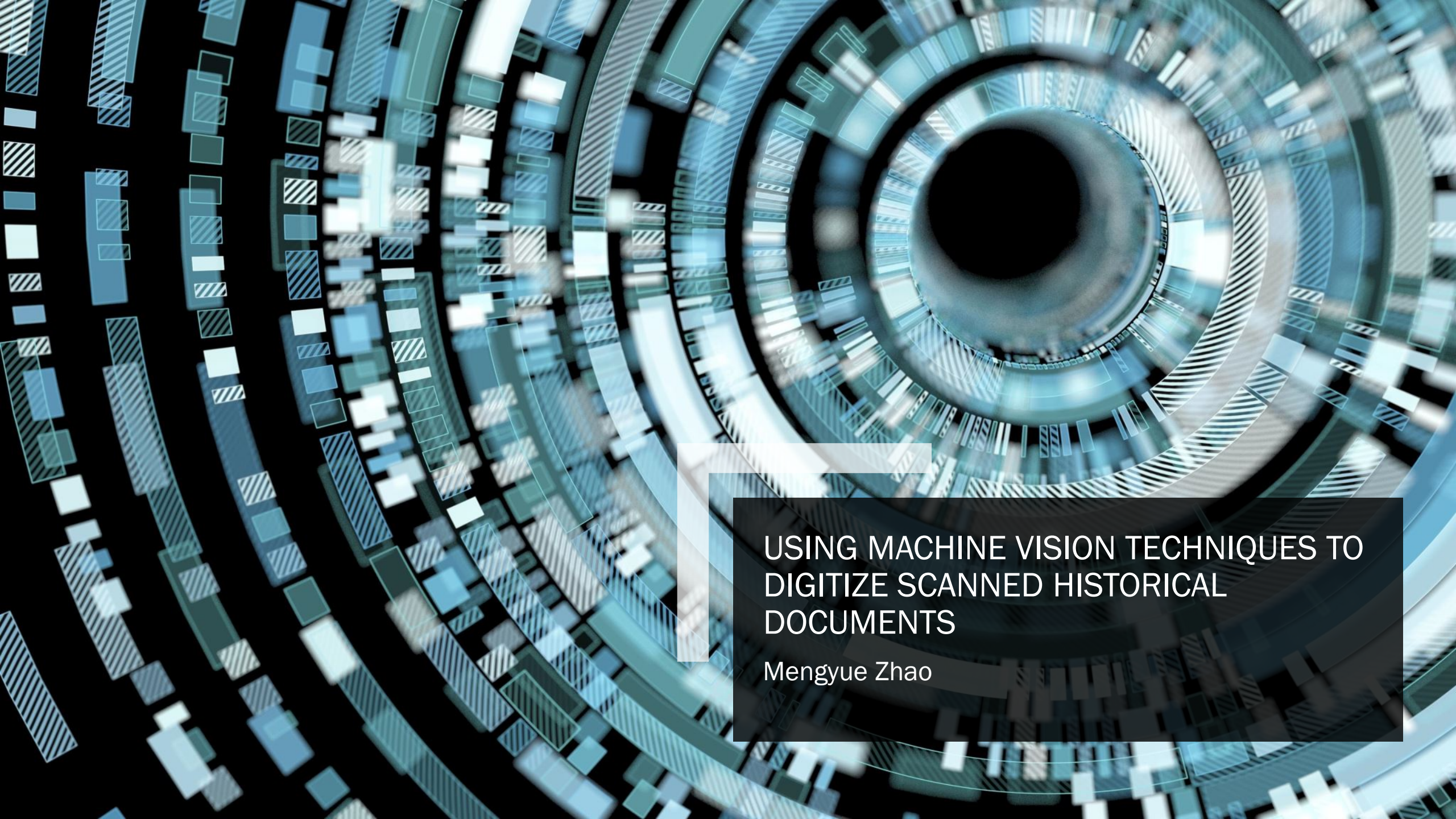


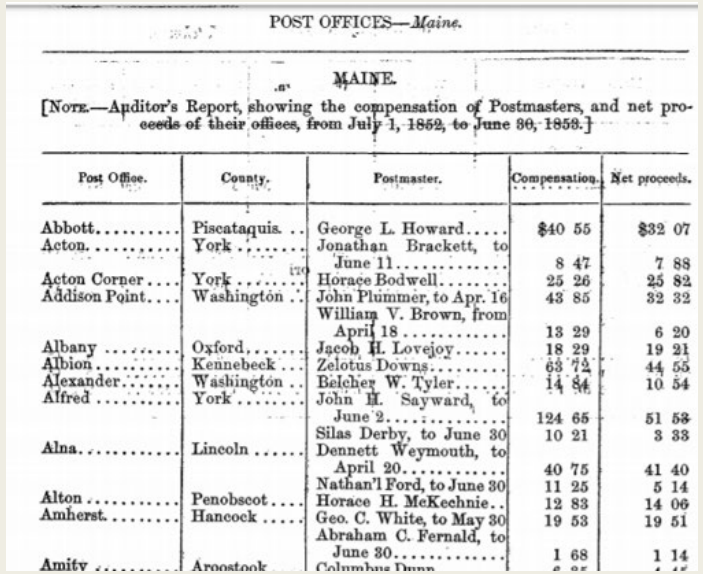
USING MACHINE VISION TECHNIQUES TO DIGITIZE SCANNED HISTORICAL DOCUMENTS

Mengyue Zhao

Motivations

- This is a side project of my undergraduate research work in the field of economic history. My work involves digitizing millions of historical economic indicators from scanned historical documents. The intention of this project is to allow the digitization to be conducted in a time-saving and cost-saving manner.
- The ***main challenge*** that I have encountered is that, due perhaps to the limitation of Nineteenth-Century and Twentieth-Century printing techniques, the tables are not properly aligned for ***detection by modern OCR software***. The solution is, there fore to make use of the computer vision capabilities of OpenCV and open source OCR packages such as Tesseract and Tensorflow OCR.

Table 1 Raw Input:



POST OFFICES—*Maine.*

MAINE.

[NOTE.—Auditor's Report, showing the compensation of Postmasters, and net proceeds of their offices, from July 1, 1852, to June 30, 1853.]

Post Office.	County.	Postmaster.	Compensation.	Net proceeds.
Abbott	Piscataquis	George L. Howard	$40 55	$32 07
Acton	York	Jonathan Brackett, to June 11	8 47	7 88
Acton Corner	York	Horace Bodwell	25 26	25 82
Addison Point	Washington	John Plummer, to Apr. 16	43 85	32 32
		William V. Brown, from April 18	13 29	6 20
Albany	Oxford	Jacob H. Lovejoy	18 29	19 21
Albion	Kennebeck	Zelotus Downs	63 72	44 55
Alexander	Washington	Belcher W. Tyler	14 84	10 54
Alfred	York	John H. Sayward, to June 2	124 65	51 53
		Silas Derby, to June 30	10 21	3 33
Alna	Lincoln	Dennett Weymouth, to April 20	40 75	41 40
		Nathan'l Ford, to June 30	11 25	5 14
Alton	Penobscot	Horace H. McKechnie	12 83	14 06
Amherst	Hancock	Geo. C. White, to May 30	19 53	19 51
		Abraham C. Fernald, to June 30	1 68	1 14
Amity	Aroostook	Columbus Dunn		

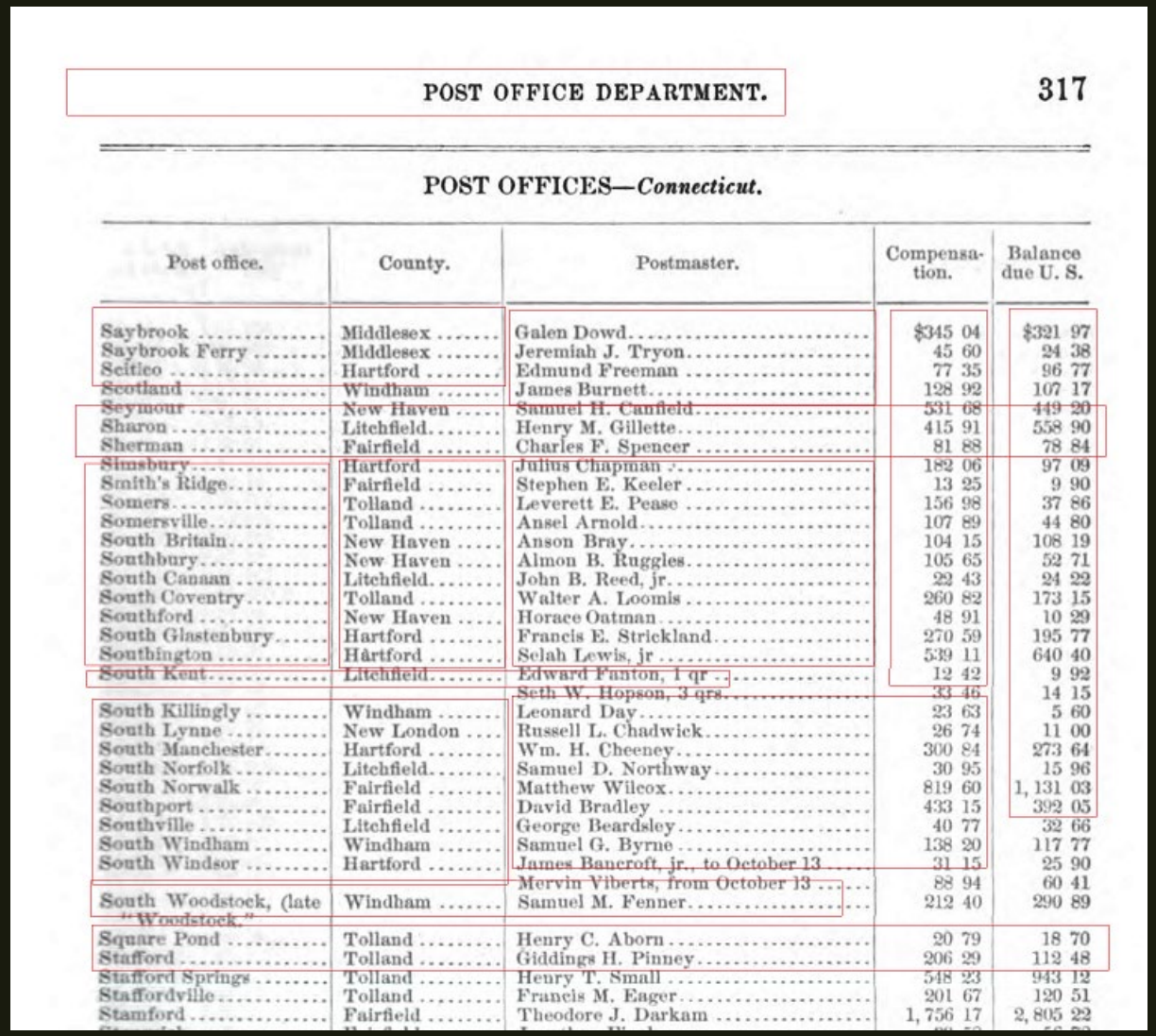
Table 2 Desired Output:

Post office	County	Postmaster	Compensation
Abbott	Piscataquis	George L. Howard	40. 55
Acton	York	Jonathan Brackett	8. 47
Acton Corner	York	Horace Bodwell	25. 26
Addison Point	Washington	John Plummer	43. 85
		William V. Brown	13. 29
Albany	Oxford	Jacob H. Lovejoy	18. 29
Albion	Kennebeck	Zelotus Downs	63. 72
Alexander	Washington	Belcher W. Tyler	14. 84
Alfred	York	John H. Sayward	124. 65
		Silas Derby	10. 21
Alna	Lincoln	Dennett Weymouth	40. 75
		Nathan' I Ford	11. 25
Alton	Penobscot	Horace H. McKechnie	12. 83
Amherst	Hancock	Geo. C. White	19. 53
		Abraham C. Fernald	1. 68
Amity	Aroostook	Columbus Dunn	6. 35
Andover	Oxford	Ezekiel Merrill	48. 51
		George J. Wardwell	5. 10
Anson	Somerset	Arthur Dinsmore	7. 73
Argyle	Penobscot	Gideon J. Newton	10. 36
Aroostook	Aroostook	Ansel T. Mooers	69. 05

Original Documents and Project Goals

- My goal is to extract the records contained in the five columns on the lefthand side, ignoring information outside of the table structure.
- Table 1: An Example of Raw Scan of Historical Document
- Table 2: Desired Output – ***Nicely Aligned*** (row-wise and column wise) table containing ***mostly correctly recognized words***.

Undesirable Results from Commercial Software:



POST OFFICE DEPARTMENT. 317

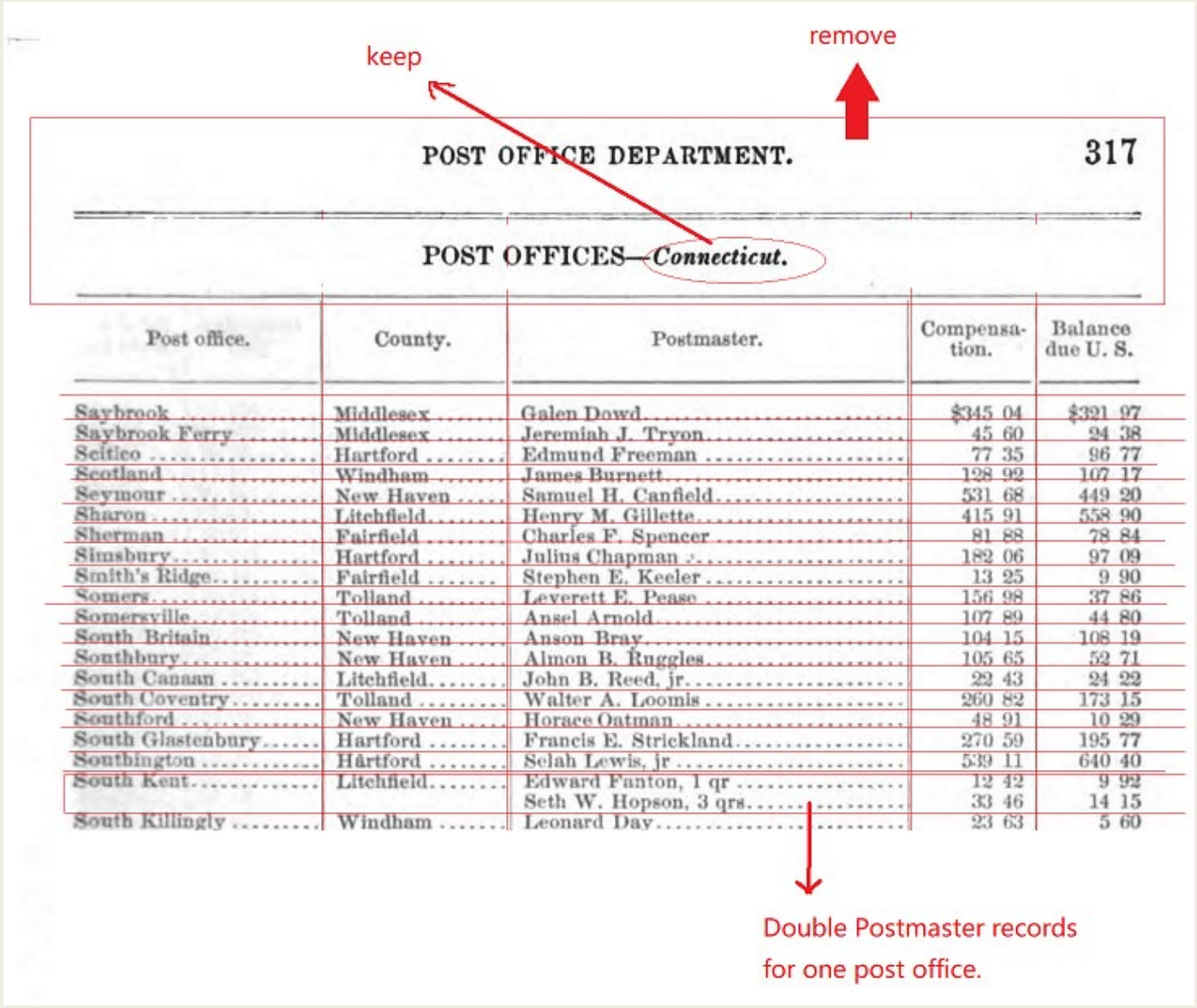
POST OFFICES—*Connecticut.*

Post office.	County.	Postmaster.	Compensation.	Balance due U. S.
Saybrook	Middlesex	Galen Dowd	$345 04	$321 97
Saybrook Ferry	Middlesex	Jeremiah J. Tryon	45 60	24 38
Scitico	Hartford	Edmund Freeman	77 35	96 77
Scotland	Windham	James Burnett	128 92	107 17
Seymour	New Haven	Samuel H. Canfield	531 68	449 20
Sharon	Litchfield	Henry M. Gillette	415 91	558 90
Sherman	Fairfield	Charles F. Spencer	81 88	78 84
Simsbury	Hartford	Julius Chapman	182 06	97 09
Smith's Ridge	Fairfield	Stephen E. Keeler	13 25	9 90
Somers	Tolland	Leverett E. Pease	156 98	37 86
Somersville	Tolland	Ansel Arnold	107 89	44 80
South Britain	New Haven	Anson Bray	104 15	108 19
Southbury	New Haven	Almon B. Ruggles	105 65	52 71
South Canaan	Litchfield	John B. Reed, jr	22 43	24 22
South Coventry	Tolland	Walter A. Loomis	260 82	173 15
Southford	New Haven	Horace Oatman	48 91	10 29
South Glastenbury	Hartford	Francis E. Strickland	270 59	195 77
Southington	Hartford	Selah Lewis, jr	539 11	640 40
South Kent	Litchfield	Edward Fanton, 1 qr	12 42	9 92
		Seth W. Hopson, 3 qrs	33 46	14 15
South Killingly	Windham	Leonard Day	23 63	5 60
South Lynne	New London	Russell L. Chadwick	26 74	11 00
South Manchester	Hartford	Wm. H. Cheeney	300 84	273 64
South Norfolk	Litchfield	Samuel D. Northway	30 95	15 96
South Norwalk	Fairfield	Matthew Wilcox	819 60	1,131 03
Southport	Fairfield	David Bradley	433 15	392 05
Southville	Litchfield	George Beardsley	40 77	32 66
South Windham	Windham	Samuel G. Byrne	138 20	117 77
South Windsor	Hartford	James Bancroft, jr., to October 13	31 15	25 90
		Mervin Viberts, from October 13	88 94	60 41
South Woodstock, (late "Woodstock.")	Windham	Samuel M. Fenner	212 40	290 89
Square Pond	Tolland	Henry C. Aborn	20 79	18 70
Stafford	Tolland	Giddings H. Pinney	206 29	112 48
Stafford Springs	Tolland	Henry T. Small	548 23	943 12
Staffordville	Tolland	Francis M. Eager	201 67	120 51
Stamford	Fairfield	Theodore J. Darkam	1,756 17	2,805 22

The red lines in the picture on the left-hand side shows the typical table structure that would be extracted by commercial OCR software.

The result contains the following problem:

- Improper row/ column alignment – the software recognize “block” of words, rather than the table structure.
- Undesired content outside of the table was recognized.



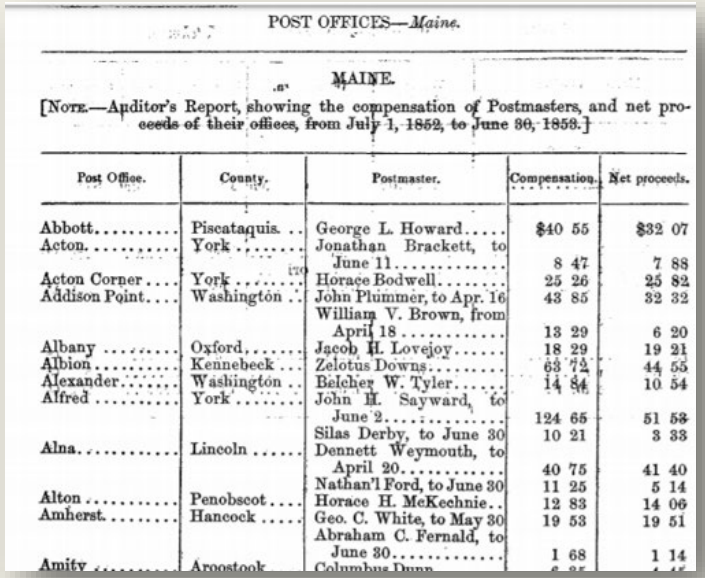
Ideal Table Structure Extraction:

- Get state name from the header, ignoring everything else on the header.
- Properly aligned rows and columns.
- Sometimes there are two records of “postmasters” (column 3) for one post office (column 1). This needs to be addressed as well.

Main Ideas for My Solution

- Step 1: Use Computer Vision to Extract Table Structures, encoding some rules/ patterns that I recognized as a human.
- Step 2: After obtaining the table structure, connect to open-source OCR libaries and read within these structures.

Figure 2. Binary Table Structure Extraction

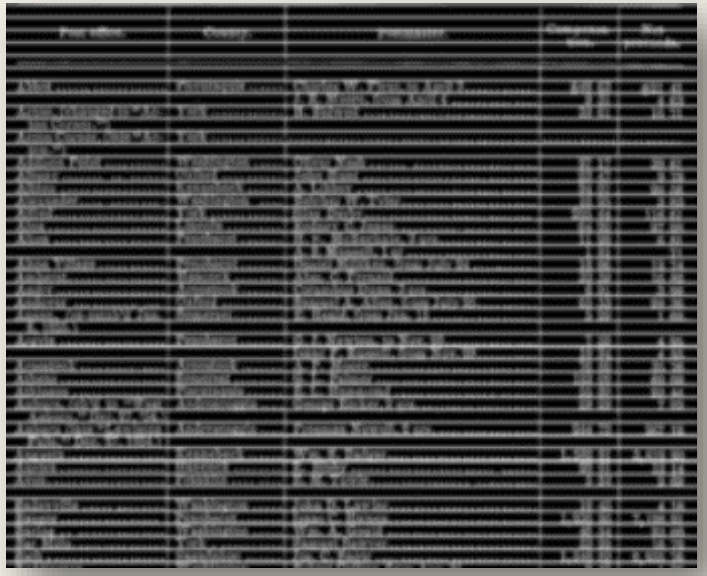


POST OFFICES—*Maine.*

MAINE.

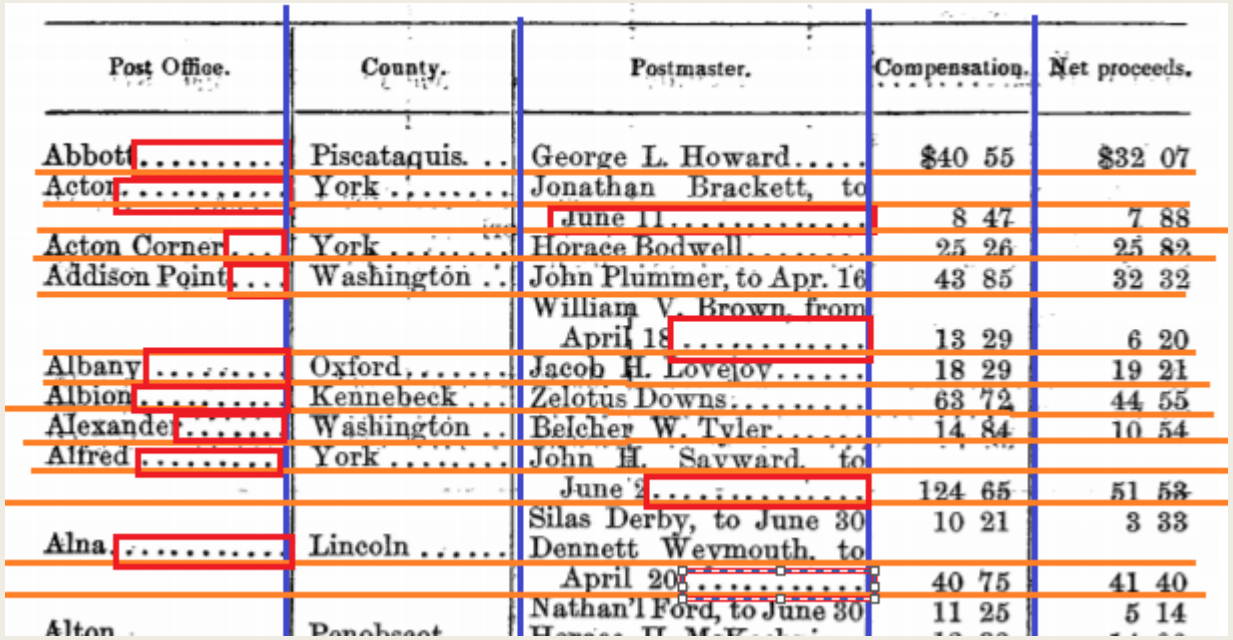
[NOTE.—Auditor's Report, showing the compensation of Postmasters, and net proceeds of their offices, from July 1, 1852, to June 30, 1853.]

Post Office.	County.	Postmaster.	Compensation.	Net proceeds.
Abbott	Piscataquis	George L. Howard	$40 55	$32 07
Acton	York	Jonathan Brackett, to June 11	8 47	7 88
Acton Corner	York	Horace Bodwell	25 26	25 82
Addison Point	Washington	John Plummer, to Apr. 16	43 85	32 32
		William V. Brown, from April 18	13 29	6 20
Albany	Oxford	Jacob H. Lovejoy	18 29	19 21
Albion	Kennebeck	Zelotus Downs	63 75	44 55
Alexander	Washington	Belcher W. Tyler	14 84	10 54
Alfred	York	John H. Sayward, to June 2	124 65	51 53
		Silas Derby, to June 30	10 21	3 33
Alna	Lincoln	Dennett Weymouth, to April 20	40 75	41 40
		Nathan'l Ford, to June 30	11 25	5 14
Alton	Penobscot	Horace H. McKechnie	12 83	14 06
Amherst	Hancock	Geo. C. White, to May 30	19 53	19 51
		Abraham C. Fernald, to June 30	1 68	1 14
Amity	Aroostook	Columbus Dunn		



Extracting Table Structure Using machine Vision

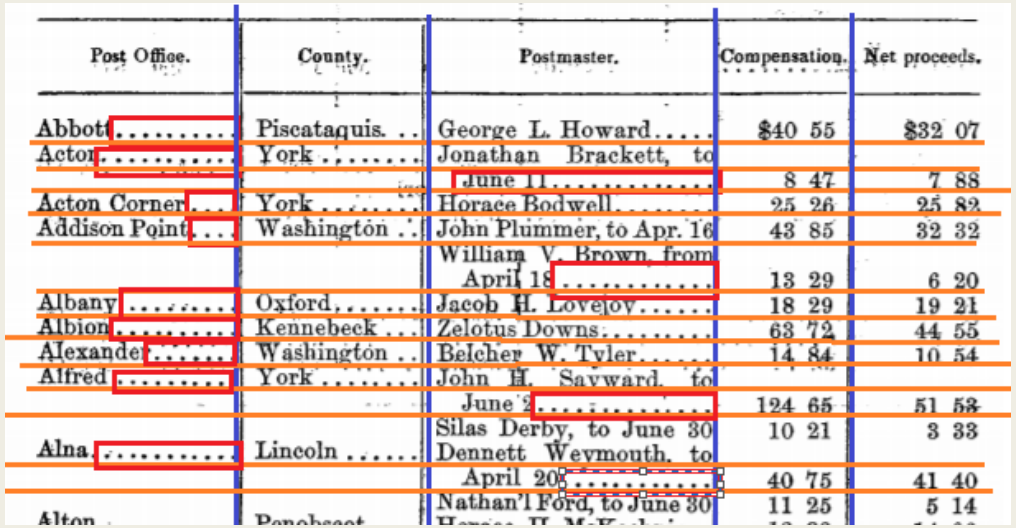
- To extract table structure, we need to do is to find the borders of the rows and the columns.
- Although the main challenge for the project is that there is no clearly bordered rows, the trailing dots (cropped in red) in the picture bellow in each line are a good reference for where the row is.
- Therefore, the first step is to recognize the trailing dots.



Post Office.	County.	Postmaster.	Compensation.	Net proceeds.
Abbott	Piscataquis.	George L. Howard	$40 55	$32 07
Acton	York	Jonathan Brackett, to June 11	8 47	7 88
Acton Corner	York	Horace Bodwell	25 26	25 82
Addison Point	Washington	John Plummer, to Apr. 16	43 85	32 32
		William V. Brown, from April 18	13 29	6 20
Albany	Oxford	Jacob H. Lovejoy	18 29	19 21
Albion	Kennebeck	Zelotus Downs	63 72	44 55
Alexander	Washington	Belcher W. Tyler	14 84	10 54
Alfred	York	John H. Sayward, to June 2	124 65	51 53
		Silas Derby, to June 30	10 21	3 33
Alna	Lincoln	Dennett Weymouth, to April 20	40 75	41 40
		Nathan'l Ford, to June 30	11 25	5 14
Alton	Penobscot			

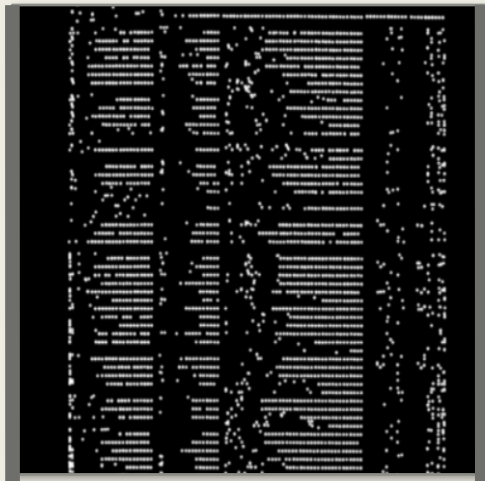
Steps for Extracting Table Structure Using Machine Vision & Human Logic

- The steps are colour coded.
- Step 1 (red):
  - *Recognize the trailing dot pattern with customised elliptical kernels. (using OpenCV).*
  - *Filter recognized dot patterns by the vertical distances between patterns, since there are unwanted noises on the document as well.*
  - *The location of the dots should give good reference for where the rows are.*
- Step 2 (orange):
  - *Draw a horizontal line through the recognized patterns, taking 2-3 recognized dots as reference points.*
  - *Move the horizontal line downward by about 1mm to get the exact position of the row border.*
- Step 3 (blue):
  - *Recognize the vertical borders with rectangle kernels.*
  - *Vertical borders are relatively easy to recognize since they are thicker, and there aren't much noises that resemble a vertical line.*



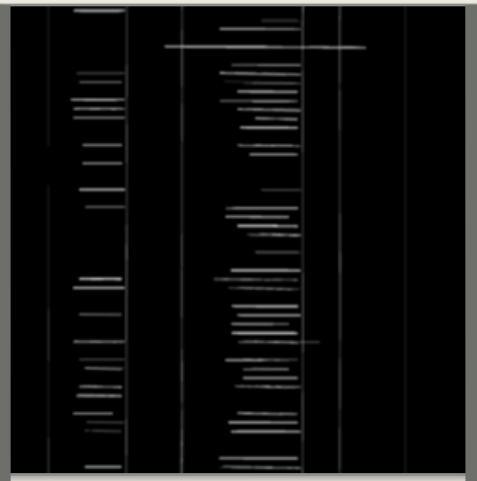
Post Office.	County.	Postmaster.	Compensation.	Net proceeds.
Abbott	Piscataquis.	George L. Howard	$40 55	$32 07
Acton	York	Jonathan Brackett, to June 11	8 47	7 88
Acton Corner	York	Horace Bodwell	25 26	25 82
Addison Point	Washington	John Plummer, to Apr. 16	43 85	32 32
		William V. Brown, from April 18	13 29	6 20
Albany	Oxford	Jacob H. Lovejoy	18 29	19 21
Albion	Kennebeck	Zelotus Downs	63 72	44 55
Alexander	Washington	Belcher W. Tyler	14 84	10 54
Alfred	York	John H. Sayward, to June 2	124 65	51 53
		Silas Derby, to June 30	10 21	3 33
Alna	Lincoln	Dennett Weymouth, to April 20	40 75	41 40
		Nathan'l Ford, to June 30	11 25	5 14
Alton	Penobscot			

Intermediate results



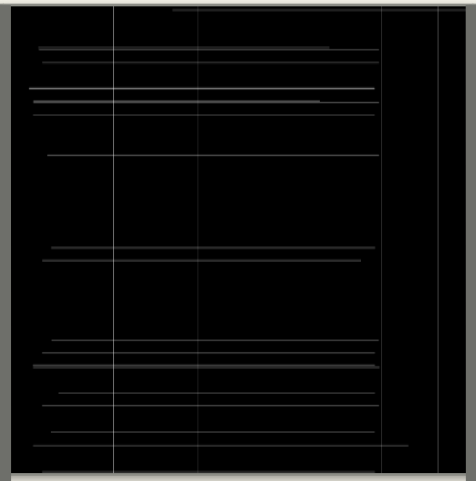
Recognize Dot Patterns

- Turn pictures into binary
- Create dot kernels/ scanners to parse through the page.
- Change the number of dots included and the size of the kernel until I get the best result. This resembles a cross-validation process.
- Dilate and contract until I have obtained the best result.



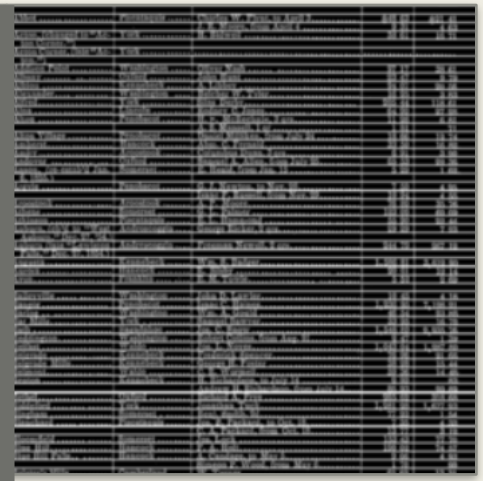
Denoise

- Keep r out horizontally continuous dot patterns that is about 8mm apart from its horizontal neighbour and remove all other dots. This will get rid of all the noises in the first picture.
- Draw lines through the resultant patterns.



Get Structure

- Get the average of the vertical coordinates of the horizontal lines in picture 2, and extract only 2-3 reference points across the page.
- Remove all lines in picture 2 and draw new lines through the reference points. The purpose of this repetitive process is to filter out as much noise as possible and get more precise reference points.



Mask the Table For Text Extraction with OCR

- Now we have a nice table structure for the OCR process.
- Note that sometimes column 1 has a one-to-many relationship with column 2 and so we might miss information for some rows in column 1. This problem could be solved by hard-coded rules after OCR.

Summary and Reflections

- This project is an example of my interest in machine vision, ability and motivation to explore all possible solutions for my problem.
- The next step of the project is to connect to open-source OCR libraries and tesseract to extract the content of the table.
- At the time that I worked with this project, I have not learned deep learning techniques such as CNN. After learning about the capability of these new techniques, I realized that some of the steps, such as creating the kernel and scanning the picture with it, could be replaced by deep learning techniques. This would potentially increase the accuracy a lot if combined with some hard-coded rules.
- Transfer learning might also be another area to explore, since local patterns (dots) and global patterns (tables) could be handled simultaneously by different layers.