Multi-Stage Framework to Boost Optical Character Recognition Performance on Low Quality Document Images

Nitin Gupta, Shashank Mujumdar, Abhinav Jain, Douglas Burdick

IBM Research

Document Intelligence Workshop at KDD 2021

Improving OCR on Low-Quality Document Images - Overview

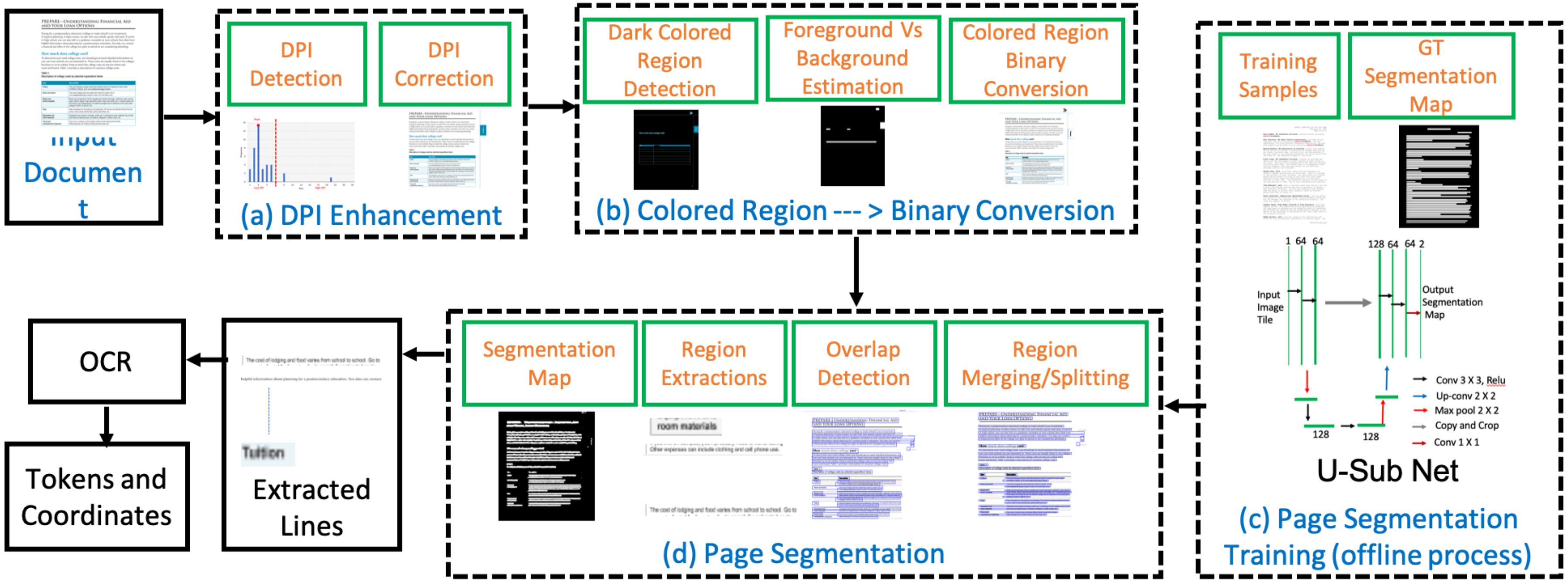
Factors affecting **Tesseract Engine**’s ability to generate binary representations and perform page segmentation :-

- Low Resolution
- Illumination Change
- Blur
- Noise
- Character Merging or Fragmentation
- Colored Regions with poor text contrast against background

We argue that any approach catering a specific issue in low-quality documents is sub-optimal in improving OCR performance. Thus, we propose a multi-stage framework that independently improves the performance of Tesseract at every stage.

Lastly, we present results on five challenging document image datasets and show superior performance against state-of-the-art baselines.

Proposed Pipeline Overview



Our framework consists of:–

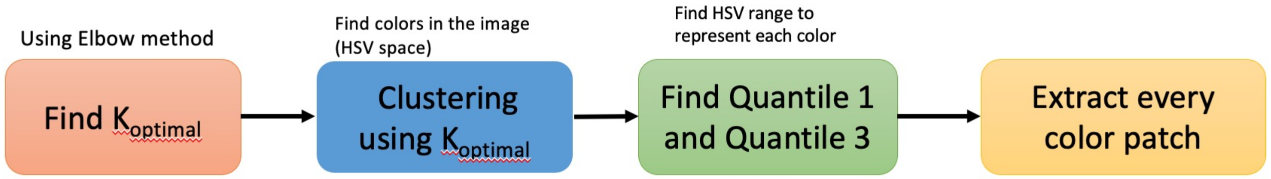
- **DPI Enhancement Module** to identify and up-scale low resolution document images
- **Colored Region Detection** Module to detect and binarize colored text regions
- **Page-Segmentation Module** to extract text lines

DPI Enhancement

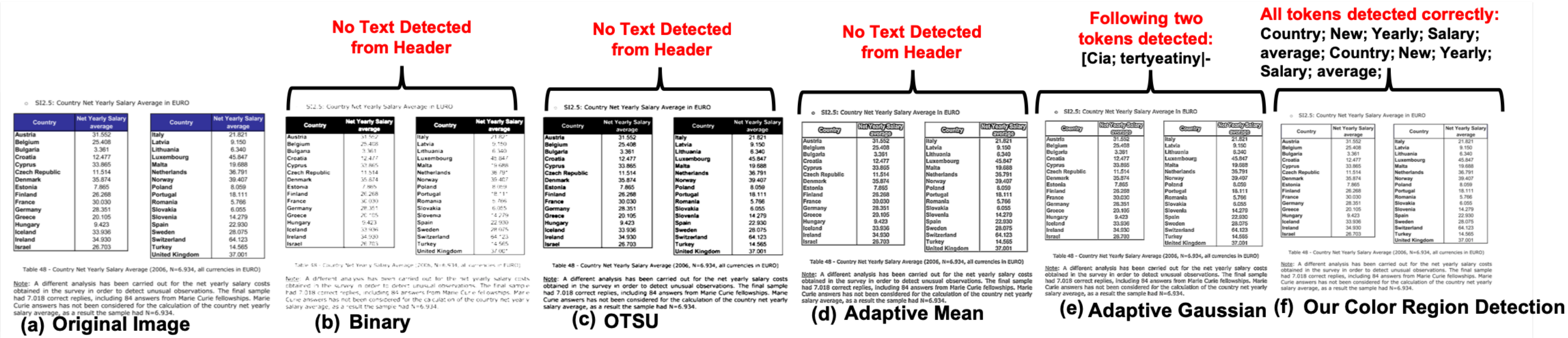
- Tesseract works best on document images with >300 DPI but suffers on 72-100 DPI low quality scans.
- We intend to re-scale the low 72-100 DPI images to 300 DPI using bi-cubic interpolation.
- From our comprehensive analysis, we identify that text lines in low 72-100 DPI images are of pixel height no more than 20 pixels. Thus, using this as a threshold, we detect and interpolate 72-100 DPI images.

Colored Region – Binary Conversion

- Existing OCR systems use Otsu Thresholding on the whole image to achieve binarization which fails on regions containing text against colored background such as headers, headings highlighted text etc.
- The local contrast observed between text and colored-regions is different than what is observed in rest of the document with mostly dark-text appearing against light background.
- We contain those local contrasts by detecting colored regions as clusters (use K-Means) in the HSV space rep the document and locally perform Otsu-thresholding in each detected region to achieve binarization.


Using Elbow method
Find $K_{optimal}$
Find colors in the image (HSV space)
Clustering using $K_{optimal}$
Find HSV range to represent each color
Find Quantile 1 and Quantile 3
Extract every color patch

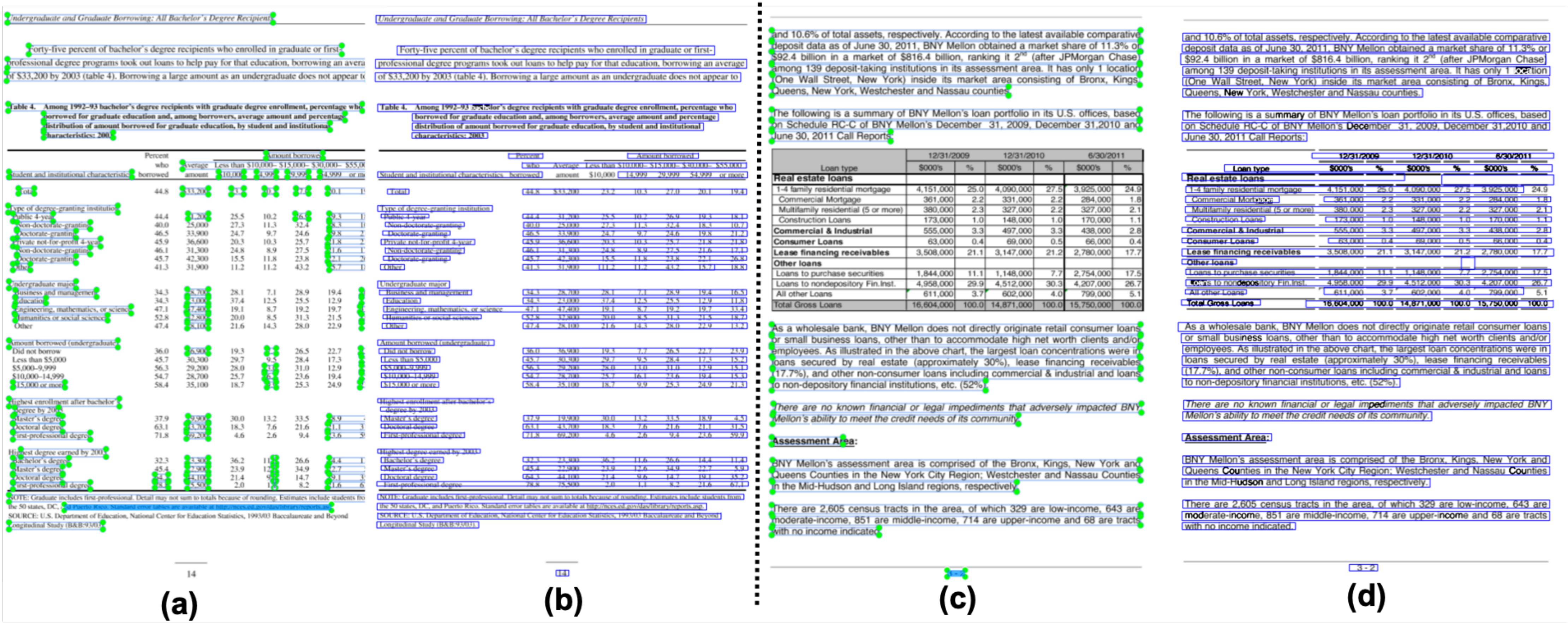
Colored Region Extraction Module


No Text Detected from Header
No Text Detected from Header
No Text Detected from Header
Following two tokens detected: [Cia; tertyeatiny|-
All tokens detected correctly: Country; New; Yearly; Salary; average; Country; New; Yearly; Salary; average;

(a) Original Image (b) Binary (c) OTSU (d) Adaptive Mean (e) Adaptive Gaussian (f) Our Color Region Detection

Page Segmentation

- With complexity and variations of 2D layouts, page segmentation is challenging
- We formulate the page segmentation as image-to-image transformation problem.
- We motivate the required architecture from U-Net, used in medical image segmentation. The output is a per-pixel probability estimate that the pixel is a part of segmented text lines. We set the threshold of the output to get the binary segmented image.
- We further perform post-processing like Region smoothing, Region Overlap Detection and Region Merg tighter bounding boxes in the segmented images.



Undergraduate and Graduate Borrowing: All Bachelor's Degree Recipients

Forty-five percent of bachelor's degree recipients who enrolled in graduate or first-professional degree programs took out loans to help pay for that education, borrowing an average of $33,200 by 2003 (table 4). Borrowing a large amount as an undergraduate does not appear to

Table 4. Among 1992–93 bachelor's degree recipients with graduate degree enrollment, percentage who borrowed for graduate education and, among borrowers, average amount and percentage distribution of amount borrowed for graduate education, by student and institutional characteristics: 2003

	Percent		Amount borrowed				
	who	Average	Less than	$10,000–	$15,000–	$30,000–	$55,000
Student and institutional characteristics	borrowed	amount	$10,000	14,999	29,999	54,999	or more
Total	44.8	$33,200	23.2	10.3	27.0	20.1	19.4
Type of degree-granting institution							
Public 4-year	44.4	31,200	25.5	10.2	26.9	19.3	18.1
Non-doctorate-granting	40.0	25,000	27.3	11.3	32.4	18.3	10.7
Doctorate-granting	46.5	33,900	24.7	9.7	24.6	19.8	21.2
Private not-for-profit 4-year	45.9	36,600	20.3	10.3	25.7	21.8	21.8
Non-doctorate-granting	46.1	31,300	24.8	8.9	27.5	21.6	17.1
Doctorate-granting	45.7	42,300	15.5	11.8	23.8	22.1	26.8
Other	41.3	31,900	11.2	11.2	43.2	15.7	18.8
Undergraduate major							
Business and management	34.3	28,700	28.1	7.1	28.9	19.4	16.5
Education	34.3	23,000	37.4	12.5	25.5	12.9	11.8
Engineering, mathematics, or science	47.1	47,400	19.1	8.7	19.2	19.7	33.4
Humanities or social sciences	52.8	32,800	20.0	8.5	31.3	21.5	18.7
Other	47.4	28,100	21.6	14.3	28.0	22.9	13.2
Amount borrowed (undergraduate)							
Did not borrow	36.0	36,900	19.3	7.7	26.5	22.7	23.9
Less than $5,000	45.7	30,300	29.7	9.5	28.4	17.3	15.2
$5,000–9,999	56.3	29,200	28.0	13.0	31.0	12.9	15.1
$10,000–14,999	54.7	28,700	25.7	16.1	23.6	19.4	15.3
$15,000 or more	58.4	35,100	18.7	9.9	25.3	24.9	21.3
Highest enrollment after bachelor's degree by 2003							
Master's degree	37.9	19,900	30.0	13.2	33.5	18.9	4.5
Doctoral degree	63.1	43,700	18.3	7.6	21.6	21.1	31.5
First-professional degree	71.8	69,200	4.6	2.6	9.4	23.6	59.9
Highest degree earned by 2003							
Bachelor's degree	32.3	23,300	36.2	11.6	26.6	14.4	11.4
Master's degree	45.4	22,900	23.9	12.6	34.9	22.7	5.9
Doctoral degree	64.3	44,100	21.4	9.6	14.7	19.1	35.2
First-professional degree	78.8	75,500	2.0	1.1	8.2	21.6	67.1

NOTE: Graduate includes first-professional. Detail may not sum to totals because of rounding. Estimates include students from the 50 states, DC, and Puerto Rico. Standard error tables are available at http://nces.ed.gov/das/library/reports.asp.

SOURCE: U.S. Department of Education, National Center for Education Statistics, 1993/03 Baccalaureate and Beyond Longitudinal Study (B&B:93/03).

14

and 10.6% of total assets, respectively. According to the latest available comparative deposit data as of June 30, 2011, BNY Mellon obtained a market share of 11.3% or $92.4 billion in a market of $816.4 billion, ranking it 2nd (after JPMorgan Chase) among 139 deposit-taking institutions in its assessment area. It has only 1 location (One Wall Street, New York) inside its market area consisting of Bronx, Kings, Queens, New York, Westchester and Nassau counties.

The following is a summary of BNY Mellon's loan portfolio in its U.S. offices, based on Schedule RC-C of BNY Mellon's December 31, 2009, December 31,2010 and June 30, 2011 Call Reports:

	12/31/2009		12/31/2010		6/30/2011	
Loan type	$000's	%	$000's	%	$000's	%
Real estate loans						
1-4 family residential mortgage	4,151,000	25.0	4,090,000	27.5	3,925,000	24.9
Commercial Mortgage	361,000	2.2	331,000	2.2	284,000	1.8
Multifamily residential (5 or more)	380,000	2.3	327,000	2.2	327,000	2.1
Construction Loans	173,000	1.0	148,000	1.0	170,000	1.1
Commercial & Industrial	555,000	3.3	497,000	3.3	438,000	2.8
Consumer Loans	63,000	0.4	69,000	0.5	66,000	0.4
Lease financing receivables	3,508,000	21.1	3,147,000	21.2	2,780,000	17.7
Other loans						
Loans to purchase securities	1,844,000	11.1	1,148,000	7.7	2,754,000	17.5
Loans to nondepository Fin.Inst.	4,958,000	29.9	4,512,000	30.3	4,207,000	26.7
All other Loans	611,000	3.7	602,000	4.0	799,000	5.1
Total Gross Loans	16,604,000	100.0	14,871,000	100.0	15,750,000	100.0

As a wholesale bank, BNY Mellon does not directly originate retail consumer loans or small business loans, other than to accommodate high net worth clients and/or employees. As illustrated in the above chart, the largest loan concentrations were in loans secured by real estate (approximately 30%), lease financing receivables (17.7%), and other non-consumer loans including commercial & industrial and loans to non-depository financial institutions, etc. (52%).

There are no known financial or legal impediments that adversely impacted BNY Mellon's ability to meet the credit needs of its community.

Assessment Area:

BNY Mellon's assessment area is comprised of the Bronx, Kings, New York and Queens Counties in the New York City Region; Westchester and Nassau Counties in the Mid-Hudson and Long Island regions, respectively.

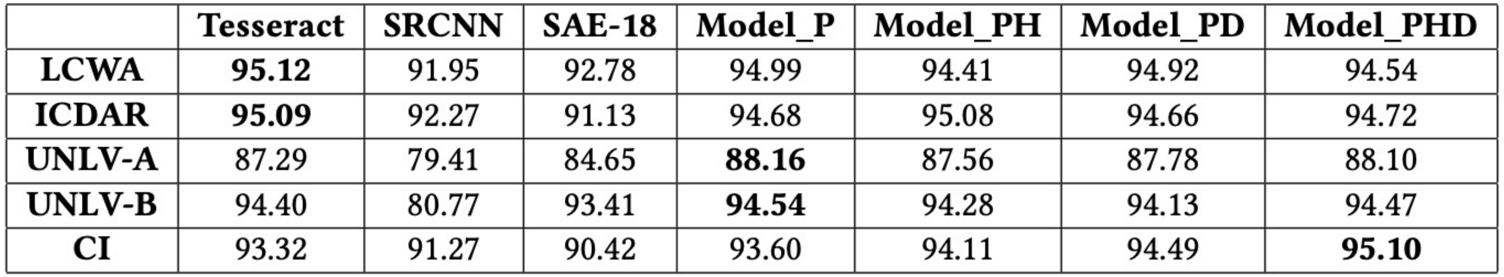
There are 2,605 census tracts in the area, of which 329 are low-income, 643 are moderate-income, 851 are middle-income, 714 are upper-income and 68 are tracts with no income indicated.

3 - 2

Visual Illustration of Tesseract (a and c) and Proposed (b and d) Page Segmentation Pipeline on 72 D

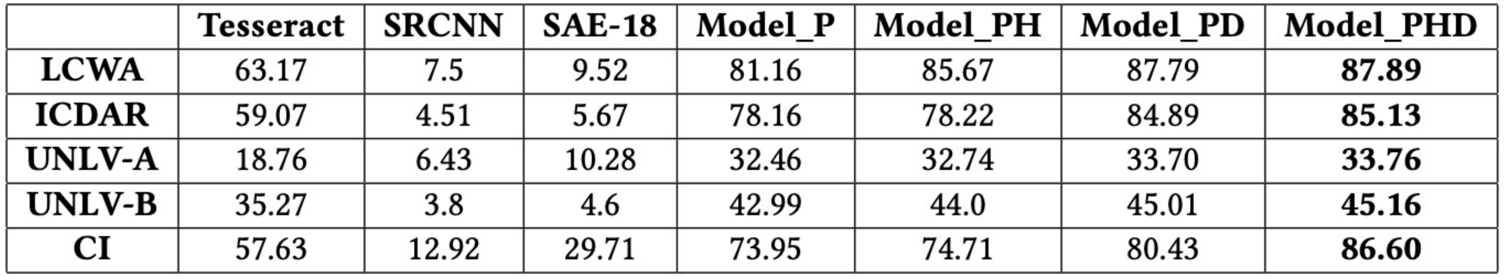
Results

Compare the OCR accuracy by matching the tokens extracted from the processed image with those from GT at the corresponding location returned by Tesseract Engine



	Tesseract	SRCNN	SAE-18	Model_P	Model_PH	Model_PD	Model_PHD
LCWA	**95.12**	91.95	92.78	94.99	94.41	94.92	94.54
ICDAR	**95.09**	92.27	91.13	94.68	95.08	94.66	94.72
UNLV-A	87.29	79.41	84.65	**88.16**	87.56	87.78	88.10
UNLV-B	94.40	80.77	93.41	**94.54**	94.28	94.13	94.47
CI	93.32	91.27	90.42	93.60	94.11	94.49	**95.10**

Table 1: OCR Accuracy Percentage on 300 DPI Images.



	Tesseract	SRCNN	SAE-18	Model_P	Model_PH	Model_PD	Model_PHD
LCWA	63.17	7.5	9.52	81.16	85.67	87.79	**87.89**
ICDAR	59.07	4.51	5.67	78.16	78.22	84.89	**85.13**
UNLV-A	18.76	6.43	10.28	32.46	32.74	33.70	**33.76**
UNLV-B	35.27	3.8	4.6	42.99	44.0	45.01	**45.16**
CI	57.63	12.92	29.71	73.95	74.71	80.43	**86.60**

Table 2: OCR Accuracy Percentage on 72 DPI Images.

Document Intelligence Workshop at KDD 2021

Thank You