Using Image Processing and Deep-Learning to Explore Digitized Historical Documents

A COLLABORATORY BETWEEN THE LIBRARY OF CONGRESS AND THE IMAGE ANALYSIS FOR ARCHIVAL DISCOVERY (AIDA) LAB AT THE UNIVERSITY OF NEBRASKA, LINCOLN, NE

Chulwoo Pack
Yi Liu

Overview

- Part 1: Aida Project: Poem Recognition
  - Part 1.1: Segmentation
  - Part 1.2: Recognition
- Part 2: Document Image Quality Assessment (DIQA)
- Part 3: Zoning
- Part 4: Deep Learning
- Part 5: Five Collaboratory Projects with Library of Congress

AIDA | Objective

- Exploring what more we can do with the millions of images that represent the digitized cultural record—particularly digital images of textual materials—and we are interested in the types of discovery that serious attention to digital images might yield
- Generate data about *visual features* from the newspaper pages and then use those extracted features within a computational system, such as artificial neural network

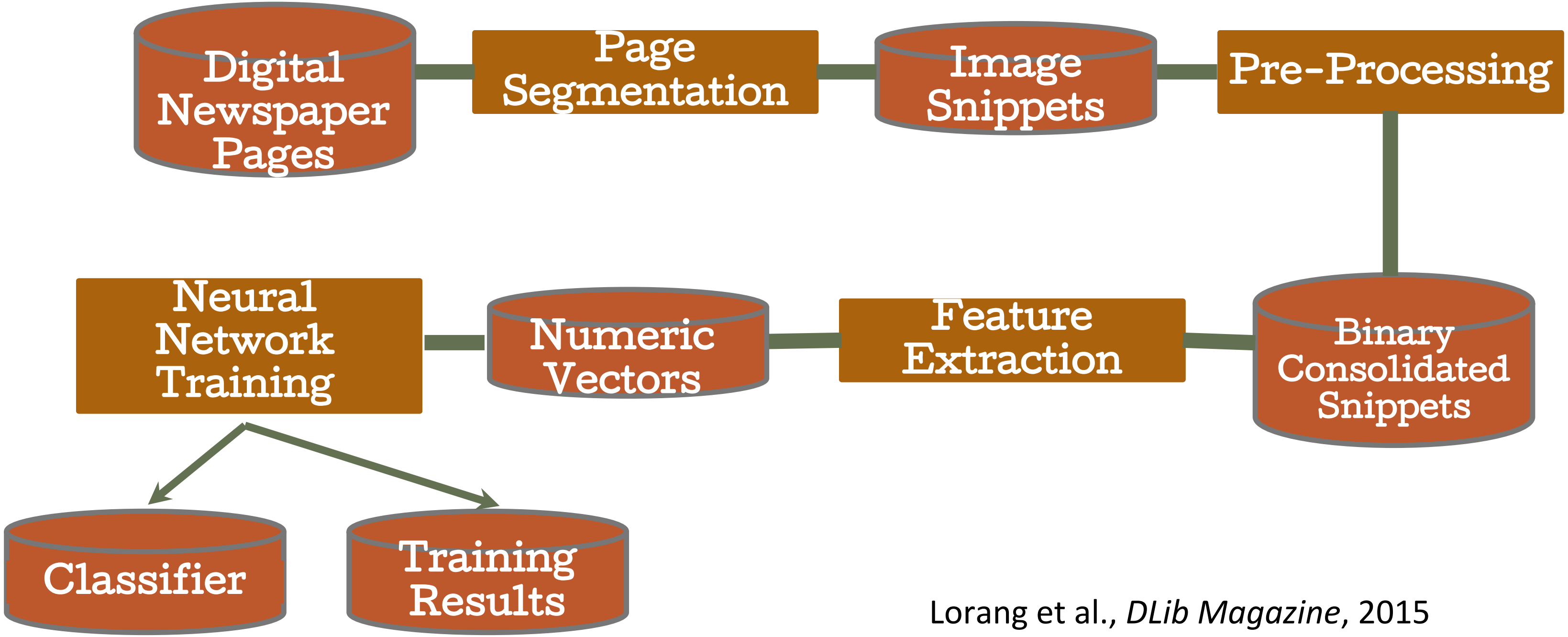
Aida

Part 1: Poem Recognition

Objectives | Identifying existence of poem in a page

Applications | metadata generation, discover-/search-ability, visualization, etc.

Poem Recognition | Workflow



Lorang et al., *DLib Magazine*, 2015

Poem Recognition |Segmentation

INTUITIVE STRATEGY

- Generate page image "snippets"
  - find the newspaper columns present on the page
  - cut each column into a series of column snippets of a fixed width:height ratio

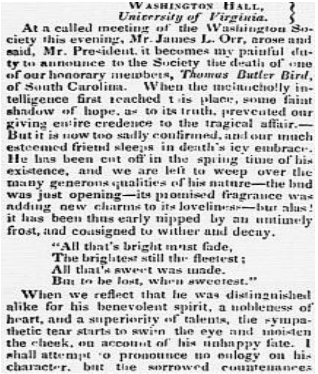
- Take the snippet, determine whether it featured poetic content, and the determine more locally where on the page the poetic content appeared

Aida

Poem Recognition | Segmentation

HOWEVER …

- ❑ **Noticed a variety of factors influence our ability to create good image snippets**



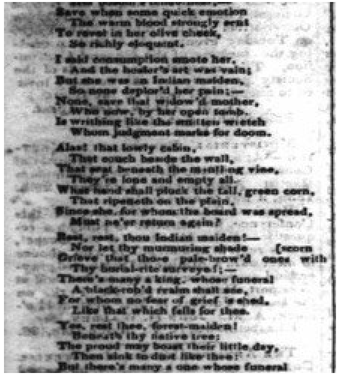
WASHINGTON HALL,
University of Virginia.

At a called meeting of the Washington Society this evening, Mr. James L. Orr, arose and said, Mr. President, it becomes my painful duty to announce to the Society the death of one of our honorary members, *Thomas Butler Bird*, of South Carolina. When the melancholy intelligence first reached this place, some faint shadow of hope, as to its truth, prevented our giving entire credence to the tragical affair.—But it is now too sadly confirmed, and our much esteemed friend sleeps in death's icy embrace. He has been cut off in the spring time of his existence, and we are left to weep over the many generous qualities of his nature—the bud was just opening—its promised fragrance was adding new charms to its loveliness—but alas! it has been thus early nipped by an untimely frost, and consigned to wither and decay.

"All that's bright must fade,
The brightest still the fleetest;
All that's sweet was made,
But to be lost, when sweetest."

When we reflect that he was distinguished alike for his benevolent spirit, a nobleness of heart, and a superiority of talents, the sympathetic tear starts to swim the eye and moisten the cheek, on account of his unhappy fate. I shall attempt to pronounce no eulogy on his character, but the sorrowed countenances

good quality



Save when some quick emotion
The warm blood strongly sent
To revel in her olive cheek,
So richly eloquent.

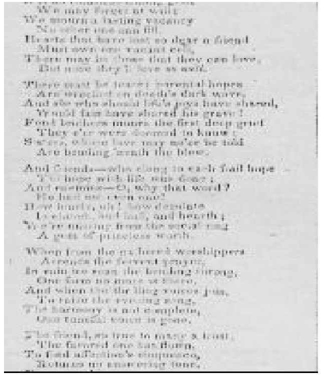
I said consumption smote her,
And the healer's art was vain;
But she was an Indian maiden,
So none deplor'd her pain;
None, save that widow'd mother,
Who sees, by her open tomb,
Is writhing like the smitten wretch
Whom judgment marks for doom.

Alas! that lowly cabin,
That couch beside the wall,
That seat beneath the mantling vine,
They're lone and empty all.
What hand shall pluck the tall, green corn,
That ripeneth on the plain,
Since she, for whom the board was spread,
Must ne'er return again?

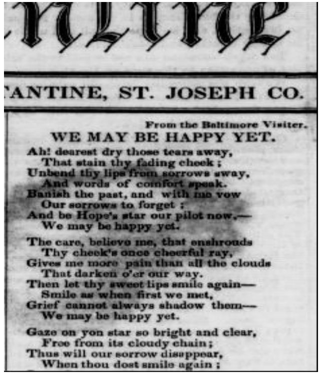
Rest, rest, thou Indian maiden!—
Nor let thy murmuring shade [scorn
Grieve that those pale-brow'd ones with
Thy burial-site surveyed;—
There's many a king, whose funeral
A black-rob'd realm shall see,
For whom no tear of grief is shed,
Like that which falls for thee.

Yes, rest thee, forest-maiden!
Beneath thy native tree;
The proud may boast their little day,
Then sink to dust like thee;
But there's many a one whose funeral

bleed-through



low contrast



ANTINE, ST. JOSEPH CO.

From the Baltimore Visiter.
WE MAY BE HAPPY YET.

Ah! dearest dry those tears away,
That stain thy fading cheek;
Unbend thy lips from sorrows sway,
And words of comfort speak.
Banish the past, and with me vow
Our sorrows to forget;
And be Hope's star our pilot now,—
We may be happy yet.

The care, believe me, that enshrouds
Thy cheek's once cheerful ray,
Gives me more pain than all the clouds
That darken o'er our way.
Then let thy sweet lips smile again—
Smile as when first we met,
Grief cannot always shadow them—
We may be happy yet.

Gaze on yon star so bright and clear,
Free from its cloudy chain;
Thus will our sorrow disappear,
When thou dost smile again;

occluding "blobs"

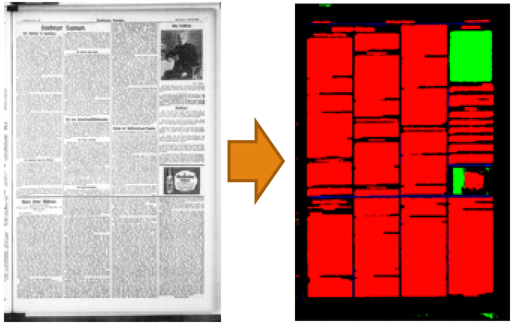
Aida

Poem Recognition |Segmentation

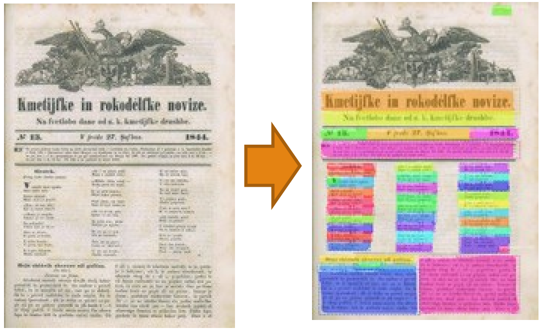
ONGOING STRATEGIES

- More sophisticated traditional image processing techniques; Connected component analysis (CCA), Voronoi-diagram
- Deep-learning-based approach; dhSegment, Mask-RCNN

CCA + Voronoi-diagram



dhSegment



Mask-RCNN

Poem Recognition | Recognition

Which one has poem?

Her beaming face seemed formed to bless—
Her eyes bespoke a soul of worth—
First at the shrine of knowledge bent—
First at the altar of her God—
On Virtue's arm she proudly leant
As up bright wisdom's path she trod.

The centre gem, the pearl of price,
Amid inferior jewels set :—
Her brightness dim'd alluring vice—
Her sweetness swept away regret—
The pure were gladdened by her smile—
The noblest her affections sought—
Her youthful bosom knew no guile—
Her generous mind no damning thought.

Yet shades of grief would often come
Across her spirit, at the hour
When wild Bees cease their drowsy hum,
And evening closed the tender flower ;
Then would she wander from the rest
By Hudson's sleeping moon-lit wave,
And weep for her, whose guilty breast
Had sent her forth the world to brave.

Years rolled, and time had lulled to sleep,
The deep emotions of her soul,
And tho' she oft went forth to weep,
Her reason held its high control—
A few short years—and far away
She hoped to spend life's gloomy hours,
And list to nature's music play,
And rest amid the fairest bowers.

him in reporting the bill.

Immediately upon the appointment of the committee, and the reference to it of the important subjects treated of in the Message of the President, and the Report of the Secretary of the Treasury, the committee found that the Treasury of the United States was, very soon, to be in want of means to meet the current demands upon it, without regard to any further transfer to the States.—They also found that this fourth instalment of the deposites with the States was to become payable on the first day of October, and amounted to about nine and one-third millions of dollars.

The state of the Treasury, as developed by the Report of the Secretary of the Treasury, was, as he now recollected, and he thought he could not be materially mistaken, that, at the time when the statement appended to that report was made up, about the first day of the present month, (he believed the exact date was the 28th of August,) there was in the Treasury, subject to draft, available and unavailable, but eight millions one hundred and some odd thousand dollars. The report was printed, and upon the table of every Senator, and would verify his correctness in this particular. This amount was exclusive of the sums already deposited with the States, being some twenty-eight millions.

To arrive at what would be the condition of the Treasury on the first of October, the expenses of the present month, which, from drafts already made and anticipated, were estimated at about two and a half millions, must be deducted from the eight millions, one hundred and odd thousands ; thus leaving in the Treasury, subject to draft, on the first day of October, less than six millions, without the transfer of a dollar to the States towards the October instalment. This, too, included all the funds in the Treasury subject to draft for payments, or transfers to the States, whether available or not, upon the drafts of the Treasurer ; the funds on de-

Aida

Poem Recognition |Recognition | Basis of Features

Feature Extraction

- **Left column width**
  - length of background pixels prior to the first object pixel for each row
- **Right column width**
  - length of background pixels after the final object pixel for each row
- **Row depth**
  - number of each sequence of continuous background pixels in each column
- **Margin statistics**
  - computed from the list of the Left Column Widths

Aida

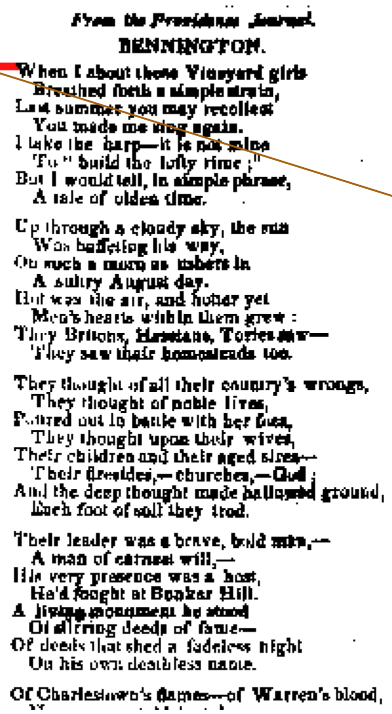
Poem Recognition |Recognition | Basis of Features

Feature Extraction

- **Jaggedness statistic**
  - measures the number of background pixels after the final object pixel in each row
- **Stanza statistic**
  - looking for gap between stanzas using a list of Row Depths
- **Row length statistic**
  - length of continuous sequence of object pixels

Aida

Poem Recognition |Recognition | Basis of Features


From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
Breathed forth a simple strain,
Last summer you may recollect
You made me sing again.
I take the harp—it is not mine
To " build the lofty rime ;"
But I would tell, in simple phrase,
A tale of olden time.

Up through a cloudy sky, the sun
Was buffeting his way,
On such a morn as ushers in
A sultry August day.
Hot was the air, and hotter yet
Men's hearts within them grew :
They Britons, Hessians, Tories saw—
They saw their homesteads too.

They thought of all their country's wrongs,
They thought of noble lives,
Poured out in battle with her foes,
They thought upon their wives,
Their children and their aged sires—
Their firesides,—churches,—God ;
And the deep thought made hallowed ground,
Each foot of soil they trod.

Their leader was a brave, bold man,—
A man of earnest will,—
His very presence was a host,
He'd fought at Bunker Hill.
A living monument he stood
Of stirring deeds of fame—
Of deeds that shed a fadeless light
On his own deathless name.

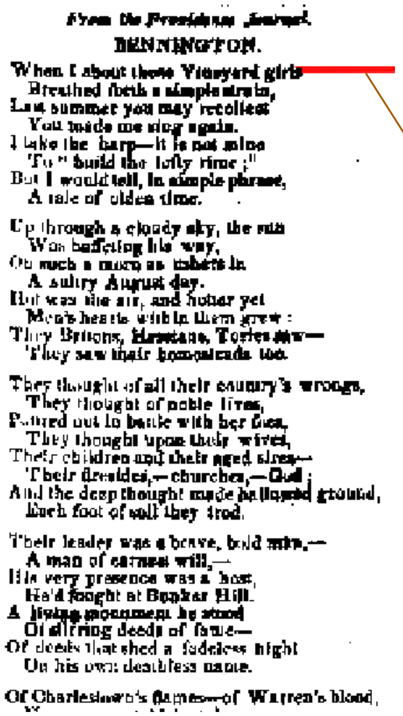
Of Charlestown's flames—of Warren's blood,

Left Column Widths

Length of background pixels prior to the first object pixel for each row

Aida

Poem Recognition | Recognition | Basis of Features



From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
Breathed forth a simple strain,
Last summer you may recollect
You made me sing again.
I take the harp—it is not mine
"To build the lofty rhyme;"
But I would tell, in simple phrase,
A tale of olden time.

Up through a cloudy sky, the sun
Was buffeting his way,
On such a morn as ushers in
A sultry August day.
Hot was the air, and hotter yet
Men's hearts within them grew:
They Britons, Hessians, Tories saw—
They saw their homesteads too.

They thought of all their country's wrongs,
They thought of noble lives,
Poured out in battle with her foes,
They thought upon their wives,
Their children and their aged sires—
Their firesides,—churches,—God;
And the deep thought made hallowed ground,
Each foot of soil they trod.

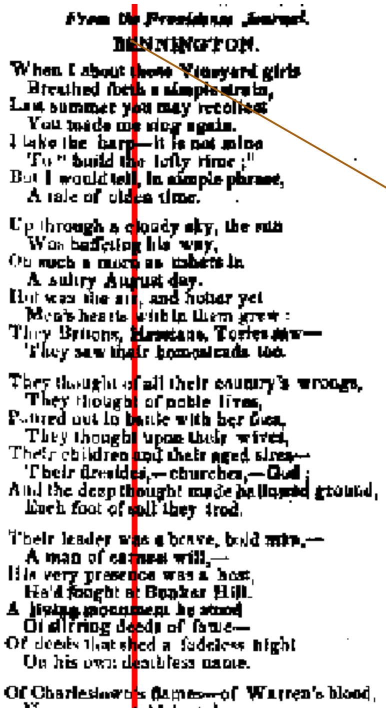
Their leader was a brave, bold man,—
A man of earnest will,—
His very presence was a host,
He'd fought at Bunker Hill.
A living monument he stood
Of stirring deeds of fame—
Of deeds that shed a fadeless light
On his own deathless name.

Of Charlestown's flames—of Warren's blood,

Right Column Widths

Length of background pixels after the final object pixel for each row

Poem Recognition |Recognition | Basis of Features


From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
Breathed forth a simple strain,
Last summer you may recollect
You made me sing again.
I take the harp—it is not mine
To "build the lofty rime;"
But I would tell, in simple phrase,
A tale of olden time.

Up through a cloudy sky, the sun
Was battling his way,
On such a morn as ushers in
A sultry August day.
Hot was the air, and hotter yet
Men's hearts within them grew:
They Britons, Hessians, Tories saw—
They saw their homesteads too.

They thought of all their country's wrongs,
They thought of noble lives,
Poured out in battle with her foes,
They thought upon their wives,
Their children and their aged sires—
Their firesides,—churches,—God;
And the deep thought made hallowed ground,
Each foot of soil they trod.

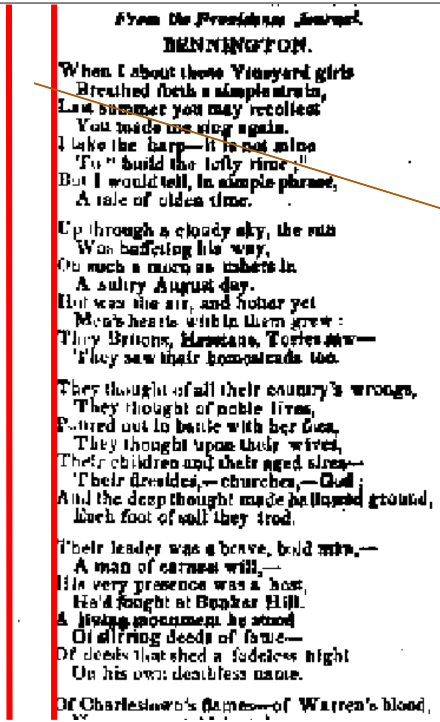
Their leader was a brave, bold man,—
A man of earnest will,—
His very presence was a host,
He'd fought at Bunker Hill.
A living monument he stood
Of stirring deeds of fame—
Of deeds that shed a fadeless light
On his own deathless name.

Of Charlestown's flames—of Warren's blood,

Row Depths

the number of each sequence of continuous background pixels in each column

Poem Recognition |Recognition | Basis of Features


From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
Breathed forth a simple strain,
Last summer you may recollect
You made me sing again.
I take the harp—it is not mine
To "build the lofty rime;"
But I would tell, in simple phrase,
A tale of olden time.

Up through a cloudy sky, the sun
Was buffeting his way,
On such a morn as ushers in
A sultry August day.
Hot was the air, and hotter yet
Men's hearts within them grew:
They Britons, Hessians, Tories saw—
They saw their homesteads too.

They thought of all their country's wrongs,
They thought of noble lives,
Poured out in battle with her foes,
They thought upon their wives,
Their children and their aged sires—
Their firesides,—churches,—God;
And the deep thought made hallowed ground,
Each foot of soil they trod.

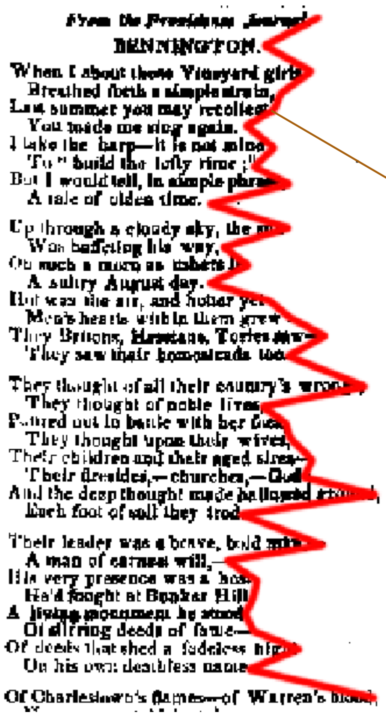
Their leader was a brave, bold man,—
A man of earnest will,—
His very presence was a host,
He'd fought at Bunker Hill.
A living monument he stood
Of stirring deeds of fame—
Of deeds that shed a fadeless light
On his own deathless name.

Of Charlestown's flames—of Warren's blood,

Margin statistics

Computed from the list of the Left Column Widths

Poem Recognition |Recognition | Basis of Features


From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
Breathed forth a simple strain,
Last summer you may recollect
You made me sing again.
I take the harp—it is not mine
To "build the lofty rime;"
But I would tell, in simple phrase
A tale of olden time.

Up through a cloudy sky, the [illegible]
Was battling his way,
On such a morn as [illegible]
A sultry August day.
Hot was the air, and hotter yet
Men's hearts within them grew—
They Britons, Hessians, Tories saw—
They saw their homesteads too.

They thought of all their country's wrongs,
They thought of noble lives
Poured out in battle with her [illegible]
They thought upon their wives,
Their children and their aged sires—
Their firesides,—churches,—God
And the deep thought made hallowed ground,
Each foot of soil they trod.

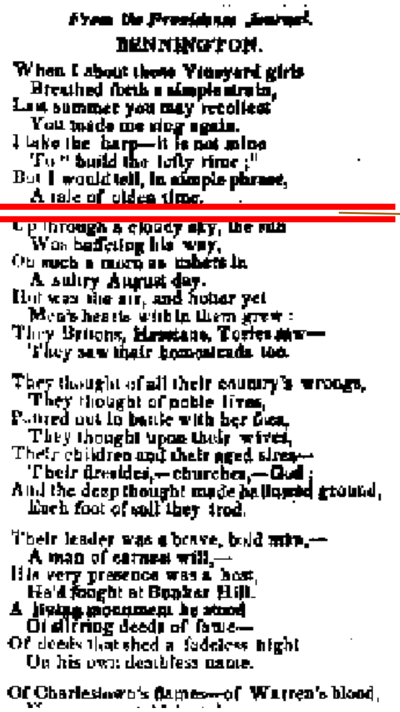
Their leader was a brave, bold man—
A man of earnest will,—
His very presence was a host,
He'd fought at Bunker Hill
A living monument he stood
Of stirring deeds of fame—
Of deeds that shed a fadeless light
On his own deathless name

Of Charlestown's flames—of Warren's blood,

Jaggedness statistics

measures the number of background pixels after the final object pixel in each row

Poem Recognition |Recognition | Basis of Features


From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
 Breathed forth a simple strain,
Last summer you may recollect
 You made me sing again.
I take the harp—it is not mine
 To "build the lofty rime;"
But I would tell, in simple phrase,
 A tale of olden time.

Up through a cloudy sky, the sun
 Was buffeting his way,
On such a morn as ushers in
 A sultry August day.
Hot was the air, and hotter yet
 Men's hearts within them grew:
They Britons, Hessians, Tories saw—
 They saw their homesteads too.

They thought of all their country's wrongs,
 They thought of noble lives,
Poured out in battle with her foes,
 They thought upon their wives,
Their children and their aged sires—
 Their firesides,—churches,—God;
And the deep thought made hallowed ground,
 Each foot of soil they trod.

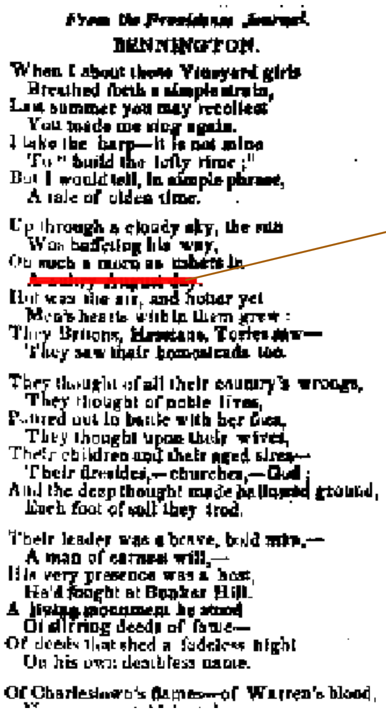
Their leader was a brave, bold man,—
 A man of earnest will,—
His very presence was a host,
 He'd fought at Bunker Hill.
A living monument he stood
 Of stirring deeds of fame—
Of deeds that shed a fadeless light
 On his own deathless name.

Of Charlestown's flames—of Warren's blood,

Stanza statistics

looking for gap between stanzas using a list of Row Depths

Poem Recognition |Recognition | Basis of Features



From the Providence Journal.

BENNINGTON.

When I about those Vineyard girls
Breathed forth a simple strain,
Last summer you may recollect
You made me sing again.
I take the harp—it is not mine
To "build the lofty rime;"
But I would tell, in simple phrase,
A tale of olden time.

Up through a cloudy sky, the sun
Was buffeting his way,
On such a morn as ushers in
A sultry August day.
Hot was the air, and hotter yet
Men's hearts within them grew:
They Britons, Hessians, Tories saw—
They saw their homesteads too.

They thought of all their country's wrongs,
They thought of noble lives,
Poured out in battle with her foes,
They thought upon their wives,
Their children and their aged sires—
Their firesides,—churches,—God;
And the deep thought made hallowed ground,
Each foot of soil they trod.

Their leader was a brave, bold man,—
A man of earnest will,—
His very presence was a host,
He'd fought at Bunker Hill.
A living monument he stood
Of stirring deeds of fame—
Of deeds that shed a fadeless light
On his own deathless name.

Of Charlestown's flames—of Warren's blood,

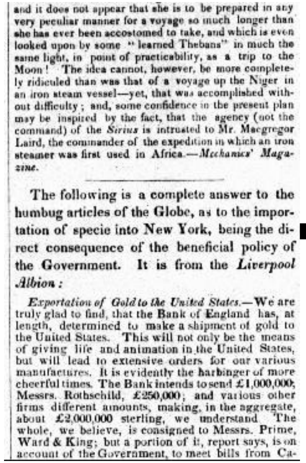
Row Length statistics

Length of continuous sequence of object pixels

Poem Recognition |Recognition

Snippet pre-processing

1. Otsu's binarization [Otsu, *IEEE TSMC* 1979]
2. Consolidation [Soh, *IAAI* 2018]

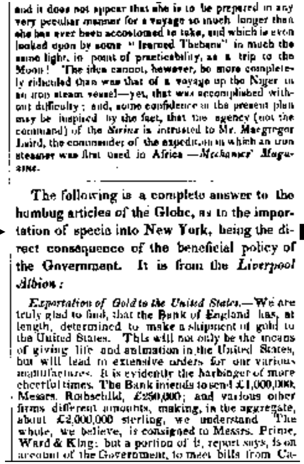


and it does not appear that she is to be prepared in any very peculiar manner for a voyage so much longer than she has ever been accostomed to take, and which is even looked upon by some "learned Thebans" in much the same light, in point of practicability, as a trip to the Moon! The idea cannot, however, be more completely ridiculed than was that of a voyage up the Niger in an iron steam vessel—yet, that was accomplished without difficulty; and, some confidence in the present plan may be inspired by the fact, that the agency (not the command) of the *Sirius* is intrusted to Mr. Macgregor Laird, the commander of the expedition in which an iron steamer was first used in Africa.—*Mechanics' Magazine.*

The following is a complete answer to the humbug articles of the Globe, as to the importation of specie into New York, being the direct consequence of the beneficial policy of the Government. It is from the *Liverpool Albion:*

Exportation of Gold to the United States.—We are truly glad to find, that the Bank of England has, at length, determined to make a shipment of gold to the United States. This will not only be the means of giving life and animation in the United States, but will lead to extensive orders for our various manufactures. It is evidently the harbinger of more cheerful times. The Bank intends to send £1,000,000; Messrs. Rothschild, £250,000; and various other firms different amounts, making, in the aggregate, about £2,000,000 sterling, we understand. The whole, we believe, is consigned to Messrs. Prime, Ward & King; but a portion of it, report says, is on account of the Government, to meet bills from Ca-

Snippet image



Binary Snippet

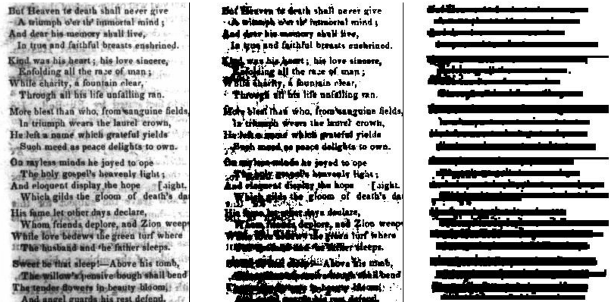
Consolidated snippet

Poem Recognition |Recognition

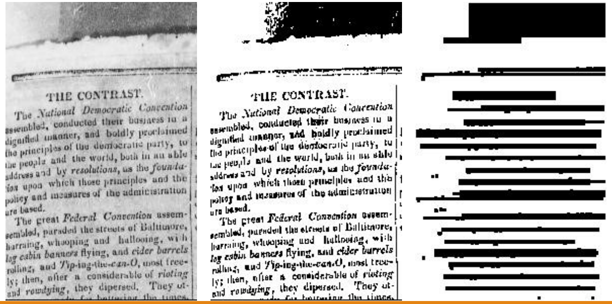
◦ Performance of feature extraction could be affected by various types of noise

Range Effects

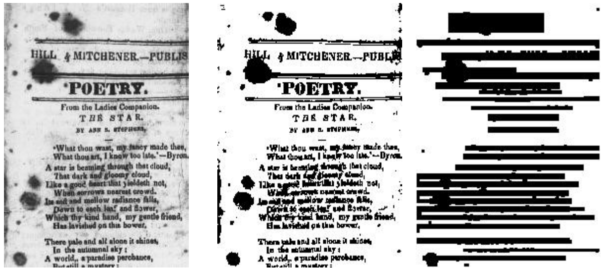
Skewed Orientation



Bleed-Through

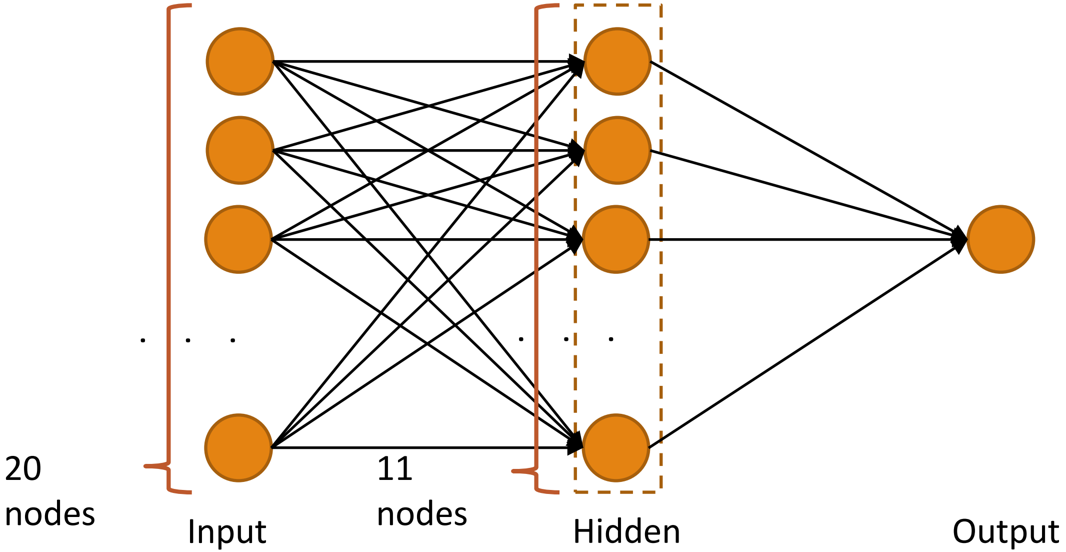


Blobs

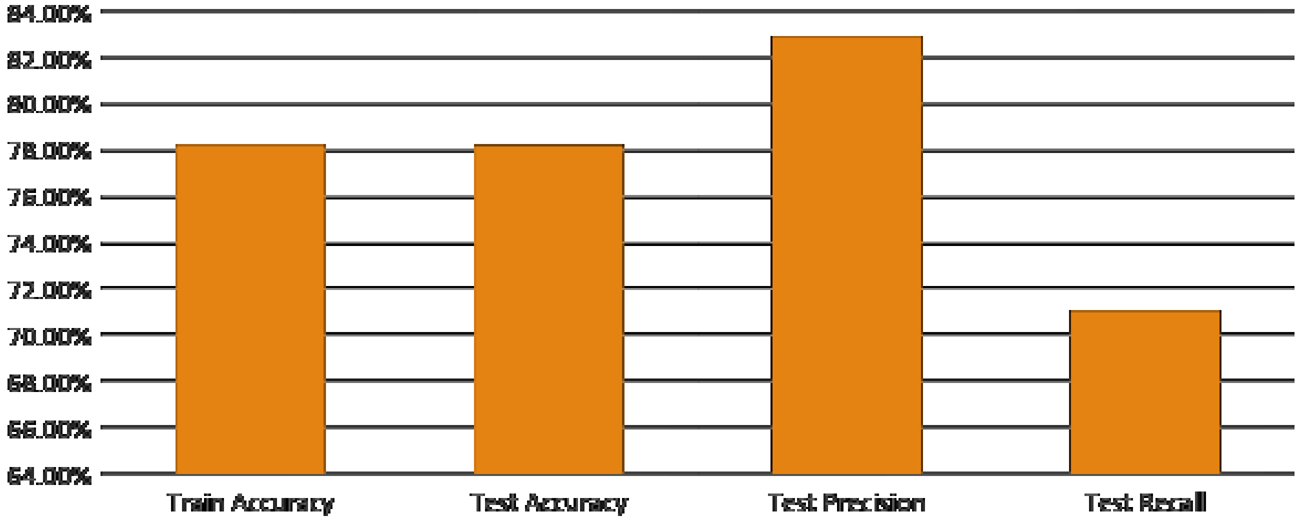


Poem Recognition |Recognition

ANN implementation from the WEKA Workbench [Eibe et al. 2016]



Poem Recognition |Recognition | performance



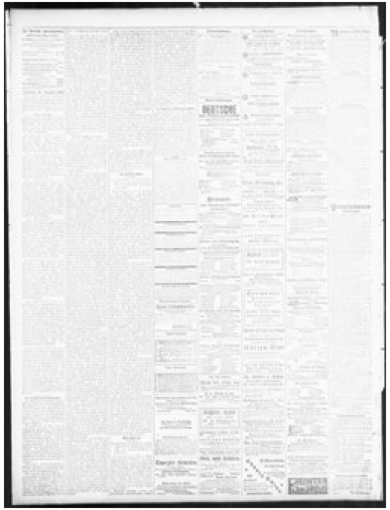
Part 2: Document Image Quality Assessment (DIQA)

Objectives | Measure visual quality of document image

Applications | metadata generation, image quality enhancement, etc.

DIQA | Objective

- Measure four main degradations inherent in digitized historical document images
- Analyze these measures in a large-scale dataset (i.e., Chronicling America) and interpret what they are saying



Contrast

THE BRYAN DAILY EAGLE

ED HALL.

THE GROCER

At Jno. B. Mike's

ALL KINDS OF

Country Produce

At Jno. B. Mike's

Range-effect

AVISOS.

CATECISMO

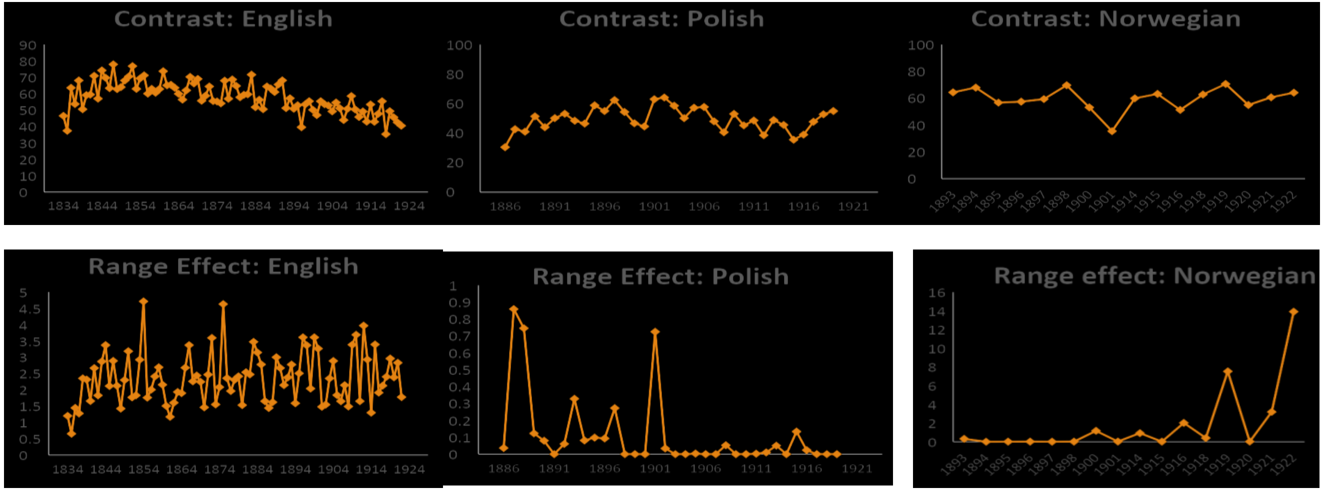
Sistema-métrico decimal

Bleed-through

Skewness

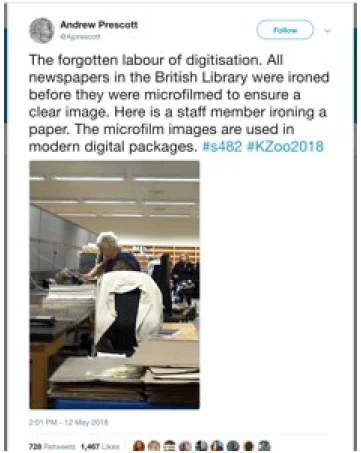
DIQA |Contrast, Range Effect

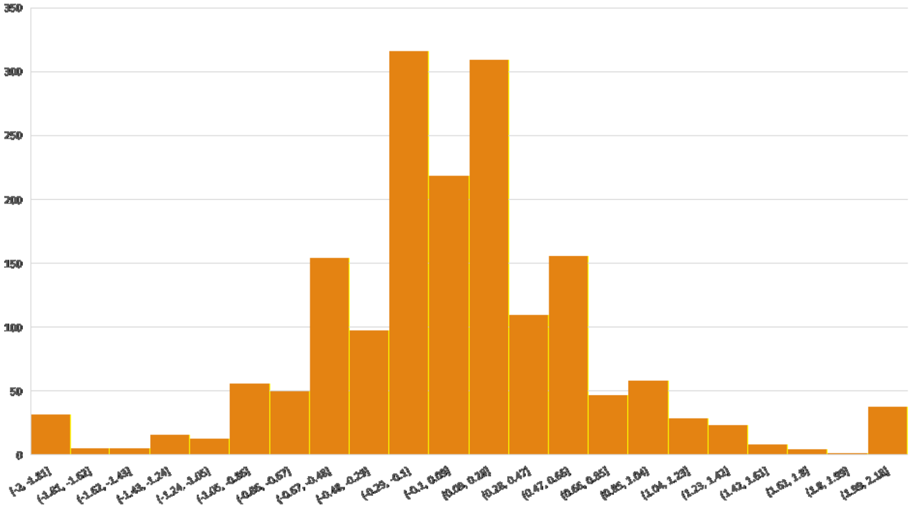
- ❑ Contrast in all languages is pretty consistent; nor does it change drastically over time
- ❑ Range effect, on the other hand, not only varies across the different languages, it also changes over time for each language



DIQA |Orientation Skew

- A more effective measure is likely to be local skew, relative no particular parts of the page, or other measures of warpedness or beveled nature of the page





Distribution of orientation skew

DIQA | Noisiness

- Assessing effects of bleed-through, blobs (e.g., stains), and other non-textual artifacts

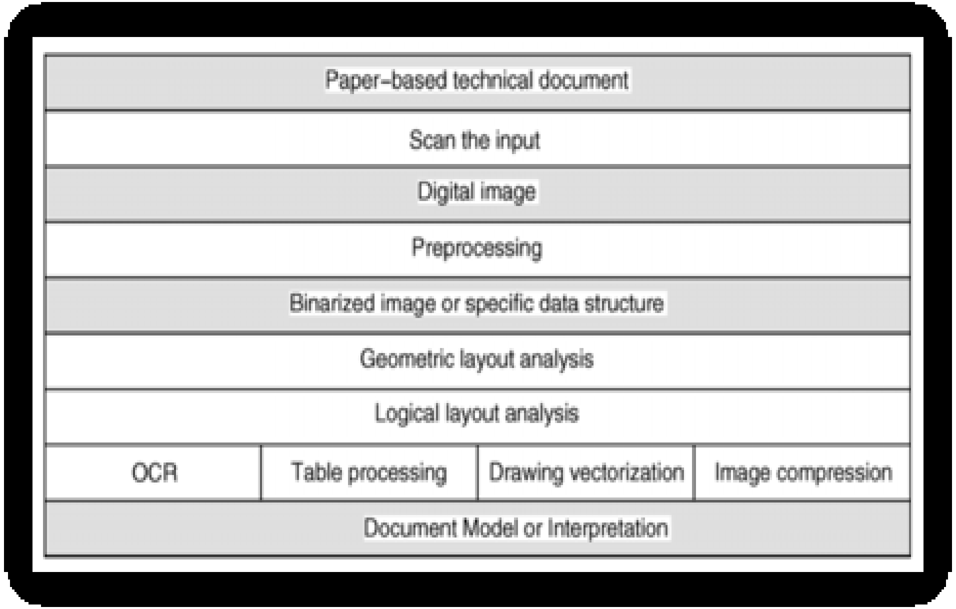
- Defects or degradations of a page, or of the digitization process based on histogram analysis—of pixels’ intensity values—of each page

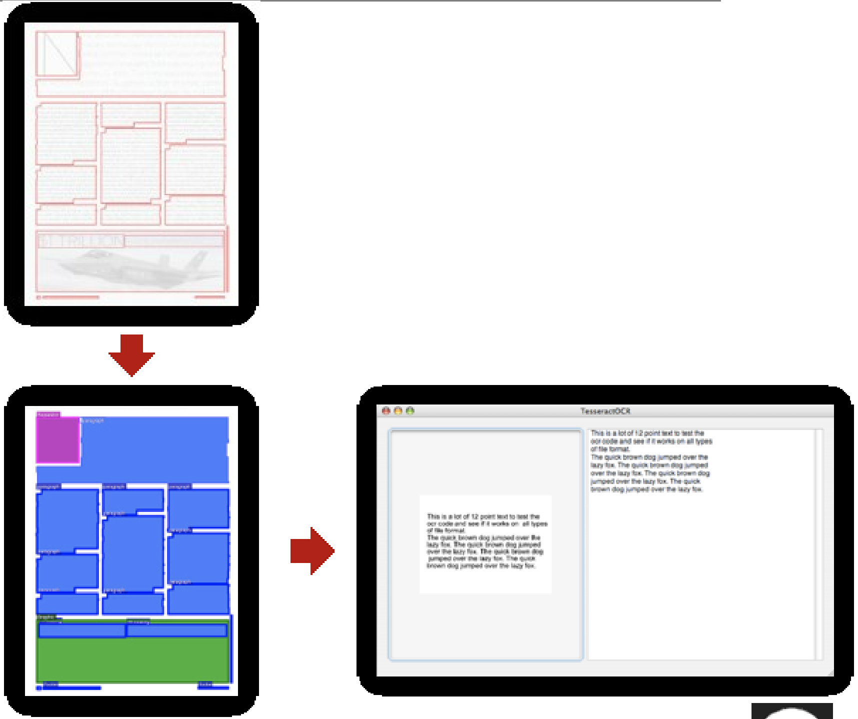
Part 3: Zoning

Objectives | Segment an image into meaningful sub-regions

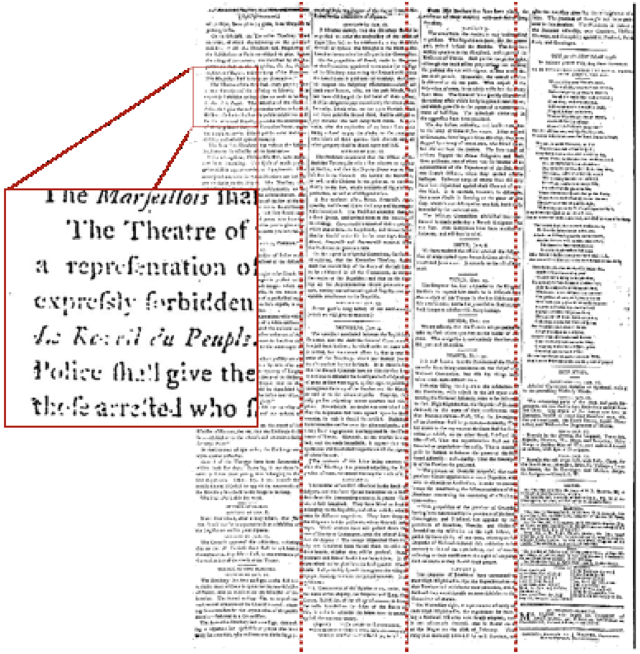
Applications | Object localization, visualization, logical layout analysis, etc.

Zoning | Background

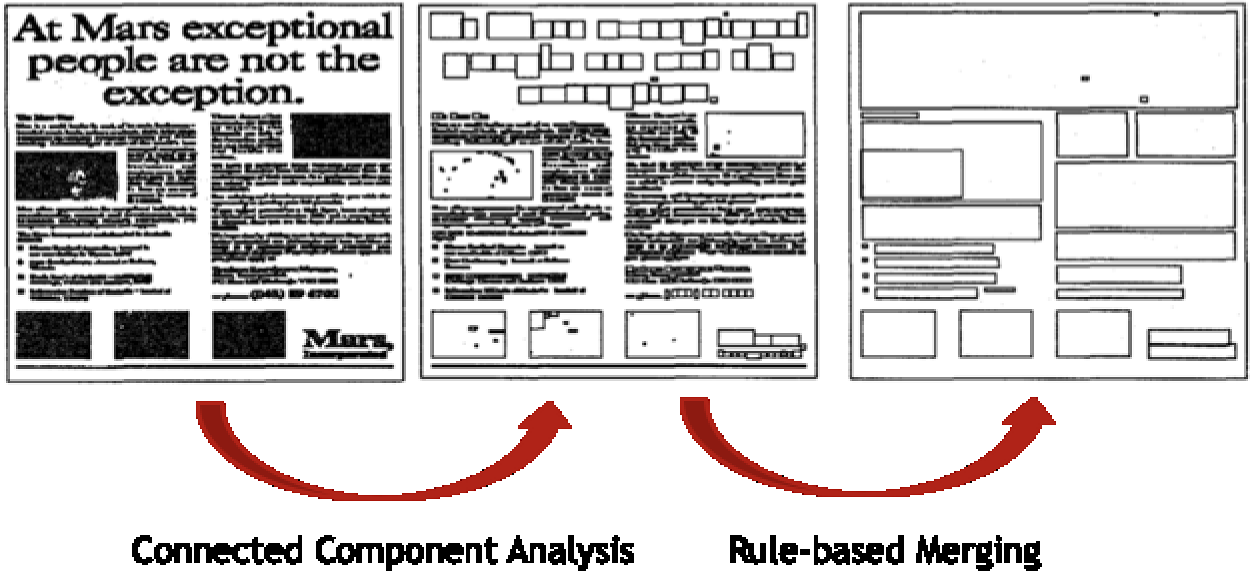




Zoning | Challenges

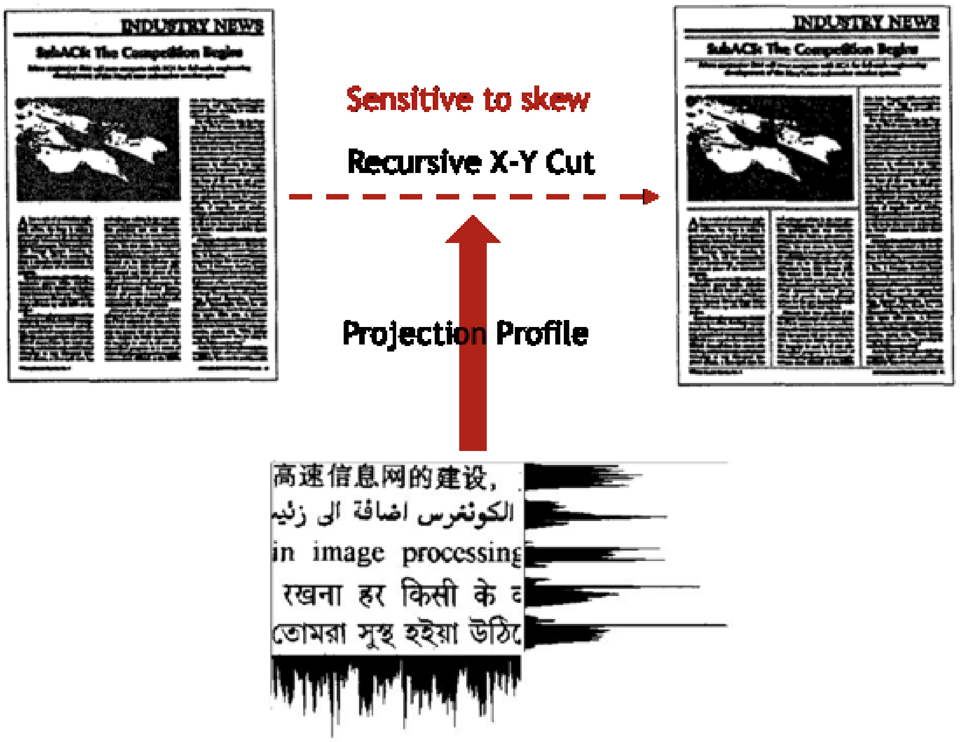


Zoning | Traditional Approaches (Bottom-up)

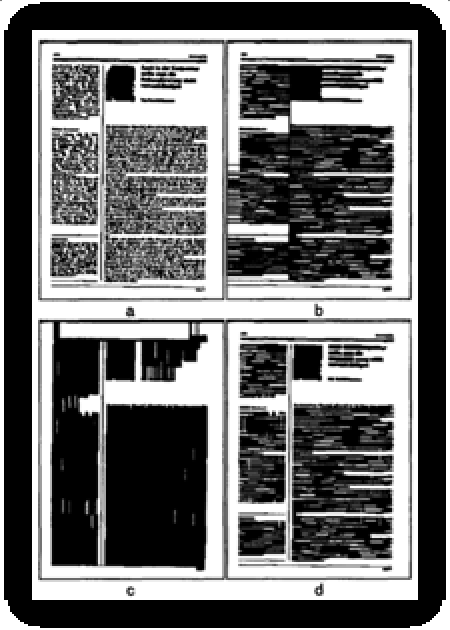


Aida

Zoning | Traditional Approaches (Top-down)



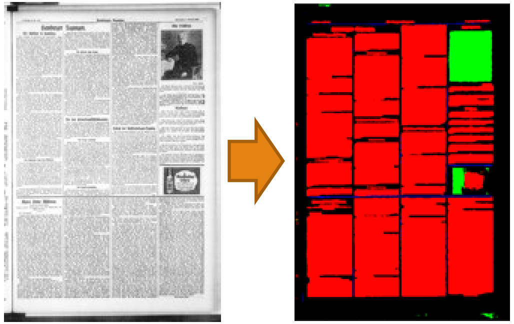
Zoning |Traditional Approaches (Hybrid)



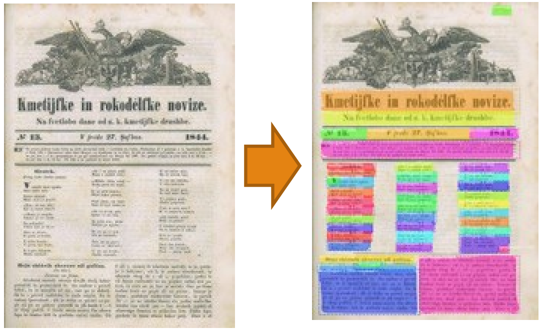
- Over-segmentation using RXYC + Merging sub-regions
- Bottom-up merging + Top-down RXYC

Zoning | State-of-the-art Approaches (Deep Learning)

- With the advent of deep learning, it has been shown that using data-driven features, instead of hand-crafted features, is more effective
- Boundary between physical layout analysis and logical layout analysis becomes ambiguous



dhSegment



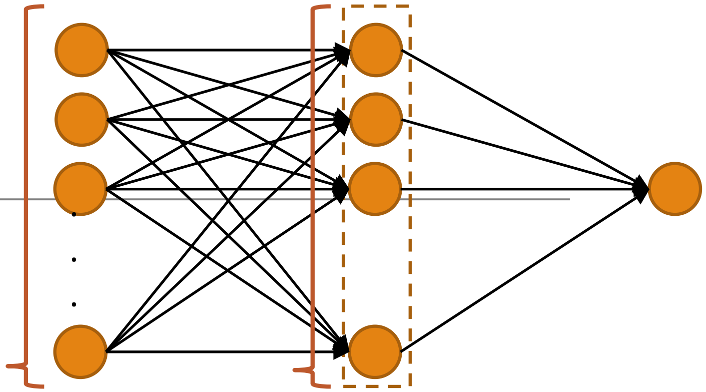
Mask-RCNN

Part 4: Deep Learning

Objectives | Improve the performance of identifying existence of poem in a page

Applications | Automated poetic content collection, article type classification

Deep Learning | Background



Recall the ANN used in Aida project

Generally speaking, Deep Learning is deep structured learning

Hence, *more* hidden layers

Depending on the classification task, there are different models

Recognizing poems in a newspaper page is an image-related classification

Hence, Convolutional Neural Network

Deep Learning | Convolutional Neural Network

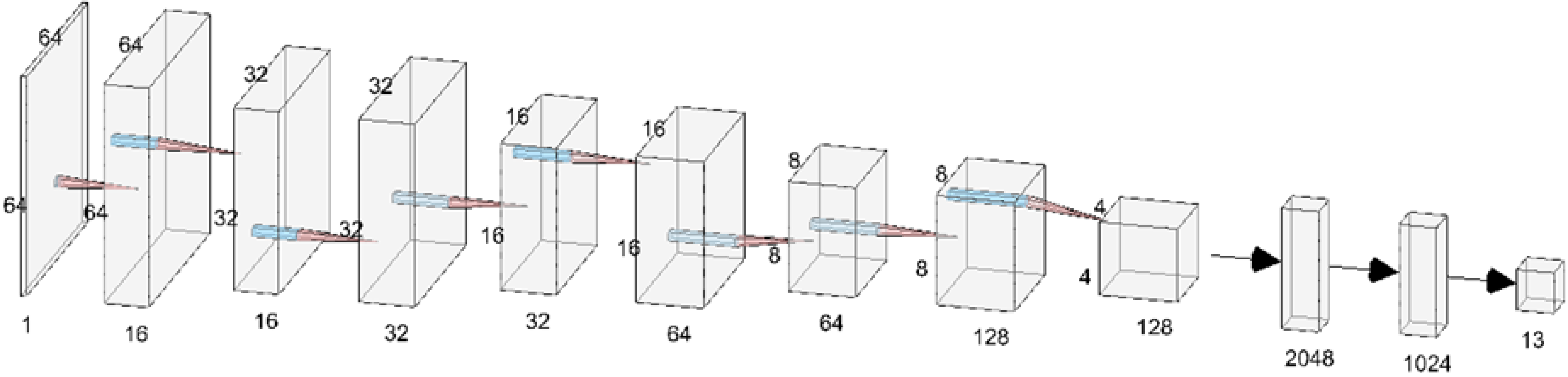
Convolutional Neural Networks (CNN) have been shown to be effective for image-related classification

- ➔ LeNet [LeCun et al.] was the start of deep CNN.
- ➔ AlexNet [Krizhevsky et al.] was inspired by LeNet, and outperformed state-of-art by large percentage on ImageNet.
- ➔ ResNet [He et al.] pushed CNN to a very deep model — 152 layer ResNet.

More and more document image related researches were attracted

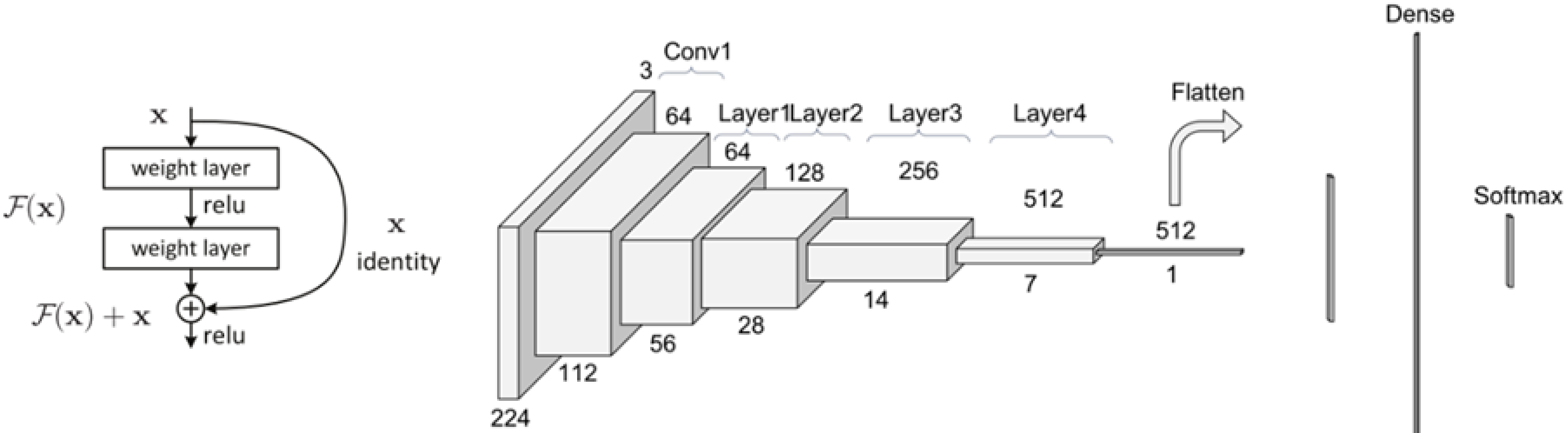
- ➔ Pondenkendath et al. applied ResNet to four tasks: handwritten style, document layout, authorship classification, font identification.

Deep Learning |Convolutional Neural Network



LeNet

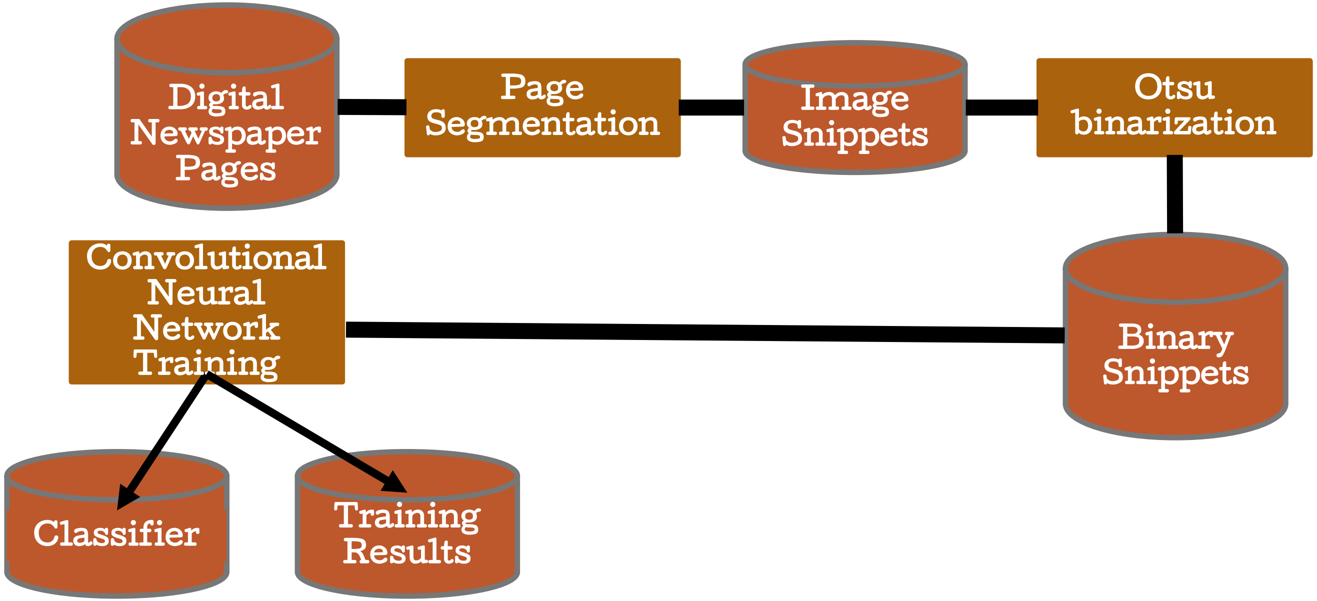
Deep Learning | Convolutional Neural Network



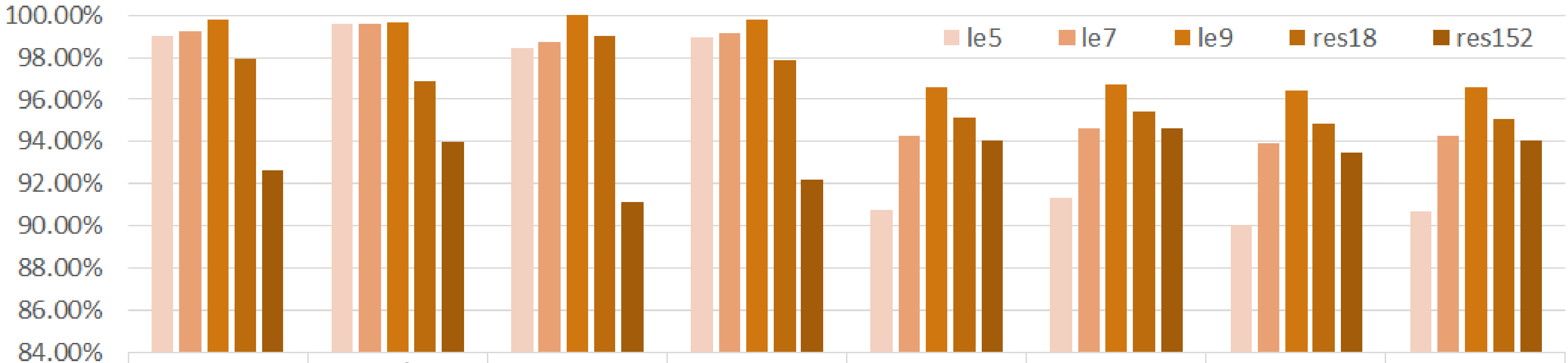
ResNet

Deep Learning |2nd Gen Aida

CNN allows to learn feature from training process

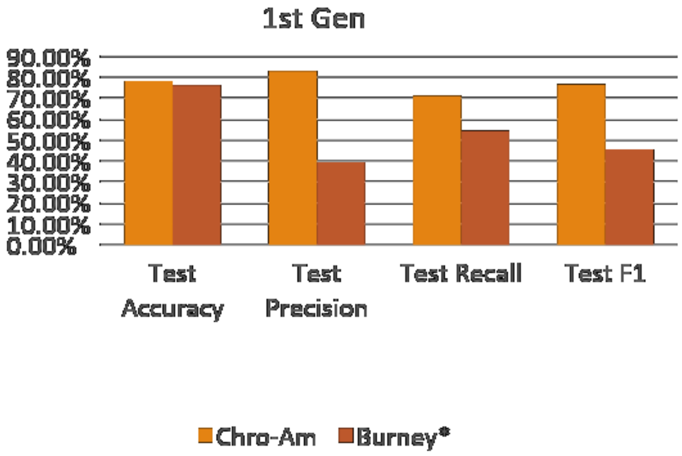


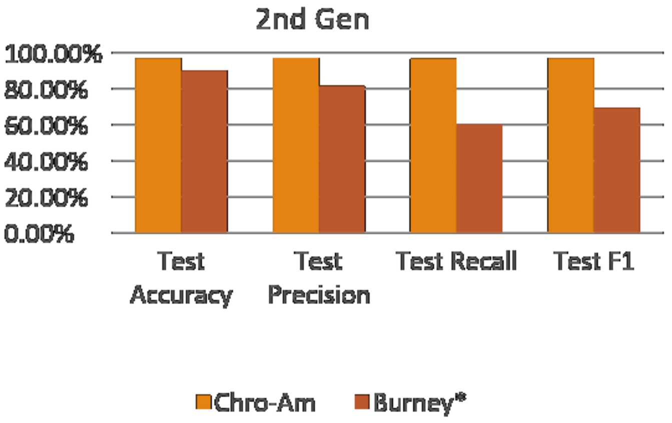
Deep Learning |2nd Gen Aida



	train accuracy	train precision	train recall	train F1	test accuracy	test precision	test recall	test F1
le5	99.00%	99.58%	98.43%	98.97%	90.74%	91.31%	90.06%	90.67%
le7	99.20%	99.58%	98.72%	99.12%	94.29%	94.61%	93.94%	94.27%
le9	**99.80%**	**99.67%**	**100.00%**	**99.83%**	**96.56%**	**96.69%**	**96.43%**	**96.56%**
res18	97.91%	96.85%	99.05%	97.90%	95.11%	95.40%	94.83%	95.10%
res152	92.60%	93.99%	91.09%	92.21%	94.09%	94.61%	93.51%	94.05%

Deep Learning | 1st vs. 2nd Gen Aida





* Burney database is not balanced, more snippets without poetic content

Deep Learning | 1st vs. 2nd Gen Aida

1st Gen AIDA
Chronicling America Database

		Ground-Truth: Poem	Ground-Truth: Not Poem
Predicted	Poem	602 (35.54%)	124 (7.32%)
	Not Poem	245 (14.46%)	723 (42.68%)
Correctly predicted poem snippets: 71.07% and not poem snippets: 85.36%			

1st Gen AIDA
Burney Collection Database

		Ground-Truth: Poem	Ground-Truth: Not Poem
Predicted	Poem	273 (10.02%)	420 (15.41%)
	Not Poem	230 (8.44%)	1802 (66.13%)
Correctly predicted poem snippets: 54.27% and not poem snippets: 81.10%			

2nd Gen AIDA
Chronicling America Database

		Ground-Truth: Poem	Ground-Truth: Not Poem
Predicted	Poem	822 (48.52%)	22 (1.30%)
	Not Poem	25 (1.48%)	825 (48.70%)
Correctly predicted poem snippets: 97.05% and not poem snippets: 97.40%			

2nd Gen AIDA
Burney Collection Database

		Ground-Truth: Poem	Ground-Truth: Not Poem
Predicted	Poem	304 (11.16%)	68 (2.50%)
	Not Poem	199 (7.30%)	2154 (79.05%)
Correctly predicted poem snippets: 60.44% and not poem snippets: 96.94%			

Deep Learning |2nd Gen Aida

2^{nd} Gen AIDA improved poetic content classification for historical newspaper by more than 10% comparing to 1^{st} gen AIDA

- 2^{nd} Gen AIDA has over 90% test accuracies on both Chronicling America and Burney database, while 1^{st} Gen AIDA cannot reach 80%.

2^{nd} Gen AIDA have potentials to generate a general classifier for other databases than the training database

- 2^{nd} Gen AIDA has over 90% test accuracy on Burney database.
- Precision and recall of 2^{nd} Gen AIDA are lower than 90% but much higher than 1^{st} Gen AIDA

Part 5: Library of Congress
Project 1. Document Segmentation

Objectives | Find and localize *Figure/Illustration/Cartoon* presented in an image

Applications | metadata generation, discover-/search-ability, visualization, etc.

Background |State-of-the-Art CNN models

- **Convolutional Neural Network** (CNN) Models (deep learning)
  - Classification [Dataset; Top-1 / Top-5]
    - 2014, VGG-16 (Classification) [ImageNet; 74.4% / 91.9%]
    - 2015, ResNet-50 (Classification) [ImageNet; 77.2% / 93.3%]
    - 2018, ResNeXt-101 (Classification) [ImageNet; 85.1% / 97.5%]
  - Segmentation [Dataset; Intersection-over-Union (IoU)]
    - 2015, U-net (Segmentation/Pixel-wise classification) [ISBI; 92.0%]

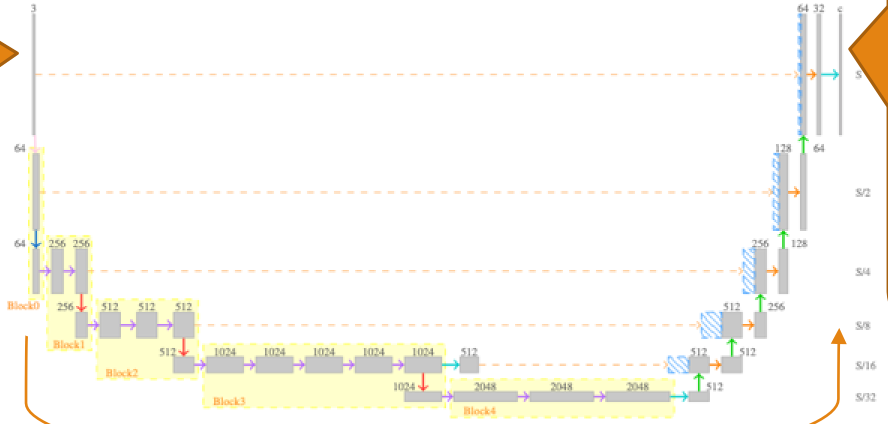
- So, we now know that CNNs achieve *remarkable* performances in both classification and segmentation tasks.
- ***What about document images then?***

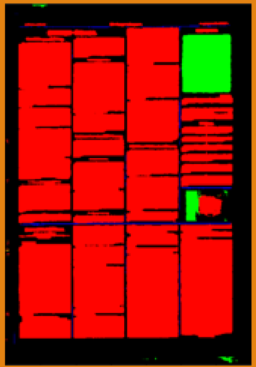
Aida

Document Segmentation | **Technical Details**

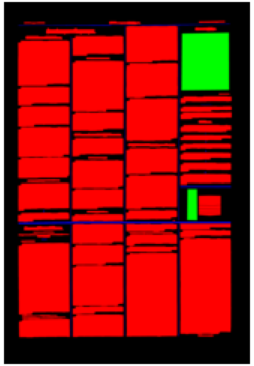
- ***Training*** is a process of finding the optimal value weights between artificial neurons that minimizes a pre-defined ***loss*** function

Input





Prediction



Ground-truth

1. Convolution & Down-sampling:
understand "***WHAT***" is present in the image (i.e., feature extraction)

2. Up-sampling:
understand "***WHERE***" it is present in the image

3. Calculate per-pixel loss
4. Update weights between neurons
5. Repeat the process

Aida

Document Segmentation | Dataset

Beyond Words

- Total of 2,635 image snippets from 1,562 pages (as of 7/24/2019)
  - 1,027 pages with single snippet
  - 512 pages with multiple snippets
- Issues
  - Inconsistency (Figure 1)
  - Imprecision (Figure 2)
  - Data imbalance (Figure 3)

Figure 1. Example of inconsistency. Note that there are more than one image snippets in the left image (i.e. input) while there is only a single annotation in the right ground-truth.

Figure 2. Example of imprecision. From left to right: (1) ground-truth (yellow: Photograph and black: background) and (2) original image. Note here that in the ground-truth, non-photograph-like (e.g., texts) components are included within the yellow rectangle region.

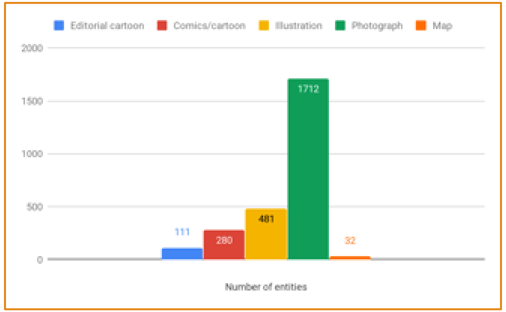


Figure 3. Number of snippets in Beyond Words. Note here the data imbalance

Document Segmentation | Dataset

European Historical Newspapers (ENP)

- Total of 57,339 image snippets in 500 pages
  - All pages have multiple snippets
- Issues
  - Data imbalance
    - Text: 43,780
    - Figure: 1,452
    - Line-separator: 11,896
    - Table: 221

Figure 4. Example of image (left) and ground-truth (right) from ENP dataset. In the ground-truth, each color represents the following components: (1) black: background, (2) red: text, (3) green: figure, (4) blue: line-separator, and (5) yellow: table.

Document Segmentation | Experimental Results

- A U-net model trained with ENP dataset shows better segmentation performance than that with Beyond Words in terms of pixelwise-accuracy and IoU score
  - IoU score is a commonly used metric to evaluate segmentation performance
  - The three issues—inconsistency, imprecision, and data imbalance—of Beyond Words dataset need to be improved for better use in training

Model	train/eval size	Classes	Weighted training	Pre-processing (Normalization)	Best Score	
					Accuracy	mIoU
BW_1500_v1	1226/306	0: Background 1: Editorial cartoon 2: Comics/cartoon 3: Illustration 4: Photograph 5: Map	No	No	0.87	0.24
BW_1500_v2			Yes [10;22;20;18;8;22]		0.88	0.26
ENP_500_v1	385/96	0: Background 1: Text 2: Figure 3: Separator 4: Table	Yes [5;10;40;10;35]	No	0.88	0.64
ENP_500_v2				Yes	0.89	0.64
ENP_500_v3			**No**	**No**	**0.91**	**0.69**
ENP_500_v4				**Yes**	**0.91**	**0.69**

*Accuracy: Pixel-wise accuracy.
*mIoU: Average intersection over union.
*Normalization: Zero mean unit variance

- Assigning different weights per class to mitigate data imbalance did *not* show performance improvement
  - ***Future Work:*** Explore a different way of weighting strategy to mitigate a data imbalance problem

Document Segmentation | Potential Applications 1

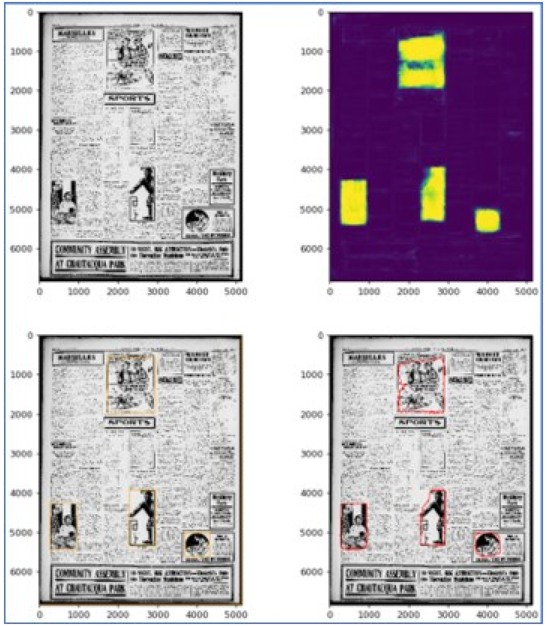


Figure 5. Segmentation result of ENP_500_v4 on Chronicling America image (sn92053240-19190805.jpg). Clockwise from top- left: (1) Input, (2) probability map for figure class, (3) detected figures in polygon, and (4) detected figures in bounding-box. In the probability map, pixels with higher probability to belong to figure class are shown with brighter color.

- ❑ Enrich page-level metadata by cataloging the types of visual components presented on a page
- ❑ Enrich collection-level metadata as well
- ❑ Visualize figures' locations on a page

Document Segmentation | Potential Applications 2

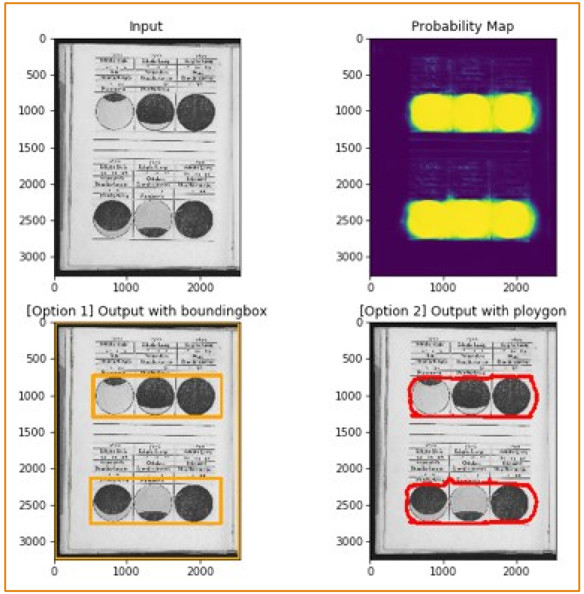


Figure 6. Successful segmentation result of ENP_500_v4 on book/printed material (https://www.loc.gov/resource/rbc0001.2013rosen0051/?sp=37).

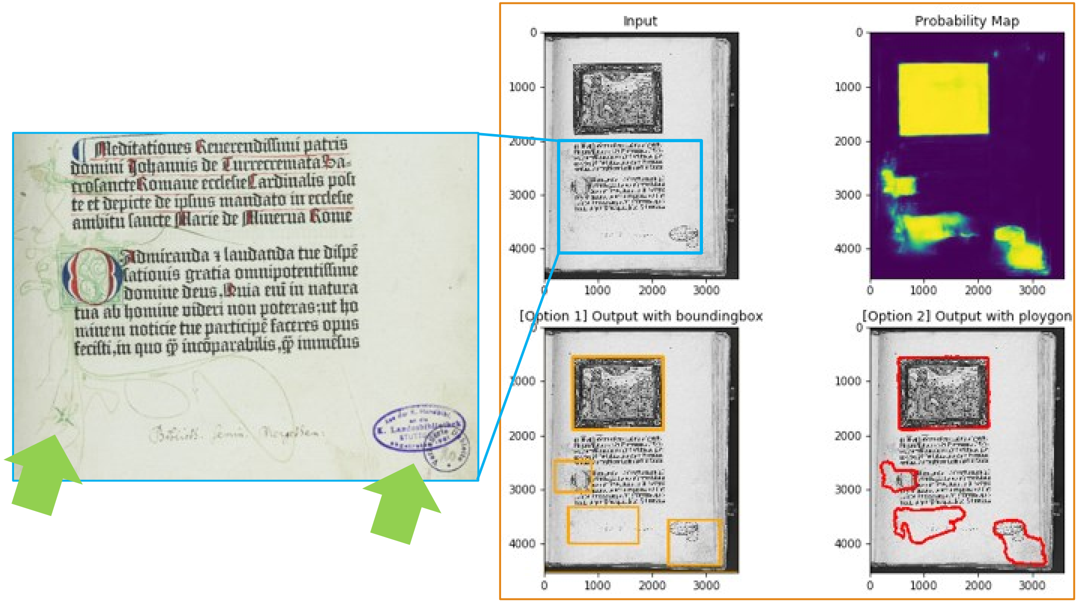


Figure 7. Failure segmentation result of ENP_500_v4 on book/printed material (https://cdn.loc.gov/service/rbc/rbc0001/2010/2010rosen0073/0005v.jpg). Note that there is light drawing or stamps (marked in green arrows) on the false positive regions.

Document Segmentation | Conclusions

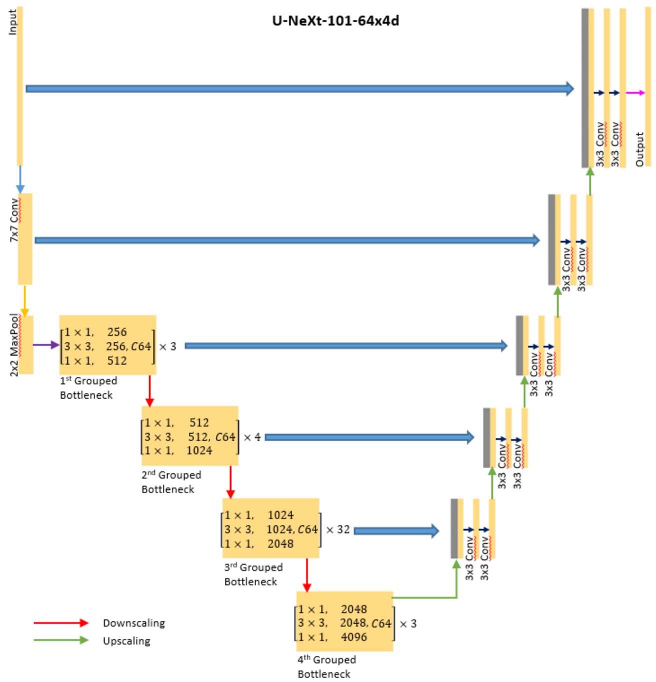
- As a preliminary experiment, a state-of-the-art CNN model (i.e., U-net) shows promising segmentation performance on ENP document image dataset,
  - There is still room for improvement with more sophisticated training strategies (e.g., weighted training, augmentation, etc.)
- To make Beyond Words dataset more as a valuable training resource for machine learning researchers, we need to address the following issues:
  - Consistency
  - Precision of the coordinates of regions

Part 5: Library of Congress
Project 2.1. Figure/Graph Extraction

Objectives | Find and localize *Figure*/Graph in a document image

Applications | Graph retrieval, document segmentation based on content type

Figure/Graph Extraction | Technical Details



An FCN (U-NeXt) is used

- U-NeXt combines ResNeXt and U-Net
  - ResNeXt101_64x4d
- Why ResNeXt101_64x4d?
  - Current state-of-art
  - Accessible pre-trained model
- **Transfer learning**
  - ResNeXt101_64x4d
  - Number of parameters:
    - 114.4 million 32.8 million

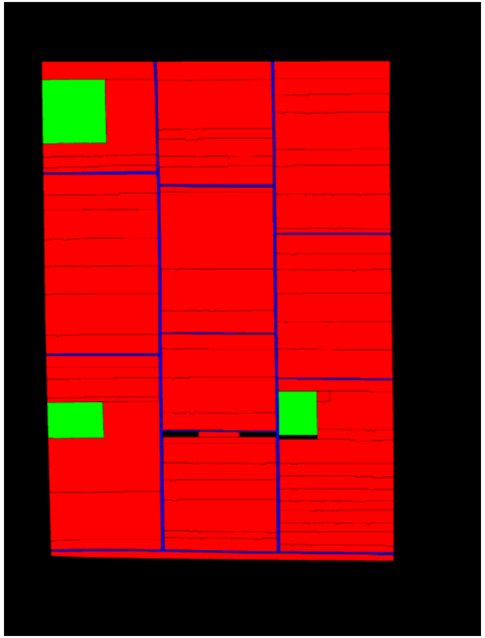
Figure/Graph Extraction | Datasets

- **ENP collection**: European newspaper collection
  - A subset used for the International Conference on Document Analysis and Recognition competition
- **Beyond Word collection**: Transcribed collection
  - But cannot be used for training directly …
    - Problem 1: missing figures in ground-truth
    - Problem 2: inaccurate ground-truth

Aida

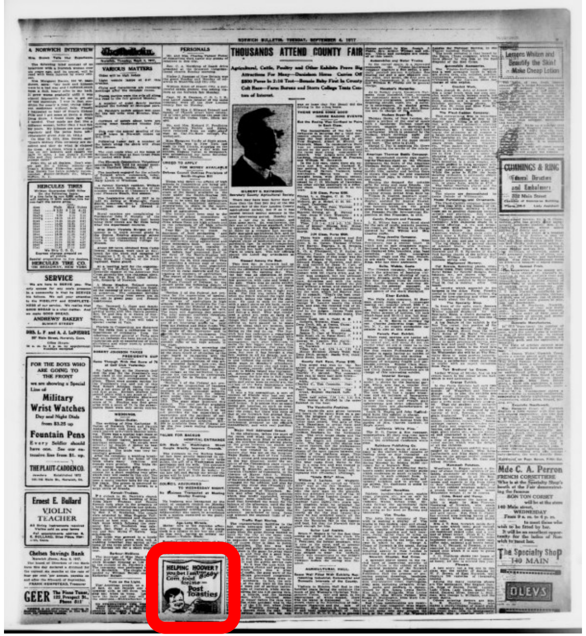
Figure/Graph Extraction | Datasets: ENP

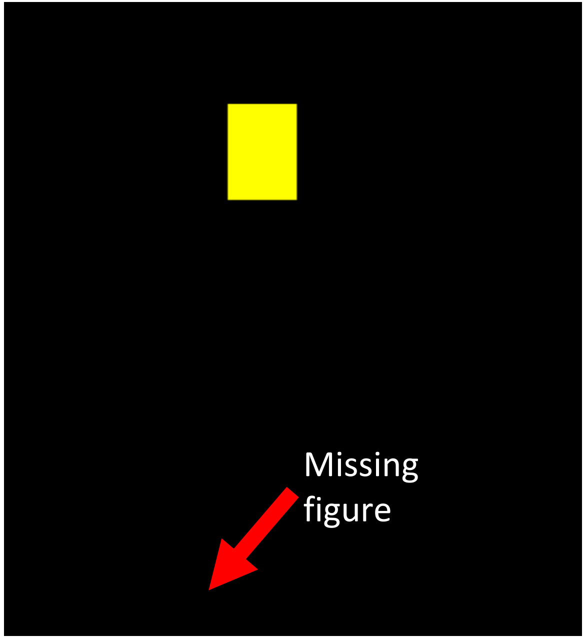
Document Image



Ground-truth

Figure/Graph Extraction | Datasets: Beyond Words





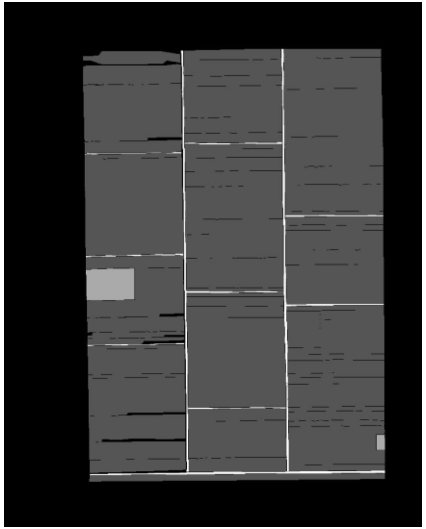
Document Image

Ground-truth

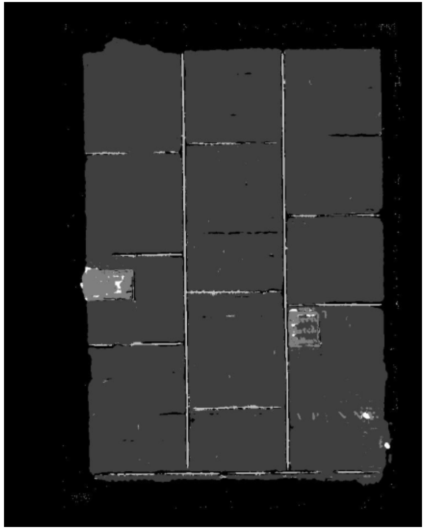
Figure/Graph Extraction | Preliminary Results

- ❑ Transfer parameters from pre-trained ResNeXt101 64x4d
- ❑ Trained on ENP dataset

Document Image



Ground truth

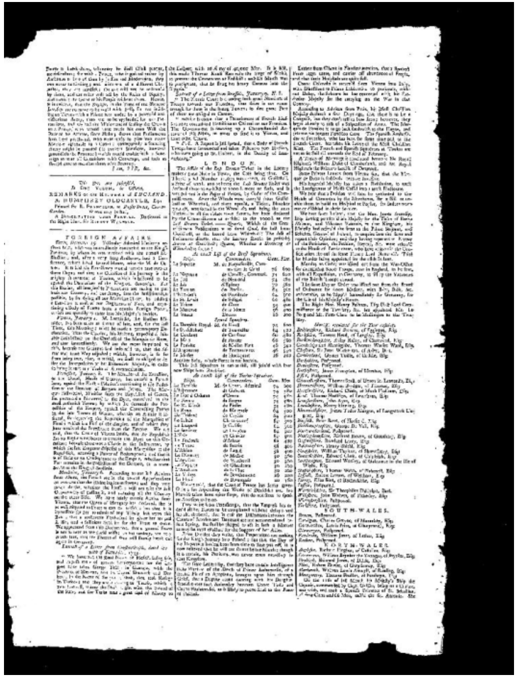


Predictio n

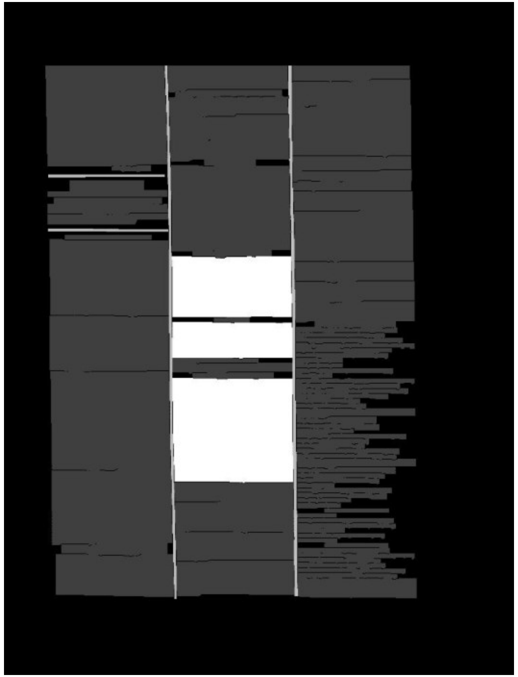
Figure/Graph Extraction | Conclusions

- Promising preliminary results
- Potential applications
  - Segmentation based on content type to increase item-level accessibility
  - Retrieval of figures/graphs for further study
- Challenges
  - U-NeXt still needs more iterations of training
  - Preliminary training indicates that tables may be the hardest type to extract

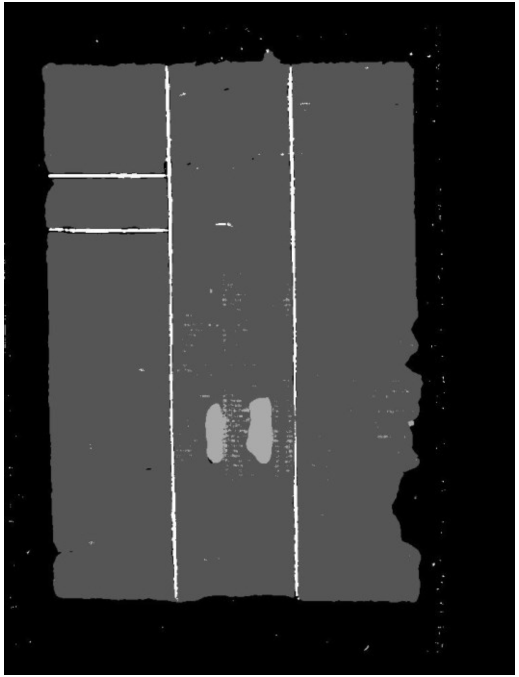
Figure/Graph Extraction | Challenge



Document Image



Ground truth



Predictio n

Aida

Part 5: Library of Congress
Project 2.1. Text Extraction from Figure/Graph

Objectives | Extract texts from figure/graph

Applications | Metadata generation, OCR for figure/graph caption

Aida

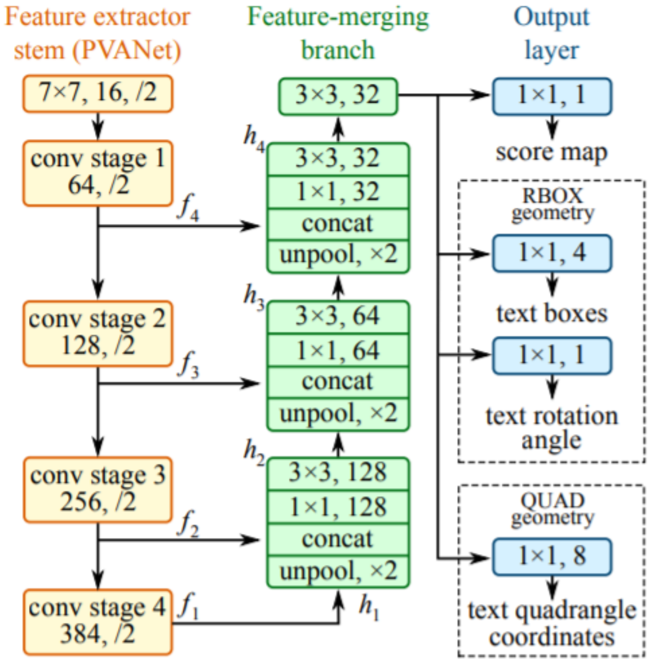
Text Extraction from Figure/Graph | Technical Details

EAST text detector

- EAST: Efficient and Accurate Scene Text detector
- HyperNet + U-Net
- Detect texts in graphic images in any direction

Why applicable?

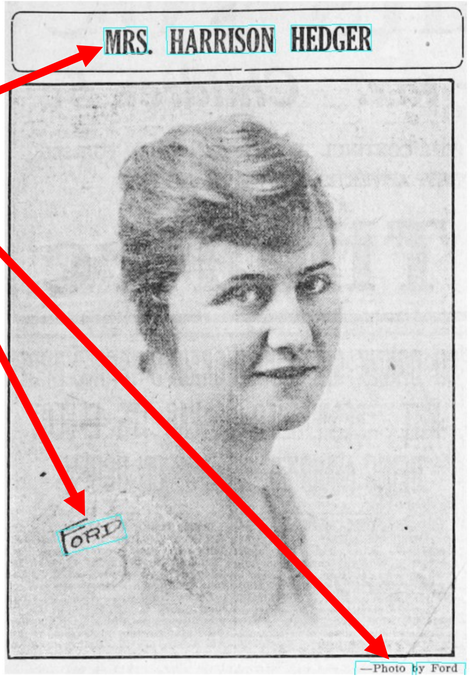
- figures/illustrations are snippets of a graphic region



Aida

Text Extraction from Figure/Graph | Preliminary Results

Detected
Texts



- ❑ Performance on detecting texts in newspaper figure/graph is good
- ❑ Texts location is recorded

Text Lines

- 6 text lines
- { "x0": 62, "y0": 608, "x1": 135, "y1": 588, "x2": 143
- { "x0": 188, "y0": 33, "x1": 312, "y1": 31, "x2": 313,
- { "x0": 331, "y0": 31, "x1": 423, "y1": 30, "x2": 423,
- { "x0": 116, "y0": 34, "x1": 166, "y1": 33, "x2": 166,
- { "x0": 405, "y0": 755, "x1": 470, "y1": 757, "x2": 47
- { "x0": 475, "y0": 756, "x1": 531, "y1": 757, "x2": 53

Text Extraction from Figure/Graph | Conclusions

- Promising preliminary results
- Potential application
  - Perform OCR on detected text regions for higher accuracy
  - Extract OCR-ed words in detected text regions as metadata

Aida

Part 5: Library of Congress
Project 3. Document Type Classification

Objectives | (1) Classify a given image into one of *Handwritten/Typed/Mixed* type; (2) Classify a given image into one of *Scanned/Microfilmed*

Applications | metadata generation, discover-/search-ability, cataloging, etc.

Document Type Classification | Technical Details

*Note that we do not need up-sampling in this task, since **WHERE** is not our concern*

- A simple VGG-16 is used (Figure 8)
  - Afzal et al. reported that most of state-of-the-art CNN models yielded around 89% of accuracy on document image classification task
- **Transfer learning?**
  - Why don't we initialize our model's weights from a model that has been already trained on a large-scale data, such as *ImageNet* (about 14M images)?
  - ***Why?*** (1) training a model from the scratch (i.e., the value of weights between neurons are initialized to random number) takes too much time; (2) we have too small a dataset to train a model

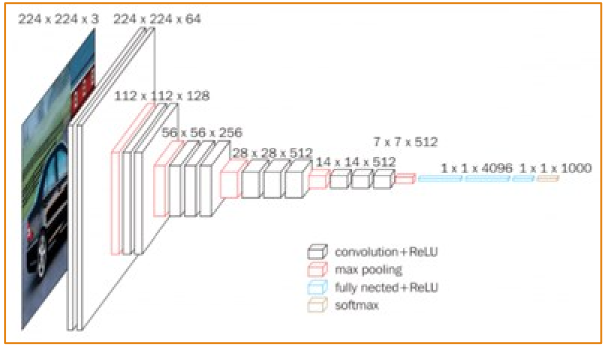


Figure 8. Architecture of original VGG-16. In our project, the last softmax layer is adjusted to have a shape of 3, which is the number of our target classes; handwritten, typed, and mixed

Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

Document Type Classification | Datasets

- We have two datasets:
  - Experiment 1: *RVL-CDIP* (400,000 document images with 16 different balanced classes); publicly available
  - Experiment 2: *suffrage_1002* (1,002 document images with 3 different balanced classes); manually compiled from ***By the People: Suffrage*** campaign (Table 1)

	handwritten	**typed**	**mixed**	Total
train	267	267	267	801
validation	33	33	33	99
test	33	33	33	99
Total	333	333	333	999

Table 1. Configuration of *suffrage_1002* dataset.

Figure 9. Example document images from each 16 different classes

Aida

Document Type Classification | Datasets

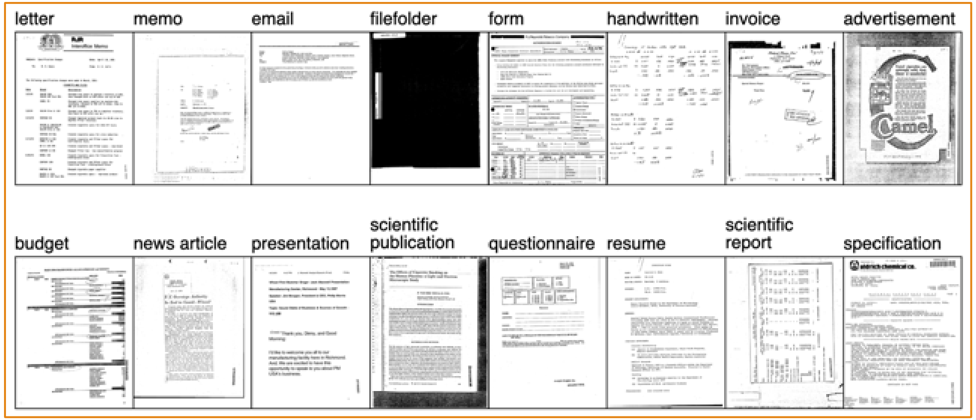


Figure 9. Example document images from each 16 different classes in *RVL_CDIP* dataset

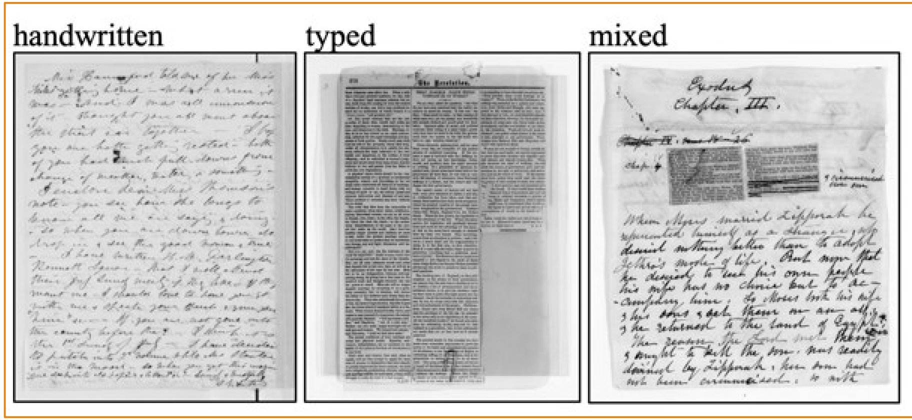


Figure 10. Example document images from each 3 different classes in *suffrage_1002* dataset

Document Type Classification | Experimental Results

Table 1. Precision, recall, and f1-score of *VGG-16* trained on *RVL_CDIP* dataset. The alphabetic labels are corresponding to the following labels: *letter*, *form*, *email*, ***handwritten***, *advertisement*, *scientific report*, *scientific publication*, *specification*, *file folder*, *news article*, *budget*, *invoice*, *presentation*, *questionnaire*, *resume*, and *memo*. Our class of interest, ***handwritten***, is bolded.

(unit: %)	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	88	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

Table 2. Precision, recall, and f1-score of *VGG-16* on *suffrage_1002* testing set.

(unit: %)	**handwritten**	**typed**	**mixed**	**Avg**
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

- Experiment 1: We obtained a model trained on a large-scale document image dataset, *RVL-CDIP* with promising classification performance, as shown in Table 1
  - ***Implication***: Features learned from natural images (ImageNet) are general enough to apply to document images
  - Now we can utilize this model by retraining it with our own *suffrage_1002* dataset in Experiment 2
- Experiment 2: The retrained model shows even better classification performance, as shown in Table 2

Aida

Document Type Classification | **Conclusions**

- In both experiments, the state-of-the-art CNN model is capable of classifying document images with promising performance
  - ***Potential Applications***: help tagging an image type
- A main *challenge*: classifying a mixed type document image, as shown in Figure 11
  - ***Future Work:*** Perform a confidence level analysis to mitigate this problem
- ***Future Work:*** We expect that the classification performance can be further improved with a larger large-scale dataset

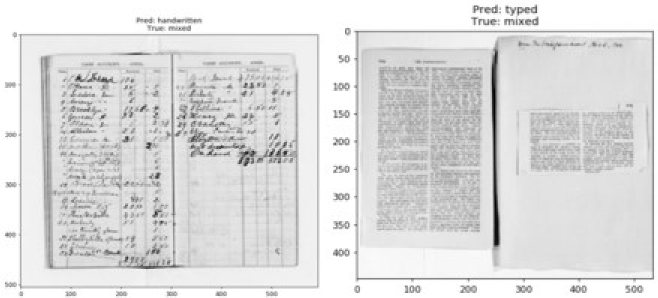


Figure 11. Failure prediction cases. On the left example, a typed region is relatively smaller than that of handwriting. On the right example, a handwriting region is relatively smaller than that of typing.

Afzal, M. Z., Kölsch, A., Ahmed, S., & Liwicki, M. (2017, November). Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*(Vol. 1, pp. 883-888). IEEE.

Part 5: Library of Congress Project 4. Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People

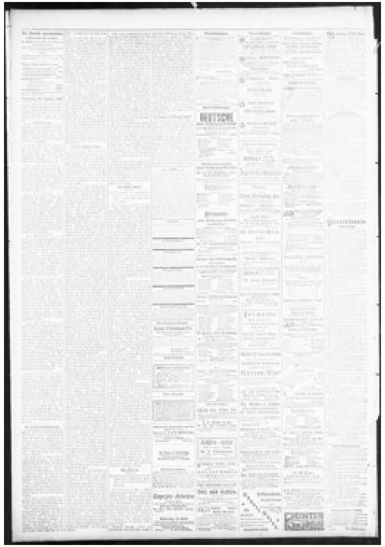
Applications | Providing quality scores for machine reading on four criteria: (1) *skewness*, (2) *contrast*, (3) *range-effect*, and (4) *bleed-through*

Aida

Quality Assessment | Technical Details

- Objective quality assessment on four criteria
  - *Skewness, Contrast, Range-effect, Bleed-through*
  - *Based on the DIQA programs developed at Aida @ UNL (previously tested using Chronicling America's repository of archived newspaper pages*
  - *Not* directly machine learning related
- **Why**?
  - Help identify images that need pre-processing
  - Reduce unnecessary workload for pre-processing images
  - Indicate general qualities of the dataset

Objective Quality Assessment | Examples



Contrast

THE BRYAN DAILY EAGLE

ED HALL

THE GROCER

At Jno. B. Mike's

ALL KINDS OF Country Produce

At Jno. B. Mike's

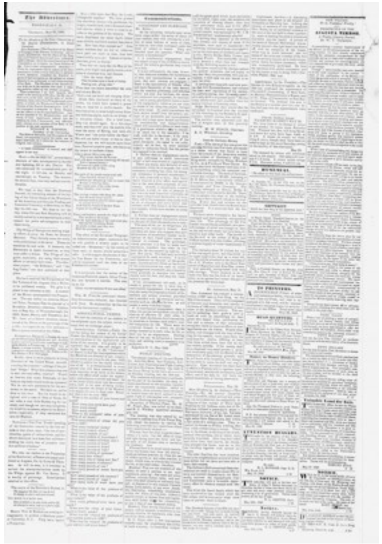
Range-effect

AVISOS.

CATECISMO

Sistema-métrico decimal

Bleed-through

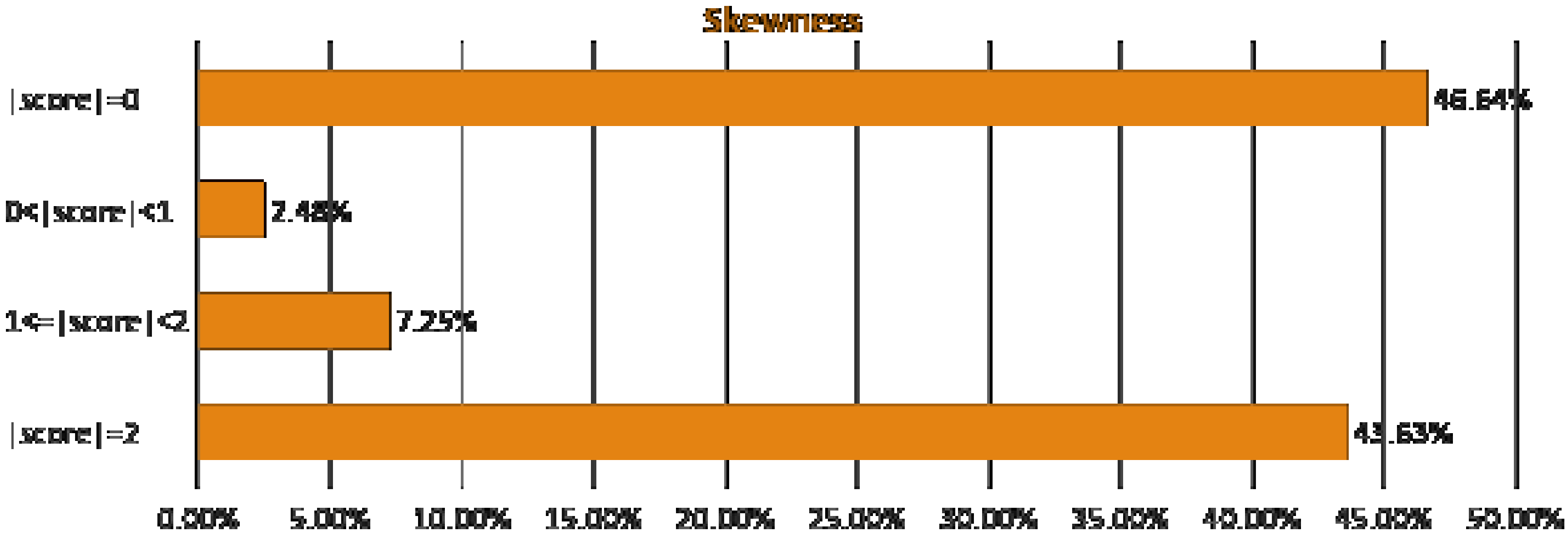


Skewness

Quality Assessment | Datasets

- The Civil War collection within By the People:
  - 36003 images were downloaded
  - 35990 images passed the DIQA program
    - *13 images failed as they barely had texts (see examples later)*

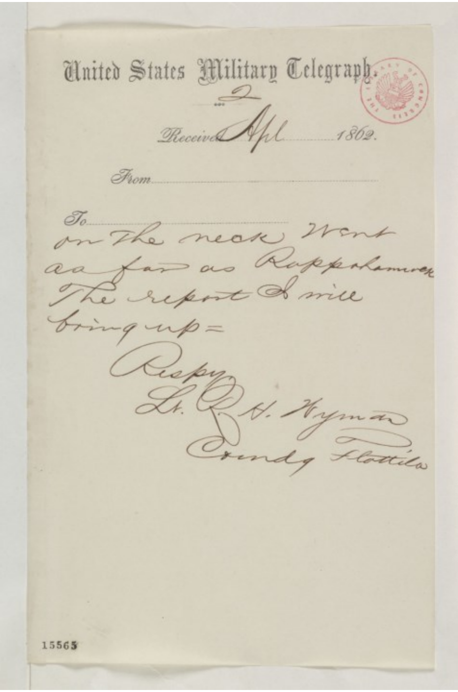
Quality Assessment | Experimental Results



Quality Assessment | Observations

- There were 46% images had the perfect score (zero) on skewness assessment
- But, there were also 43% images had the largest score (two)
- This suggest the skewness of the dataset may be divided

- However, a large portion of the dataset was hand-written
  - The skewness evaluation was depending on vertical aligned text line ends
  - Hand-written lines that were unjustified on left/right margin may result in a faulty score


United States Military Telegraph.

2

Received Apl 1862.

From

To

on the neck went as far as Rappahannock The report I will bring up =

Respy

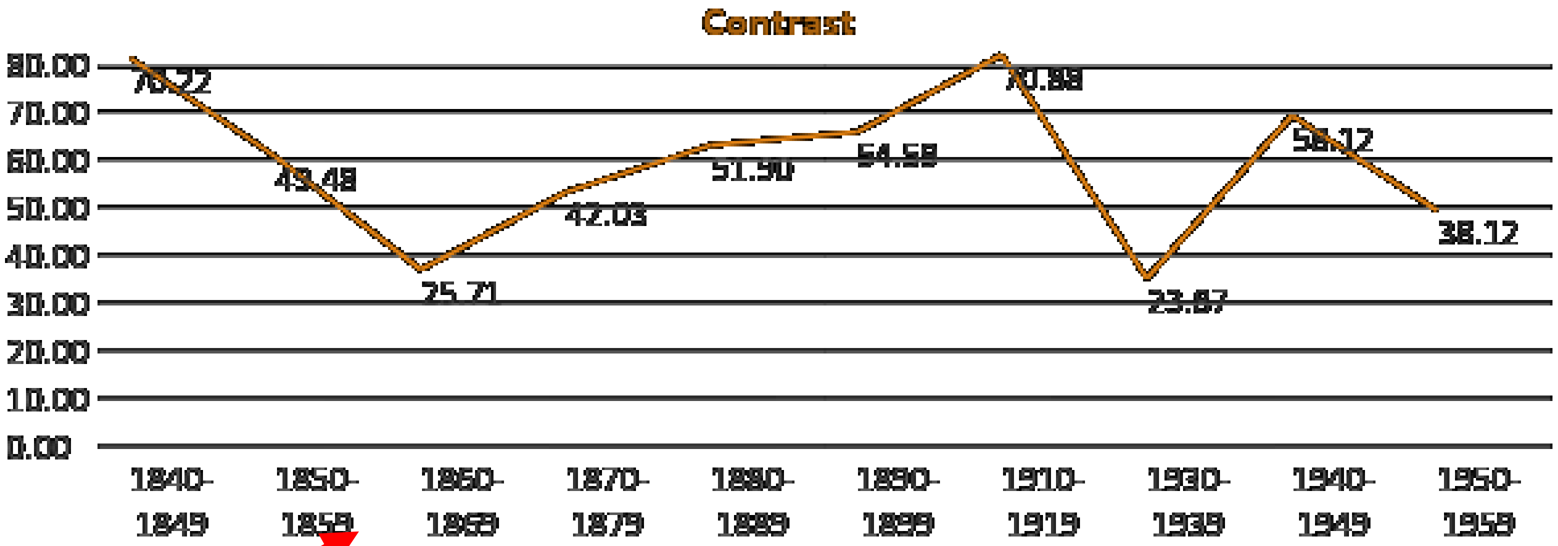
Lt. R. H. Hyman

Comdg Flotilla

15565

Aida

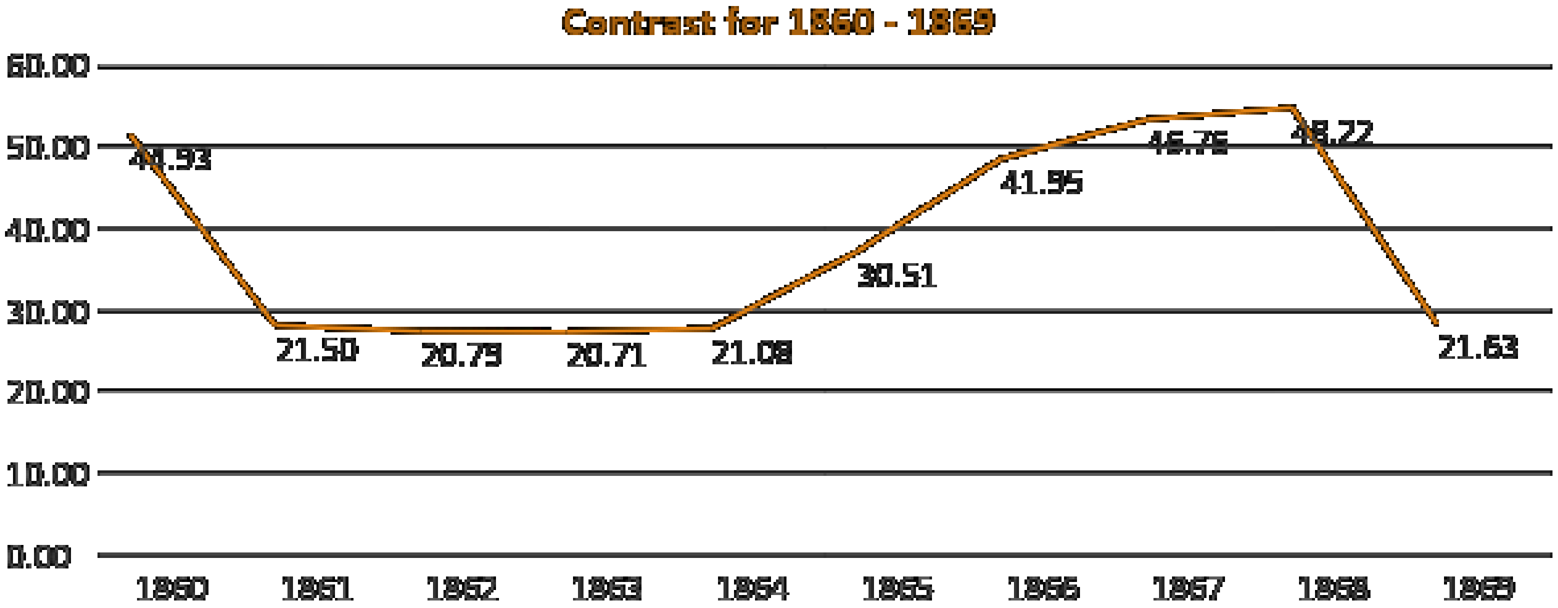
Quality Assessment | Experimental Results



~90% of images in the dataset falls within this range

Aida

Quality Assessment | Experimental Results



Quality Assessment | Observations

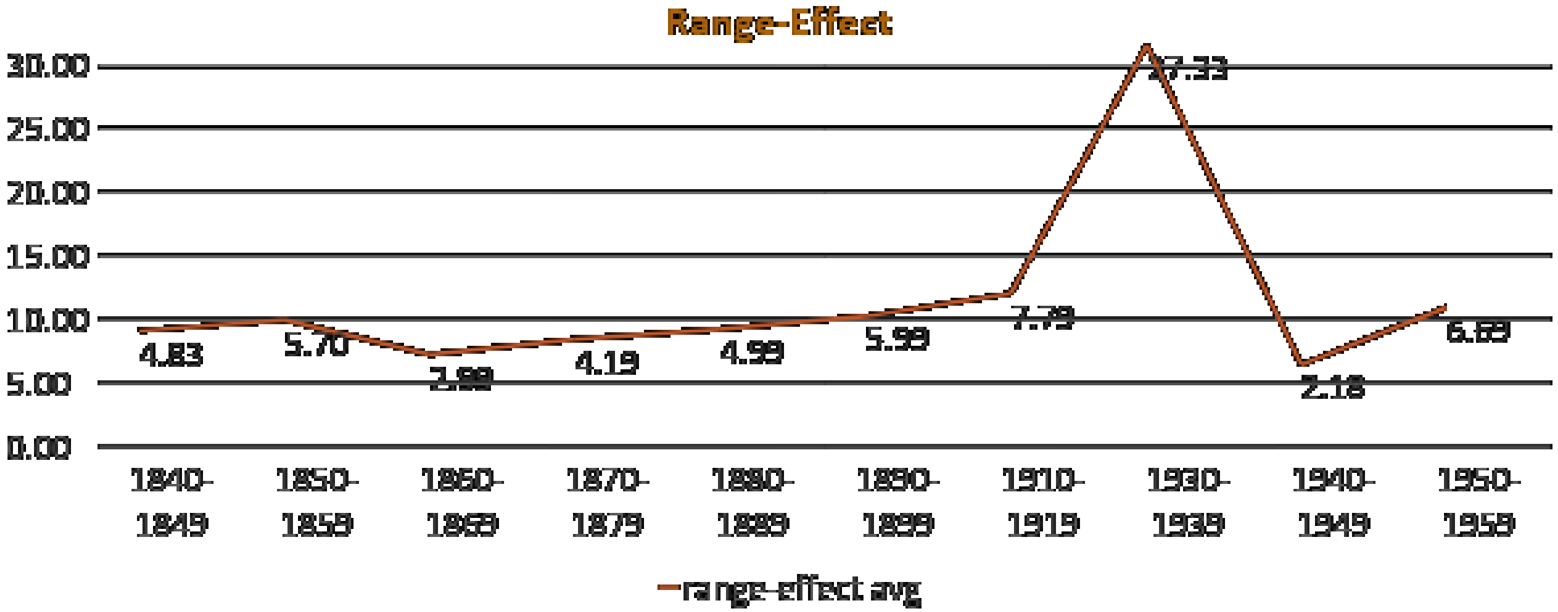
- Based on previous work of Aida, contrast score less than 40 may cause troubles for reading

- The first chart shows the average contrast was good
- But ~90% images fall in year range from 1860 to 1869

- The second chart break the year range to year-wise analysis
- Images from 1961 to 1964 seem to have contrast issues

Aida

Quality Assessment | Experimental Results



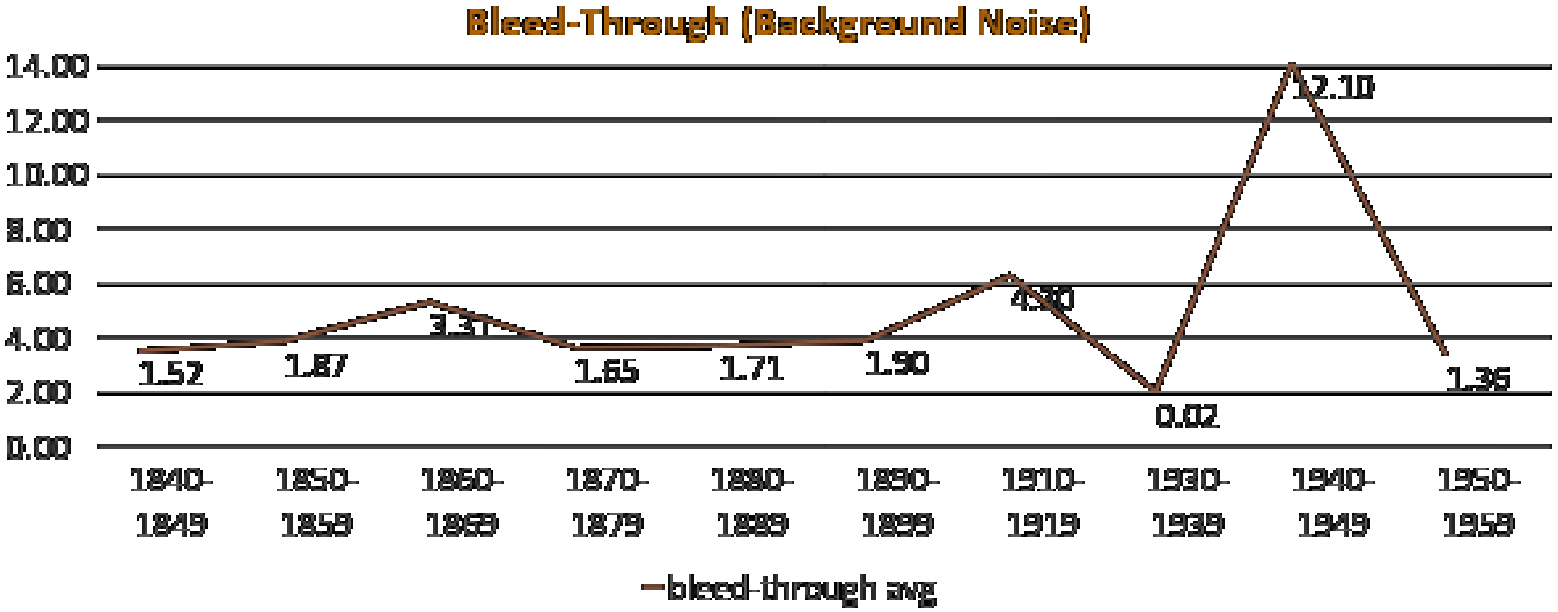
Aida

Quality Assessment | Observations

- Based on DIQA on Chronicling America, range-effect score that is smaller than 3 is good
- Statistic data indicates the database averagely has quality issues on range effect

Aida

Quality Assessment | Experimental Results

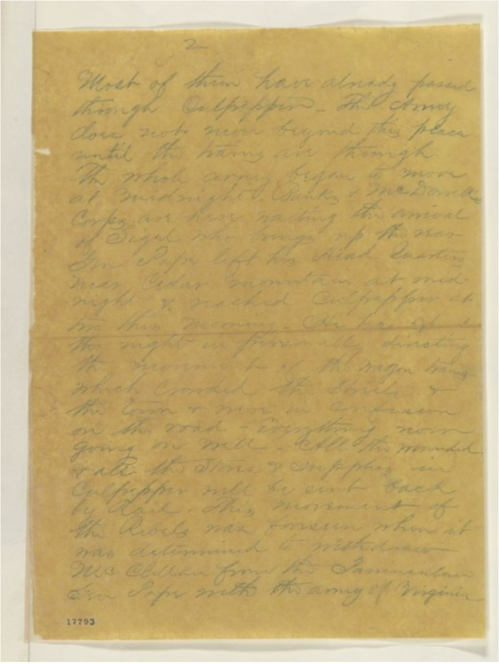


Quality Assessment | Observations

- Unfortunately, there is no magic number to say which score is good
- But rather than 76 images from 1940 to 1949, other images has relatively lower score (better quality) on background noise

Quality Assessment | Potential Issues

- ❑ Numerous images with yellowish background and faded inks
- ❑ They are hard to read even to human eye
  - ❑ Contrast could be lowered
  - ❑ Skewness could be almost impossible to compute

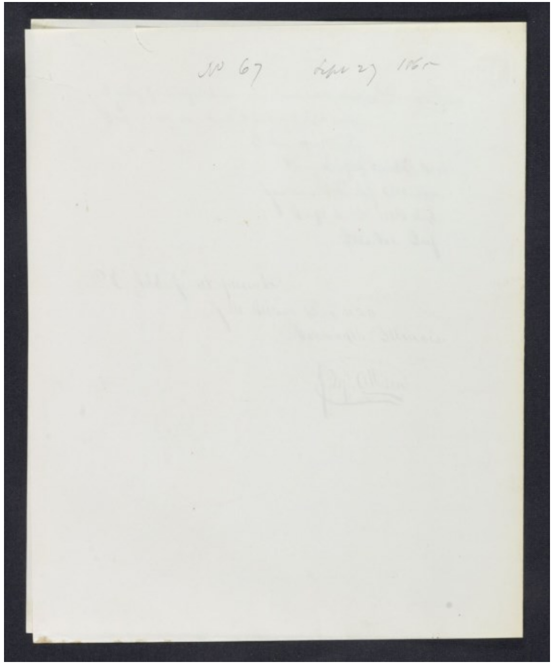

2

Most of them have already passed
through Culpepper - The Army
does not move beyond this place
until the trains are through
The whole army began to move
at midnight. Banks & McDowell
Corps are here waiting the arrival
of Sigel who brings up the rear
Gen Pope left his Head Quarters
near Cedar mountain at mid-
night & reached Culpepper at
[illegible] this morning - He has spent
the night in personally directing
the movement of the wagon trains
which crowded the streets &
the town & were in confusion
on the road - Everything now
going on well - All the wounded
& all the stores & supplies in
Culpepper will be sent back
by Rail - This movement of
the Rebels was foreseen when it
was determined to withdraw
McClellan from the Peninsula
Gen Pope with the army of Virginia

17793

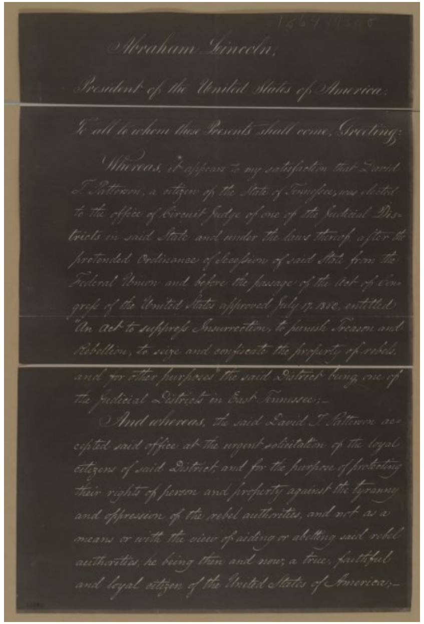
Quality Assessment | Potential Issues

- Numerous images are covers or labels of a series
- These images are largely blank
  - Contrast is poor
  - Histogram equalization might be able to enhance the quality



Quality Assessment | Potential Issues

- There are color-inverted images from microfilm
  - Renders bleed-through assessment useless



Abraham Lincoln,

President of the United States of America:

To all to whom these Presents shall come, Greeting:

Whereas, it appears to my satisfaction that David T. Patterson, a citizen of the State of Tennessee, was elected to the office of Circuit Judge of one of the Judicial Districts in said State and under the laws thereof, after the pretended Ordinance of Secession of said State from the Federal Union and before the passage of the Act of Congress of the United States approved July 17, 1862, entitled "An Act to suppress Insurrection, to punish Treason and Rebellion, to seize and confiscate the property of rebels, and for other purposes" the said District being one of the Judicial Districts in East Tennessee;—

And whereas, the said David T. Patterson accepted said office at the urgent solicitation of the loyal citizens of said District, and for the purpose of protecting their rights of person and property against the tyranny and oppression of the rebel authorities, and not as a means or with the view of aiding or abetting said rebel authorities, he being then and now, a true, faithful and loyal citizen of the United States of America;—

Part 5: Library of Congress
Project 5. Digitization Type Differentiation: Microfilm or Scanned

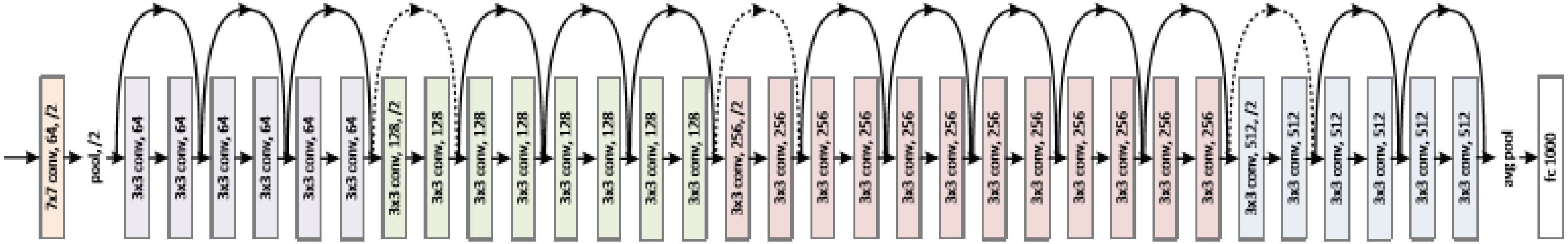
Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*

Applications | Metadata generation, pre-processing policy selection

Aida

Digitization Type Differentiation | Technical Details

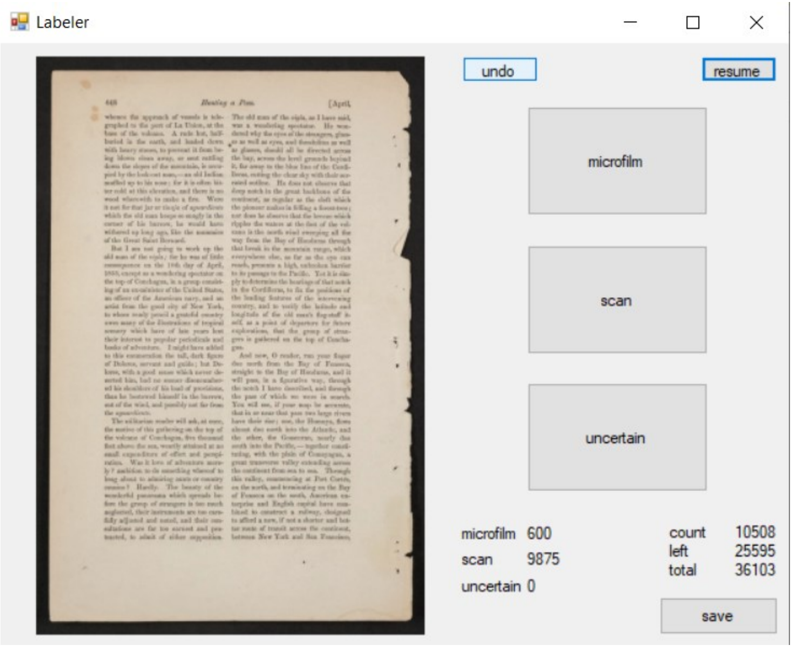
- Pre-trained ResNeXt is adopted
- Attached output layers are two dense layers with a 1D output vector
- **The pre-trained ResNeXt can classify images to 1000 different categories**
- The pre-trained ResNeXt is a good feature extractor
  - Number of parameters: 94.1 million ⍰ 12.6 million



Digitization Type Differentiation | Datasets

- Created from the Civil War collection within By the People
- A manually created database by ***randomly*** choosing 600 images on scanned materials and 600 images on microfilm materials
- The randomization was performed by shuffling the entire list of 36,003 images in the collection
- The randomization ensured that images in the collection have a fair chance to be chosen
- The randomization seed was fixed to ensure the experiments can be reproduced

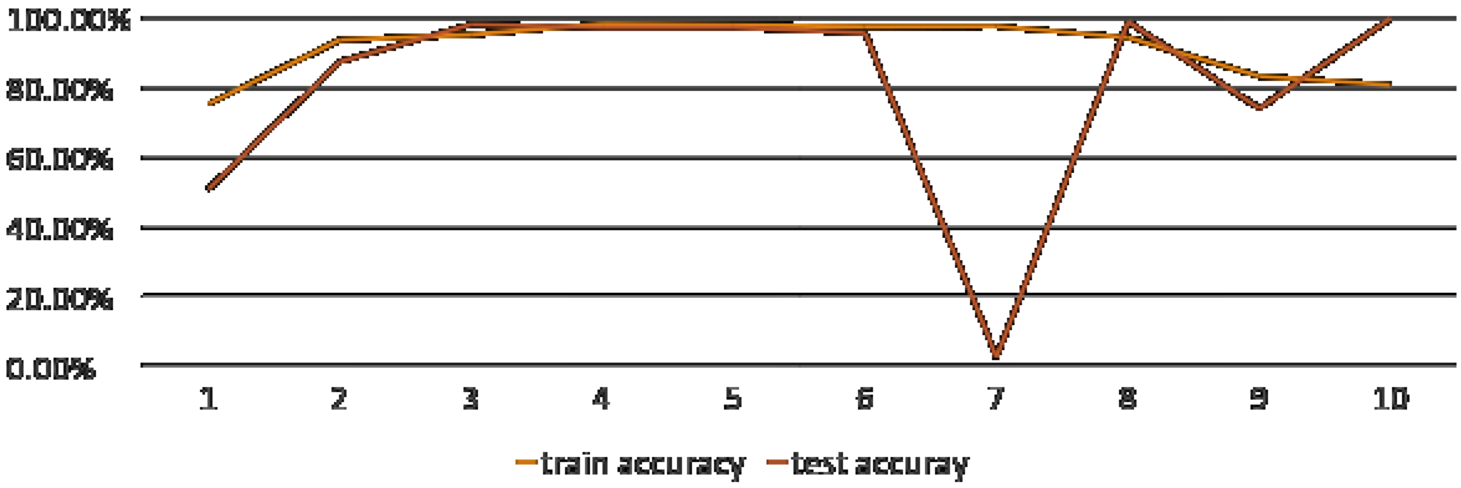
Digitization Type Differentiation | Datasets



- **Rough estimate**: Based on 10,508 images that was processed, ***ratio of images from microfilm to scanned materials is about 1:16***

Digitization Type Differentiation | Experimental Results

- With pre-trained ResNeXt,
  - It only took **one** iteration to reach more than 90% accuracy on training set, and
  - It only took **two** iterations to reach more than 90% accuracy on testing set



Aida

Digitization Type Differentiation | Experimental Results

- The best test iteration result was able to 100% correctly classify all images

		Ground Truth	
		Scanned	Microfilm
Prediction	Scanned	60	0
	Microfilm	0	60

Aida

Digitization Type Differentiation | Conclusions

- Existing pre-trained model can be easily extended to more designated tasks
- The extended model only need a small set of labeled data to reach near-perfect performance in this task
- Automated digitization type differentiation is ***readily*** achievable.

Aida

Questions?