Progress report

Yi Liu

Current Progress

Document Image Quality Assessment

In the last progress report (Document image quality assessment for digital library collections), we proposed to perform document image quality assessment (DIQA) to By the People dataset. Hence, in this report, we downloaded 36,003 images from the civil war collection (the dataset) of By the People. And we analyzed the outcome of the assessment results.

The DIQA algorithms used in this experiment were developed as part of the project Aida to assess qualities of newspaper page images from 1834 to 1922, which included four criteria: (1) skewness, (2) contrast, (3) range-effect, and, (4) bleed-through (background noise). We found that, for newspaper page images, a contrast score higher than 40 could be considered as having good contrast quality. And a range-effect score lower than three could be considered as having no or fewer range-effect issues. However, there was no clear indicator for skewness and bleed-through assessment. All we could say was that the lower the score on skewness or bleed-through, the better the quality.

In this statistical analysis, there were 35,990 out of 36003 images that successfully passed the quality assessment program. 13 images failed due to exceptions of the program caused by incorrect assumptions. We will later dig into the program to find the detailed reasons causing these exceptions.

Skewness. For skewness evaluation shown in Figure 1, there are 43.63% (15,703 out of 35990) images in the dataset with the maximum skewness score (i.e., score of 2). Hence, there are 43.63% images that are significantly skewed. There are 7.25% images that are lightly skewed (i.e., skewness ~1-2) in the dataset. Further, 2.48% of the images are trivially skewed (i.e., skewness < 1) in the dataset. And there are 43.63% images that are not skewed at all. Note that the larger the absolute value of the score, the more skewed the document image. And a positive or negative score indicated the skewness orientation. In Figure 1, "|score|" means the absolute value of the skewness score.

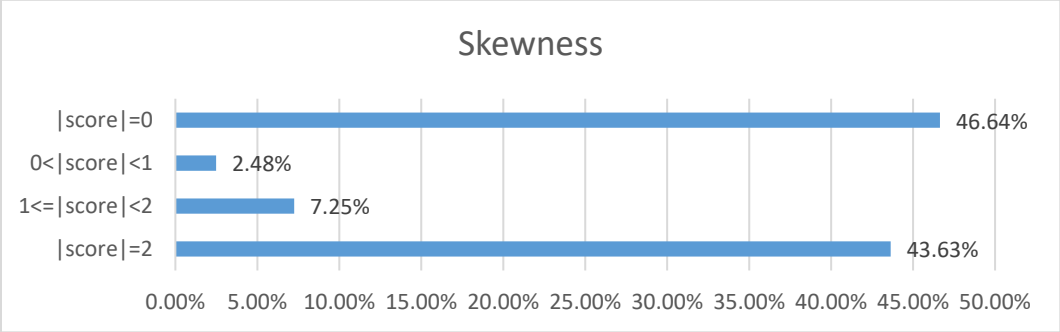


Figure 1 Skewness analysis

Contrast. For contrast evaluation, shown in Figure 2, images from 1930 to 1939 result in lowest contrast score (i.e., score of 23.87). And images from 1910 to 1919 result in highest score (i.e., score of 70.88). Note that, in this analysis, the higher the contrast the better the visual quality. Hence, based on the study

of Aida (i.e., score above 40 indicating good quality in contrast evaluation), except for images from 1860 to 1869 and from 1930 to 1939, the collection has a good contrast quality. However, there are 90% images from 1860 to 1869 in the collection. Hence, the 10-year chart (Figure 2) is not a good representation of the overall collection. As a result, we break the 10-year period from 1860 to 1869 into a year-by-year chart, shown in Figure 3.

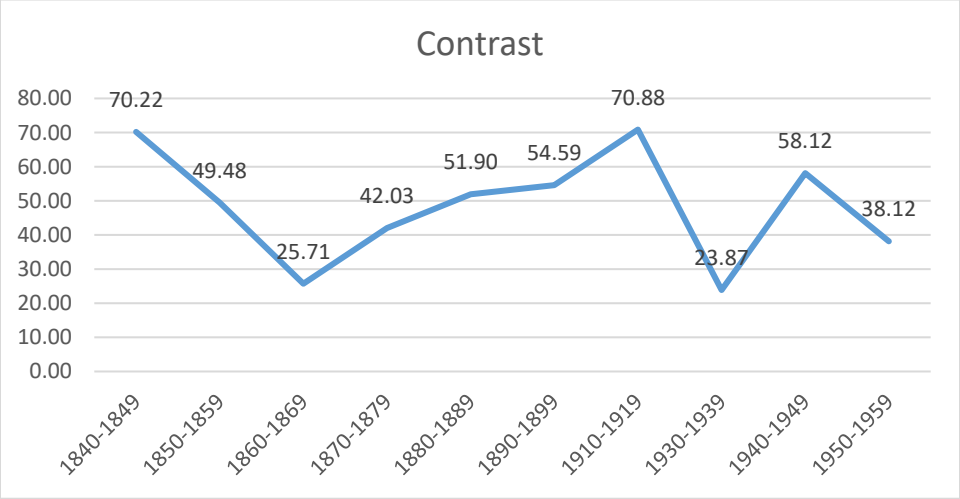


Figure 2 Contrast Score Analysis

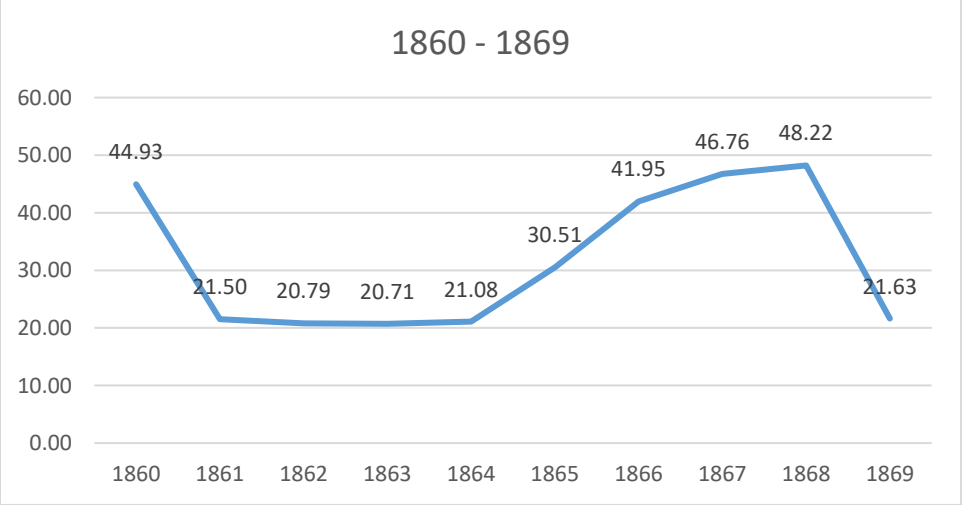


Figure 3 Contrast Score Analysis from 1860 to 1869

The breakdown chart shows that images with low score are from years 1861 to 1865. We suspect that the low score could be document images that are digitized from handwritten letters, shown in Figure 4. There are two problems among these letters that could lower the contrast score. First, the background largely suffers from yellowing. And, second, the ink is significantly faded. Further, we see that the appearance of low scores overlaps with the civil war years. Hence, the low score may also due to the degradation of the document considering the plausible challenges in newspaper preservation during the war.

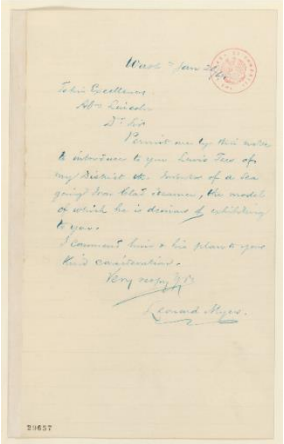


Figure 4 Handwritten letter with fading out ink

Range-effect. For range-effect evaluation, shown in Figure 5, images from all (but one) year ranges have relatively low scores. For the year range of 1930-1939, there are two with relatively high scores and hence the score of 27.33. Note that, for range-effect evaluation, the lower the score the better the quality. However, compared to our baseline study done in the Aida project that found any score below three implied good quality in range-effect evaluation, the civil war collection suffers from relatively more range-effect problems than the newspaper collection previously evaluated by Aida. This does not mean that the visual quality is necessarily visually for human perception. But it indicates that the collection could need substantial preprocessing to reduce range-effect before in-depth analysis.

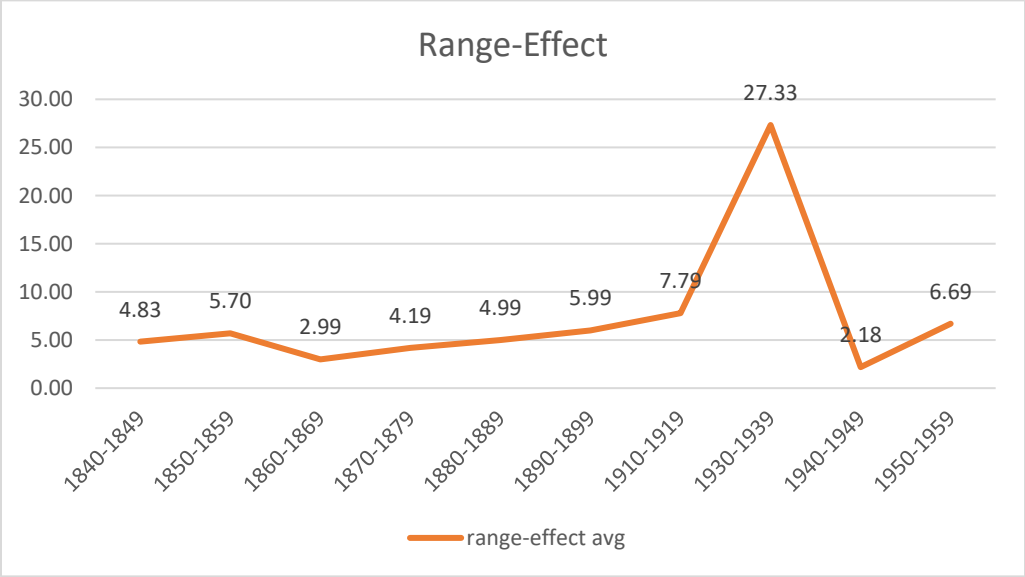


Figure 5 Range-effect Score Analysis

Bleed-through. For bleed-through (background noise) evaluation, shown in Figure 6, again, images from all year ranges (except one) have relatively low scores on bleed-through evaluation. For the year range 1940-1949, there are 76 images with high scores and hence the score of 12.10). Note that, the lower the bleed-through score the better the quality. However, a score identifying generally good quality does not exist for bleed-through evaluation. We can only confidently say that the score of zero is ideal.

Further, based on observation, the "paper yellowing" issue is a major problem in the collection. In our processing, a document image is first converted into a grayscale image by the evaluation algorithm. Hence, the yellowing paper results in a dark background after the conversion. A dark background would affect bleed-through evaluation, even, might result in a faulty evaluation. However, this does not mean that the bleed-through evaluation is not useful. Considering, in a way, the bleed-through evaluation represents the quality of background cleanliness, and thus, a high score can suggest that the background may need a noise removal process.

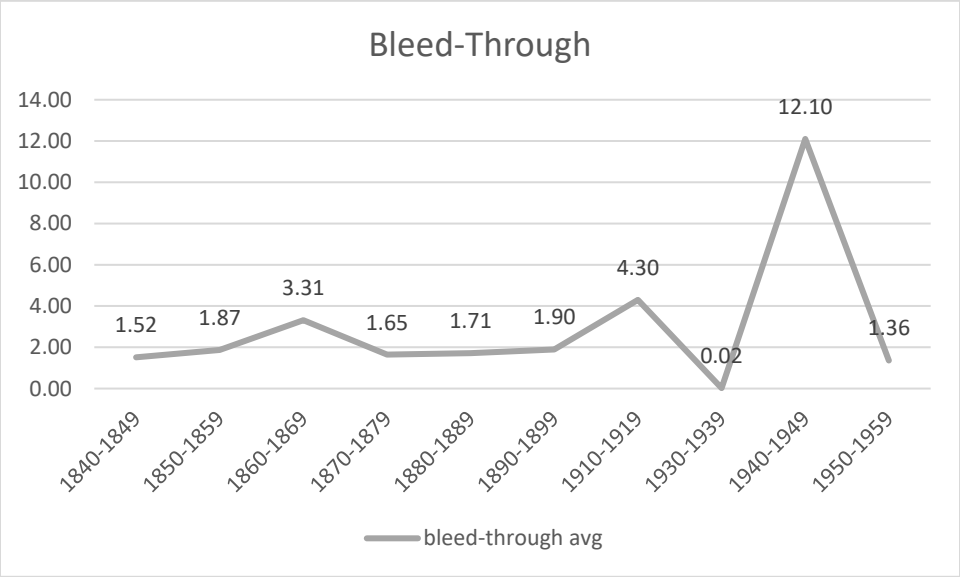


Figure 6 Bleed-through Score Analysis

Differentiation between Microfilm and Scanned images

The types of digitization that generated the document images are mixed. There are both microfilms and scanned images in the collection. As a result, techniques developed for one type might not work for the other. In our DIQA suite of image processing tools, for example, we assume that the document images were scanned images, with white or brighter pixels as background, and darker pixels as texts. However, documents from microfilm sometimes have inverted range of pixel intensities, rendering our image processing tools not effective. Hence, we propose a way to differentiate the digitization type of document to metatag them for further processing.

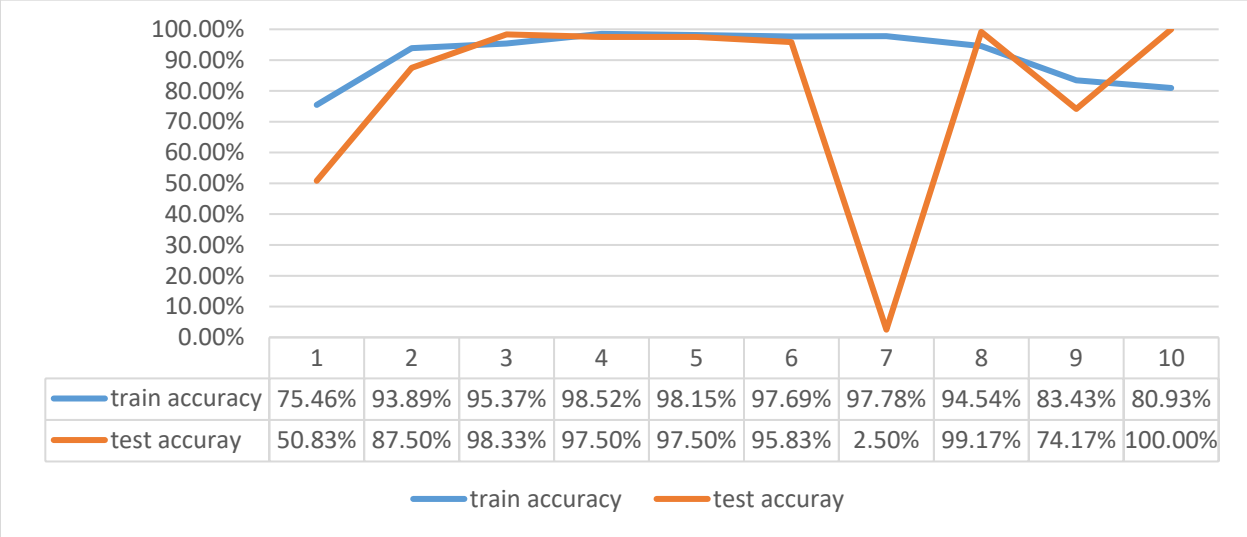
We propose to adopt the current state-of-art image classification model, called ResNeXt, to classify the digitization type of documents. In addition, to train the model, we need a set of labeled images. Hence, we manually build a database containing 1200 images from the civil war dataset. In this database, there are 600 scanned document and 600 documents from microfilm. A balanced database is built so that the training will not be biased.

Further, in a general idea of a machine learning training process, we want to keep the database as balanced as possible to prevent bias problem. This applies to not only numbers of instance for each label, but also other aspects such as skewness, contrast, range-effect, and bleed-through. In other words, we want our model to "see" as many conditions as possible during the training. Hence, during the creation of

the database, we randomize the file list to make each image in the collection has a fair chance to be included by the database.

Moreover, we also want to maintain replicability for future studies. So, the randomization was performed with a fixed random seed using a pseudo-randomization algorithm. By taking advantage of the randomization algorithm, we can reproduce the result as needed.

Shown in Figure 7, the ResNeXt model works very well on differentiating the two digitization types. The training process took only two iterations to reach over 90% accuracy. And test accuracy reached 100% correct at the 8th iteration. We see that the test accuracy at 7th iteration drops to 2.5%. This may be caused by the optimizer of the training process. The optimizer keeps a momentum to make the training process to be able to jump out of a local minimum. Hence, it may result in abnormal test accuracy. However, the test accuracy in these iterations does not necessarily affect the final performance of the classification as long as the training does not stop on these iterations.



	1	2	3	4	5	6	7	8	9	10
train accuracy	75.46%	93.89%	95.37%	98.52%	98.15%	97.69%	97.78%	94.54%	83.43%	80.93%
test accuray	50.83%	87.50%	98.33%	97.50%	97.50%	95.83%	2.50%	99.17%	74.17%	100.00%

Figure 7 Digitization Type Differentiation using ResNeXt-100 64x4d

Work That Has Been Done

Task 1

36,003 images from the civil war collection were downloaded through the website of By the People. And the downloaded image was backed up and stored in the CDRH server of the Aida team.

@cdrhdev2.unl.edu/var/local/aida/by-the-people_civil-war

Task 2

Collect information of creation/publication years of the corresponding item of the civil war collection for DIQA analysis.

@cdrhdev2.unl.edu/var/local/aida/by-the-people_civil-war/civil-war-images-info.csv

Task 3

Manually create a database containing 1200 images to perform training and evaluation for the digitization type classification.

@cdrhdev2.unl.edu/var/local/aida/by-the-people_civil-war/microfilms.txt

@cdrhdev2.unl.edu /var/local/aida/by-the-people_civil-war/scans.txt

Task 4

Adopt ResNeXt model from ImageNet-1000

Task 5

Create corresponding code to fine-tune and classify the digitization type.