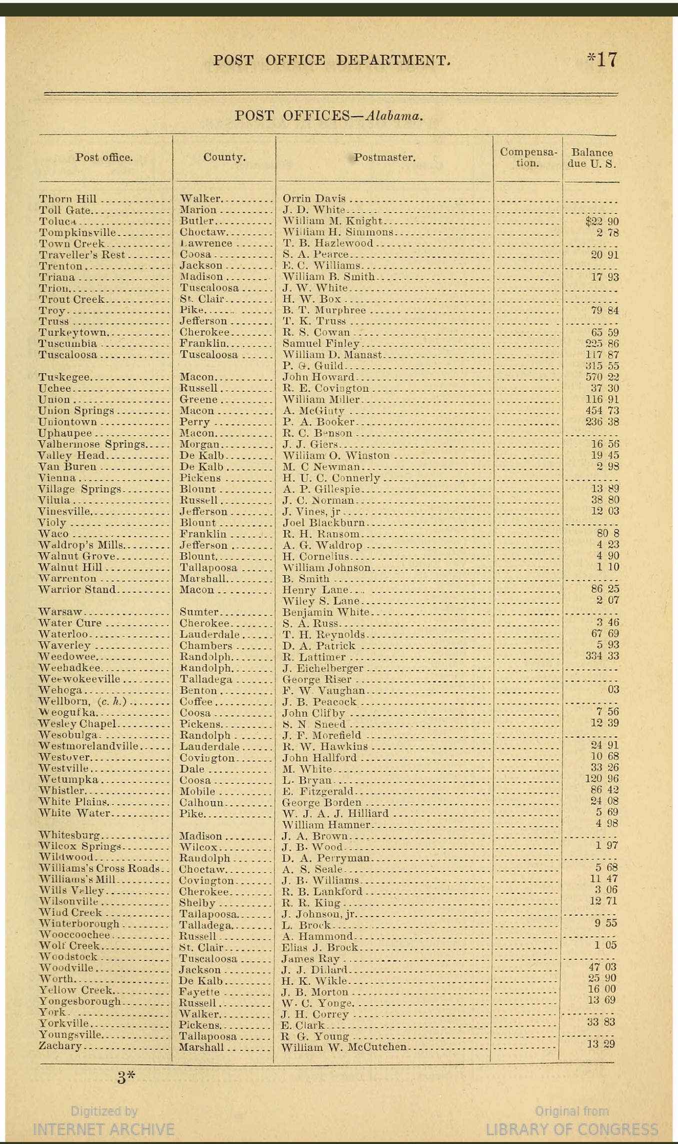
POST OFFICES—*Alabama.*

Post office.	County.	Postmaster.	Compensation.	Balance due U. S.
Blakely	Baldwin	J. H. Stanmyres		
Blake's Ferry	Randolph	J. J. Bradley		$17 78
Blocker's	Tuscaloosa	William G. Blocker		
Blount Spring	Blount	B. H. Sapp		
Blountsville	Blount	John L. Hopkins		66 38
		H. A. Gillespie		33 81
Blowing Spring	Limestone	J. Martin		
Blue Lick	Franklin	A. T. Bonds		
Blue Monntain	Calhoun	J. W. Dendy		
Blue Pond	Cherokee	A. Poore		20 20
Blue Spring	Morgan	C. Ratliff		65
Bluffport	Sumter	W. H. Sanders		
Bluff Spring	Talladega	M. Street		20 79
Boligee	Greene	S. L. Gwyn		8 85
Bowden	Talladega	J. W. Barton		
Bradford	Coosa	Thomas Harris		25 36
Braggs	Lowndes	R. D. Twombly		
Branchville	St. Clair	S. G. Hurst		32 26
Breckinridge	Conecuh	W. W. Johnson		
Brewersville	Sumter	S. J. Arrington		
Brickville	Lawrence	J. N. McCarley		20 86
Bridgeport	Jackson	William C. Glover		19 23
Bridgetown	Shelby	Thomas L. Pledger		14 23
		M. L. Inzer		3 27
Bridgeville	Pickens	M. W. Stephenson		33 06
Broken Arrow	St. Clair	F. B. Walker		
Brooklyn	Conecuh	Y. S. Hirshfelder		26 85
Brooksville	Blount	H. C. Reed		14 64
Broomtown	Cherokee	J. Mills		
Brownsville	Talladega	Daniel Brown		3 38
Bruceville	Pike	Henry Lane		
Brundidge	Pike	G. C. Collier		45 03
Brush Creek	Perry	J. Massengale		13 07
Buchanan	Randolph	A. B. Fincher		
Buckhorn	Pike	J. P. D. Kelley		
Buena Vista	Monroe	M. Patterson		34 27
Buford	Barbour	J. Thomas		
Bulger's Mills	Coosa	J. H. Hubbard		11 12
Bullock	Coffee	A. H. Justice		
Burleson	Franklin	A. A. Hughes		
Burnsville	Dallas	Josiah Dunn		
Burnt Corn	Monroe	J. N. Dennard		
Bushville	Barbour	T. C. Parker		
Butler	Choctaw	J. T. Foster		119 53
		Chas. C. Hill		5 81
Butler Spring	Butler	T. C. Watts		
Buycksville	Coosa	W. H. Spigener		
Buzbeeville	Coffee	Wm. F. Johnson		2 31
Cahaba	Dallas	R. J. Travers		354 32
Cainland	Benton	Richard Jenkins		3 85
Calhoun	Lowndes	A. McCaskill		43 10
Cambridge	Dallas	C. M. Cochran		
Camden	Wilcox	T. C. Brewer		411 07
Campbell's Home	Shelby	H. B. Jennings		
Campbell's Store	Blount	J. H. Campbell		12 15
Camp Hill	Tallapoosa	B. Conire		19 47
Camp Spring	Lawrence	William A. Wilam		
Carlowsville	Dallas	J. D. Alison		
Carrollton	Pickens	William C. Wilson		155 21
Carthage	Tuscaloosa	T. Brown		13 69
Catoma	Montgomery	Benjamin Rodgers		10 63
Cave Spring	Fayette	S. Hamil		
Cedar Bluff	Cherokee	John Lawrence		80 03
Cedar Grove	Jefferson	J. N. Benkett		1 97
		N. W. McDaniel		9 34
Cedar Plains	Morgan	J. Herring		6 84
Cedar Springs	Benton	William Bell		15 06
Central Institute	Coosa	B. T. Smith		18 20
Centre	Cherokee	W. H. Stiff		
Centre Hill	Limestone	William R. Crutcher		4 44
Centre Star	Lauderdale	B. P. Joiner		6 56
Centreville	Bibb	J. S. Kennedy		101 03
Clark Bluff	Marion	E. B. Ford		19 :6
Chambers C. H	Chambers	John A. Frazer		88 17
		Wm. Logan		96 22
Chanahatchee	Tallapoosa	J. R. Brooks		
Chandler's Springs	Talladega	S. D. Watson		26 09
Chepultepec	Blount	W. B. Deaver		21 24
Chesnut Creek	Autauga	A. J. Cooper		

Tesseract is an optical character recognition software initially maintained by HP in the 1990s, later open-sourced and now maintained by Google.

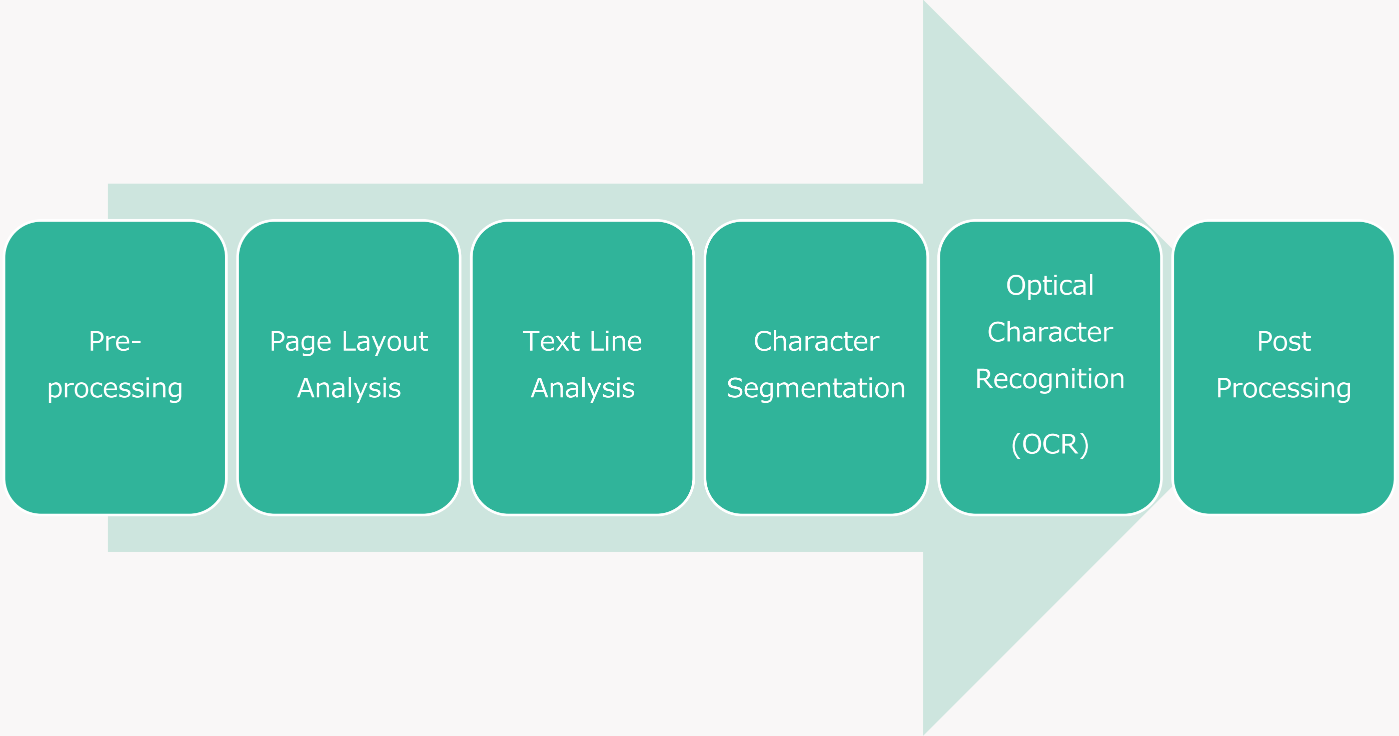
What is Tesseract and how does it work?

How does Tesseract (or OCR in general) Work?



POST OFFICES—*Alabama.*

Post office.	County.	Postmaster.	Compensation.	Balance due U. S.
Thorn Hill	Walker	Orrin Davis		
Toll Gate	Marion	J. D. White		
Toluca	Butler	William M. Knight		$22 90
Tompkinsville	Choctaw	William H. Simmons		2 78
Town Creek	Lawrence	T. B. Hazlewood		
Traveller's Rest	Coosa	S. A. Pearce		20 91
Trenton	Jackson	E. C. Williams		
Triana	Madison	William B. Smith		17 93
Trion	Tuscaloosa	J. W. White		
Trout Creek	St. Clair	H. W. Box		
Troy	Pike	B. T. Murphree		79 84
Truss	Jefferson	T. K. Truss		
Turkeytown	Cherokee	R. S. Cowan		65 59
Tuscumbia	Franklin	Samuel Finley		225 86
Tuscaloosa	Tuscaloosa	William D. Manast		117 87
		P. G. Guild		315 55
Tuskegee	Macon	John Howard		570 22
Uchee	Russell	R. E. Covington		37 30
Union	Greene	William Miller		116 91
Union Springs	Macon	A. McGinty		454 73
Uniontown	Perry	P. A. Booker		236 38
Uphaupee	Macon	R. C. Benson		
Valhermose Springs	Morgan	J. J. Giers		16 56
Valley Head	De Kalb	William O. Winston		19 45
Van Buren	De Kalb	M. C Newman		2 98
Vienna	Pickens	H. U. C. Connerly		
Village Springs	Blount	A. P. Gillespie		13 89
Vilula	Russell	J. C. Norman		38 80
Vinesville	Jefferson	J. Vines, jr		12 03
Violy	Blount	Joel Blackburn		
Waco	Franklin	R. H. Ransom		80 8
Waldrop's Mills	Jefferson	A. G. Waldrop		4 23
Walnut Grove	Blount	H. Cornelius		4 90
Walnut Hill	Tallapoosa	William Johnson		1 10
Warrenton	Marshall	B. Smith		
Warrior Stand	Macon	Henry Lane		86 25
		Wiley S. Lane		2 07
Warsaw	Sumter	Benjamin White		
Water Cure	Cherokee	S. A. Russ		3 46
Waterloo	Lauderdale	T. H. Reynolds		67 69
Waverley	Chambers	D. A. Patrick		5 93
Weedowee	Randolph	R. Lattimer		334 33
Weehadkee	Randolph	J. Eichelberger		
Weewokeeville	Talladega	George Riser		
Wehoga	Benton	F. W Vaughan		03
Wellborn, (*c. h.*)	Coffee	J. B. Peacock		
Weogufka	Coosa	John Clifby		7 56
Wesley Chapel	Pickens	S. N Sneed		12 39
Wesobulga	Randolph	J. F. Morefield		
Westmorelandville	Lauderdale	R. W. Hawkins		24 91
Westover	Covington	John Hallford		10 68
Westville	Dale	M. White		33 26
Wetumpka	Coosa	L. Bryan		120 96
Whistler	Mobile	E. Fitzgerald		86 42
White Plains	Calhoun	George Borden		24 08
White Water	Pike	W. J. A. J. Hilliard		5 69
		William Hamner		4 98
Whitesburg	Madison	J. A. Brown		
Wilcox Springs	Wilcox	J. B. Wood		1 97
Wildwood	Randolph	D. A. Perryman		
Williams's Cross Roads	Choctaw	A. S. Seale		5 68
Williams's Mill	Covington	J. B. Williams		11 47
Wills Valley	Cherokee	R. B. Lankford		3 06
Wilsonville	Shelby	R. R. King		12 71
Wind Creek	Tallapoosa	J. Johnson, jr		
Winterborough	Talladega	L. Brock		9 55
Wooccoochee	Russell	A. Hammond		
Wolf Creek	St. Clair	Elias J. Brock		1 05
Woodstock	Tuscaloosa	James Ray		
Woodville	Jackson	J. J. Dillard		47 03
Worth	De Kalb	H. K. Wikle		25 90
Yellow Creek	Fayette	J. B. Morton		16 00
Yongesborough	Russell	W. C. Yonge		13 69
York	Walker	J. H. Correy		
Yorkville	Pickens	E. Clark		33 83
Youngsville	Tallapoosa	R G. Young		
Zachary	Marshall	William W. McCutchen		13 29



Pre-processing

Page Layout Analysis

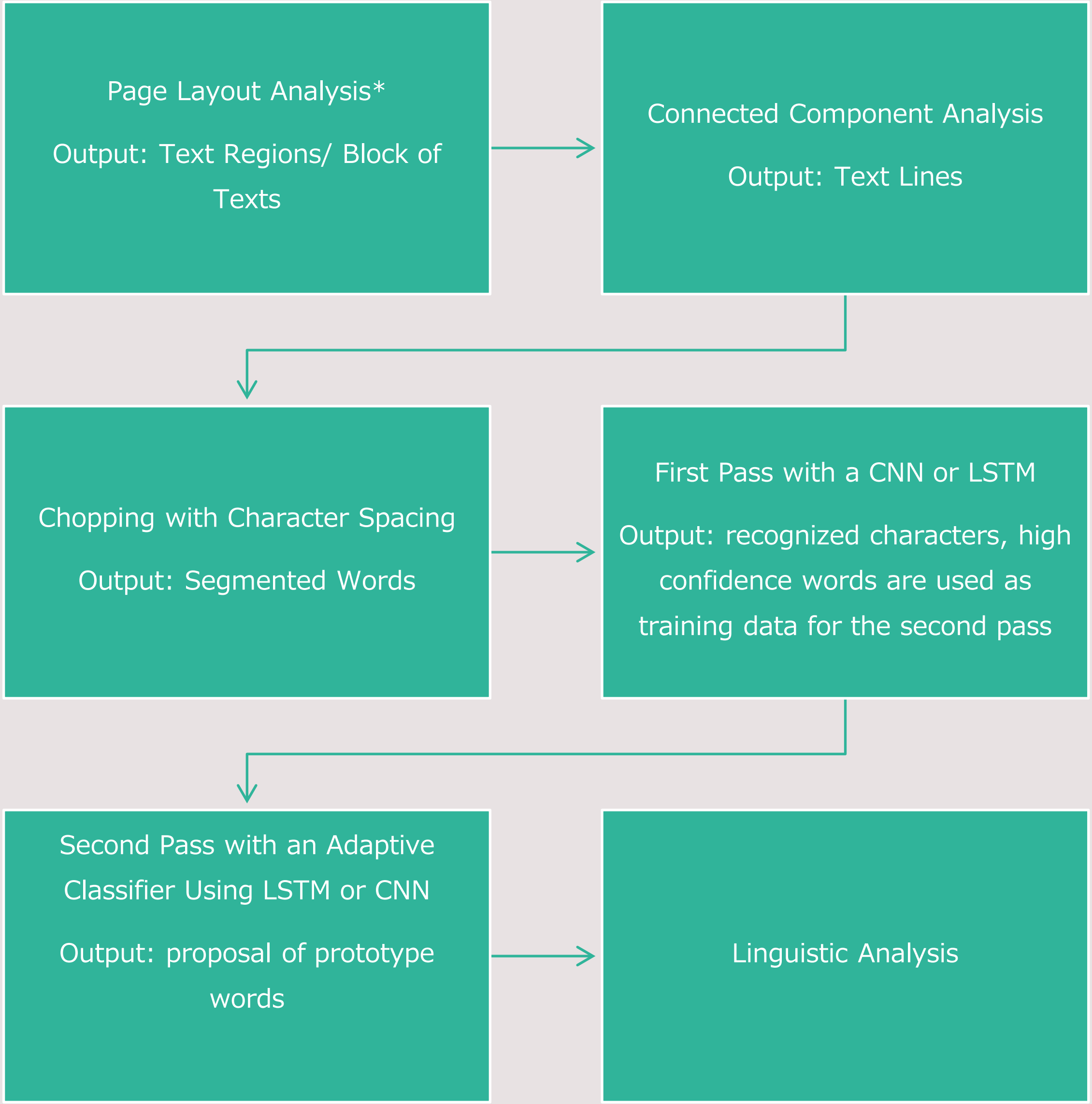
Text Line Analysis

Character Segmentation

Optical Character Recognition (OCR)

Post Processing

Tesseract Architecture



Page Layout Analysis*

Output: Text Regions/ Block of Texts

Connected Component Analysis

Output: Text Lines

Chopping with Character Spacing

Output: Segmented Words

First Pass with a CNN or LSTM

Output: recognized characters, high confidence words are used as training data for the second pass

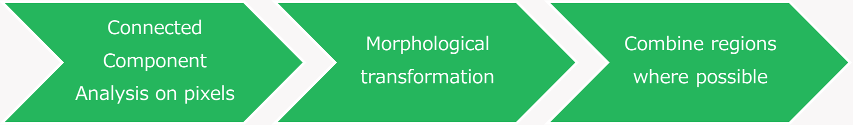
Second Pass with an Adaptive Classifier Using LSTM or CNN

Output: proposal of prototype words

Linguistic Analysis

Page Layout Analysis

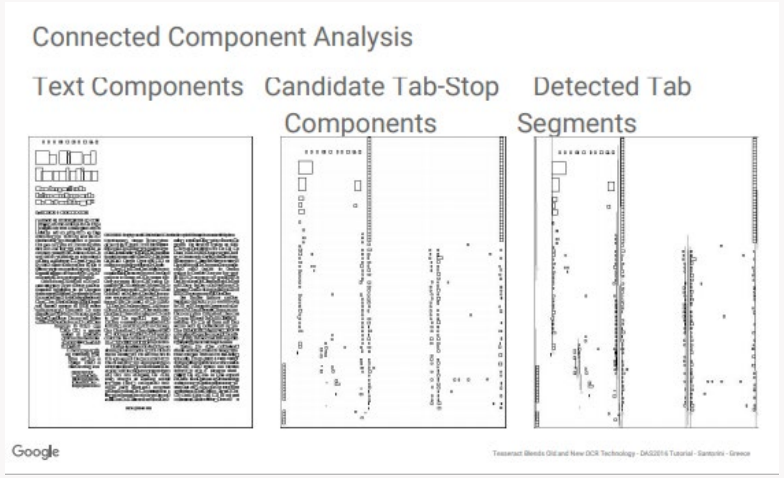
Bottom-up Approach



Connected Component Analysis on pixels

Morphological transformation

Combine regions where possible



Connected Component Analysis

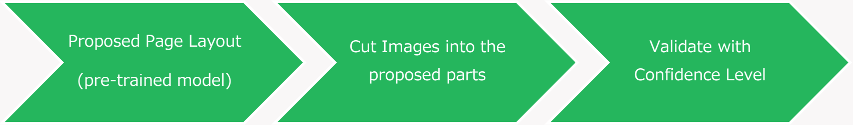
Text Components

Candidate Tab-Stop Components

Detected Tab Segments

Tesseract Blends Old and New OCR Technology - DAS2016 Tutorial - Santorini - Greece

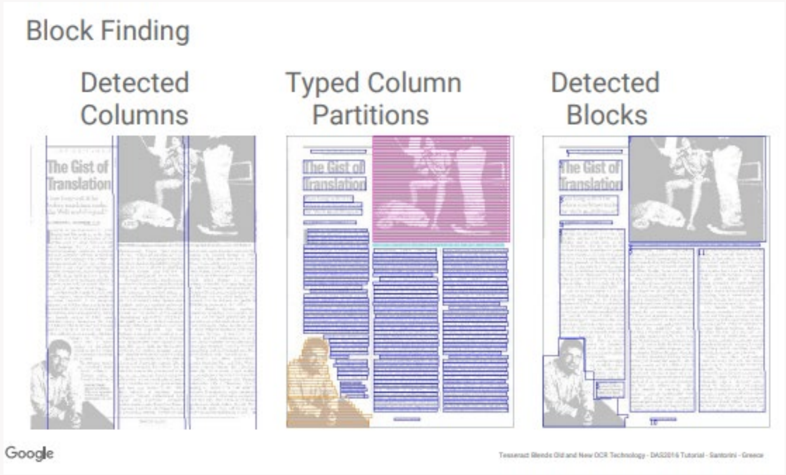
Top-down Approach



Proposed Page Layout (pre-trained model)

Cut Images into the proposed parts

Validate with Confidence Level



Block Finding

Detected Columns

Typed Column Partitions

Detected Blocks

Tesseract Blends Old and New OCR Technology - DAS2016 Tutorial - Santorini - Greece

Using LSTM in OCR

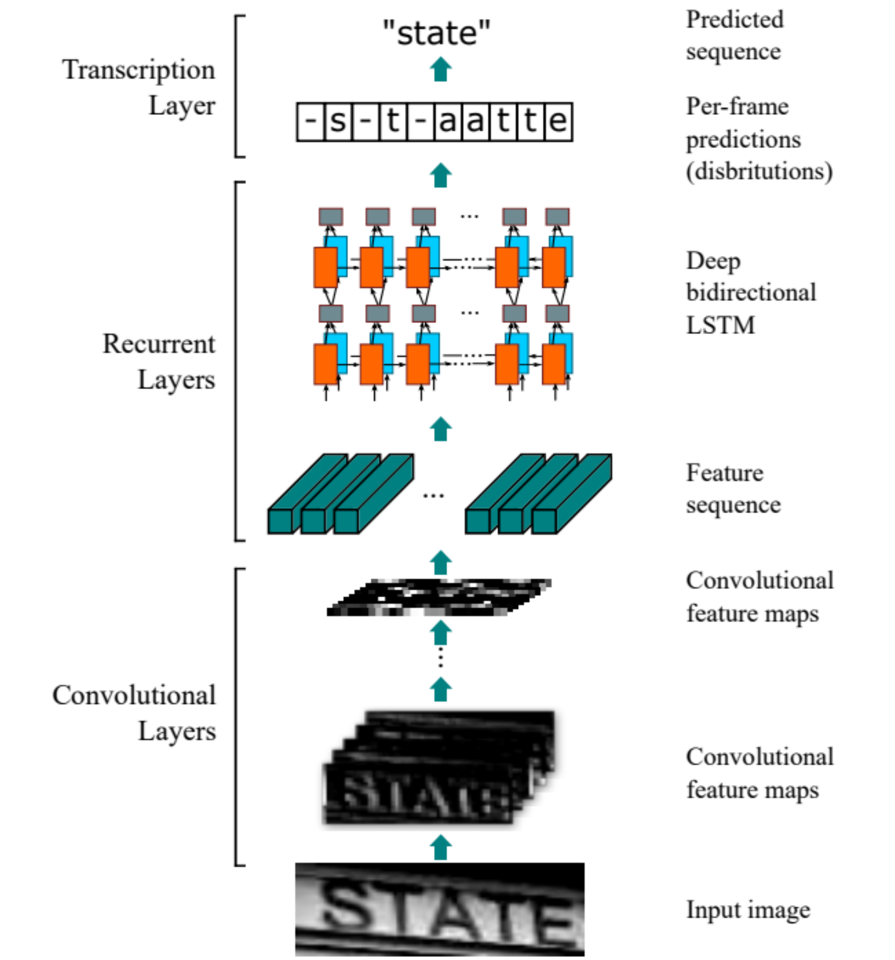


Figure 1. The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

Links

Visualized Tesseract Architecture

https://limitlessdatascience.wordpress.com/2019/07/01/tesseract-4-0-intro-installation/

How does Tesseract use LSTM

https://github.com/tesseract-ocr/docs/blob/master/das_tutorial2016/6ModernizationEfforts.pdf¥

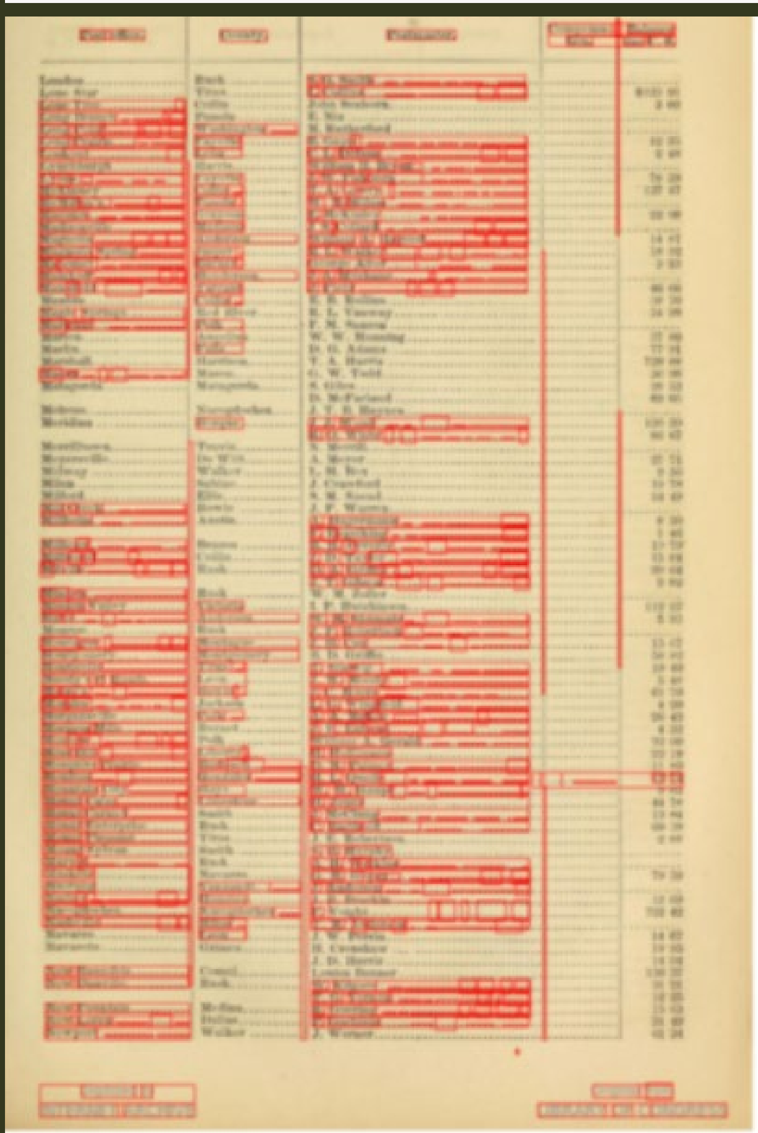
https://nanonets.com/blog/ocr-with-tesseract/

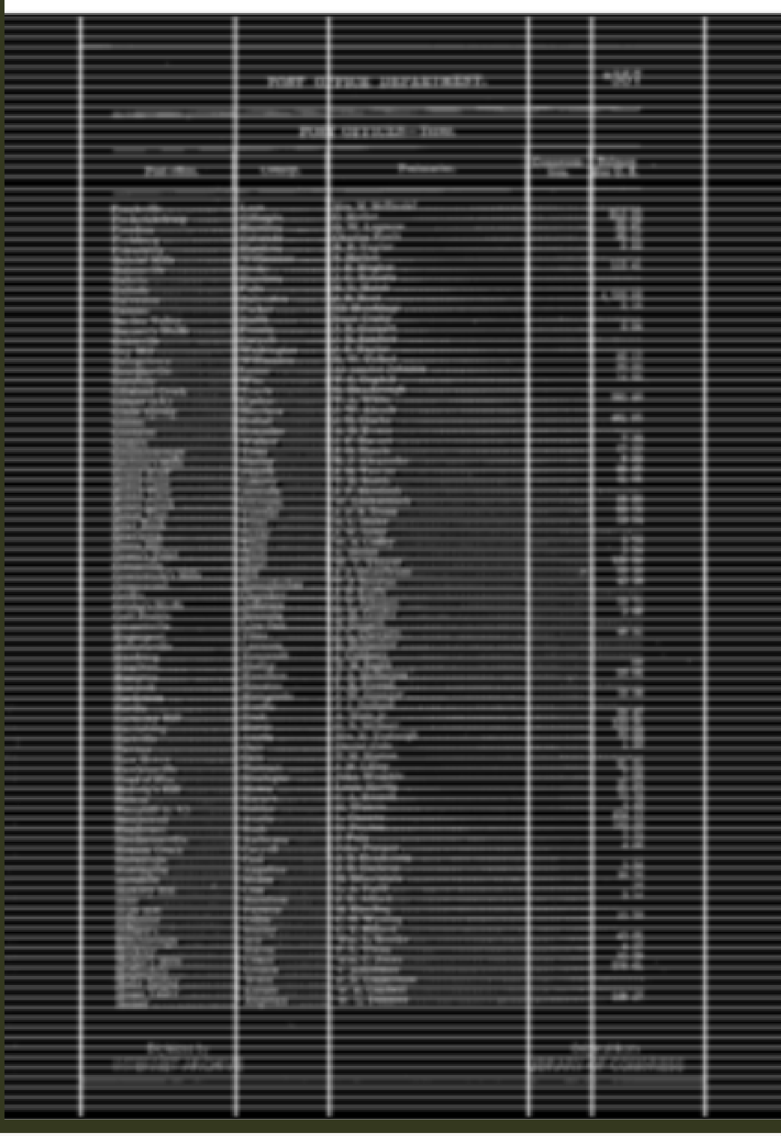
How LSTM work for OCR

https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa

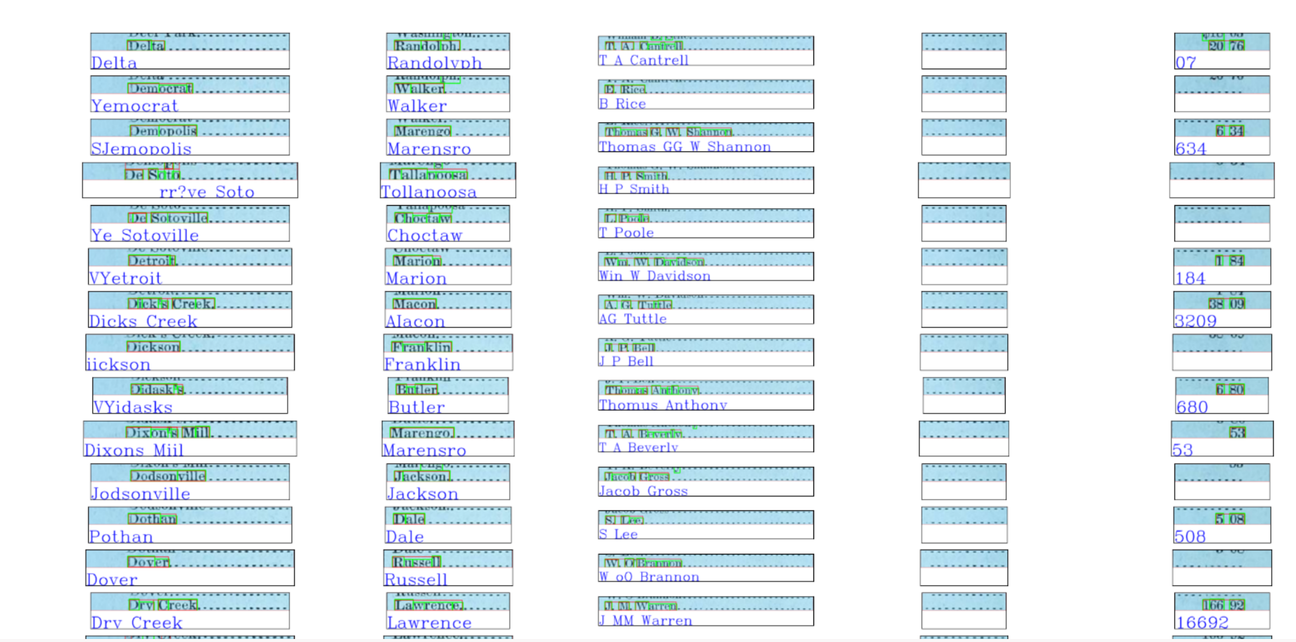
Page Segmentation Solutions

- Left: Tesseract Segmentation Results
- Right: CV Morphological Transformations





OCR Results



Delta Delta	Randolph. Randolvph	T. A. Cantrell T A Cantrell		20 76 07
Democrat Yemocrat	Walker. Walker	E. Rice B Rice		
Demopolis SJemopolis	Marengo Marensro	Thomas G. W. Shannon. Thomas GG W Shannon		6 34 634
De Soto rr?ve Soto	Tallapoosa Tollanoosa	H. P. Smith. H P Smith		
De Sotoville. Ye Sotoville	Choctaw Choctaw	L. Poole. T Poole		
Detroit. VYetroit	Marion. Marion	Wm. W. Davidson. Win W Davidson		1 84 184
Dick's Creek. Dicks Creek	Macon. Alacon	A. G. Tuttle AG Tuttle		38 09 3209
Dickson iickson	Franklin Franklin	J. P. Bell J P Bell		
Didask's. VYidasks	Butler. Butler	Thomas Anthony. Thomus Anthony		6 80 680
Dixon's Mill. Dixons Miil	Marengo. Marensro	T. A. Beverly. T A Beverlv		53 53
Dodsonville Jodsonville	Jackson. Jackson	Jacob Gross Jacob Gross		
Dothan Pothan	Dale Dale	S. Lee. S Lee		5 08 508
Dover. Dover	Russell Russell	W. O'Brannon. W oO Brannon		
Dry Creek. Drv Creek	Lawrence. Lawrence	J. M. Warren. J MM Warren		166 92 16692

Proposed Solutions

- Customized Preprocessing [done]
- Page Layout Analysis with Morphological Transformation [done]
- Page Segmentation with Morphological Transformation [done]
- Fixed-width Character Segmentation [tuning]
- Retrain Neural Network on Specific Fonts [done]
- Restrained Classifier with a Customized Dictionary (word list) [doing]
- Restricted Output Base [done]
- Post Processing [doing]