

Progress report – Document image quality assessment for digital library collections

Yi Liu

Background

To better access digital library collections in terms of improving searchability, often document images—e.g., digitized or scanned from their original paper versions or microfilms—are tagged with metadata. Typically, metadata comes in two forms: (1) metadata about the document such as the original date of the document, publication venue, and so forth—also known as ancillary data in the realm of image processing; and (2) metadata about the texts directly discerned from the document, either through manual processing or natural language processing, such as keywords found in the texts, or titles extracted from the texts. However, as researchers begin to process large quantities of document images to develop robust classifiers or to develop generalizable automated systems, there is an increasing need for a third form of metadata: (3) metadata about the image quality of the document images such as average intensity of an image, contrast, range effects, layout structure, etc., such that researchers could query and retrieve specific subsets of document images based on these qualities for testing. It is this third form of metadata that motivates this report on document image quality assessment for digital library collections.

In general, image quality assessment includes both machine and human perceptions of quality of an image. For machine perception, quality assessment evaluates difficulties to predict or categorize an image for a machine. And for human perception, quality assessment evaluates difficulties to understand and interpret an image based on the visual appearance for human. In this report, our focus is document images.

In document image quality assessment (DIQA), there are two types of quality metrics [Ye and Doermann 2013]. One is called the *objective* quality metrics that is based on the ability to accurately predict the quality of a document image. For example, an optical character recognition (OCR) accuracy prediction model based on a convolutional neural network (CNN) [Kang et al. 2014] predicts the accuracy of OCR outcome for the document image. The other one is called the *subjective* quality metrics that define document image quality with respect to human perception. For example, a rating-based method assigns a categorical label to each image, such as the mean opinion score (MOS).

Problem Definition with respect to the Chronicling America's Repository

The Chronicling America repository (including Beyond Words and By the People) has little information on the document image quality. Note that there is an OCR accuracy quality score provided in the corresponding "ocr.xml" file. However, the score is not provided for all pages. Only some of the document pages have the corresponding score within its XML file. In addition, the score shows an objective score for OCR accuracy. It cannot intuitively indicate the quality of the image for human perception. Hence, a subjective score system is required to provide more quality information on human perception for further usage. For example, a school teacher might want to find some documents for his or her classroom activities. Intuitively, s/he might want document images with a clean background, good contrast, and less

content density. With a MOS system, the search query could be as easy as searching for document images with a good background, high contrast, and low density metadata.

State-of-the-Art

[Kang et al. 2014] proposed a shallow CNN model to predict the OCR accuracy for document images. The proposed model is shown in Figure 1. Further, they proposed to use a parallel min-max pooling before the dense layer. Such min-max pooling was able to maintain filter responses characterized by max and min values to capture statistical quality information for final score prediction.

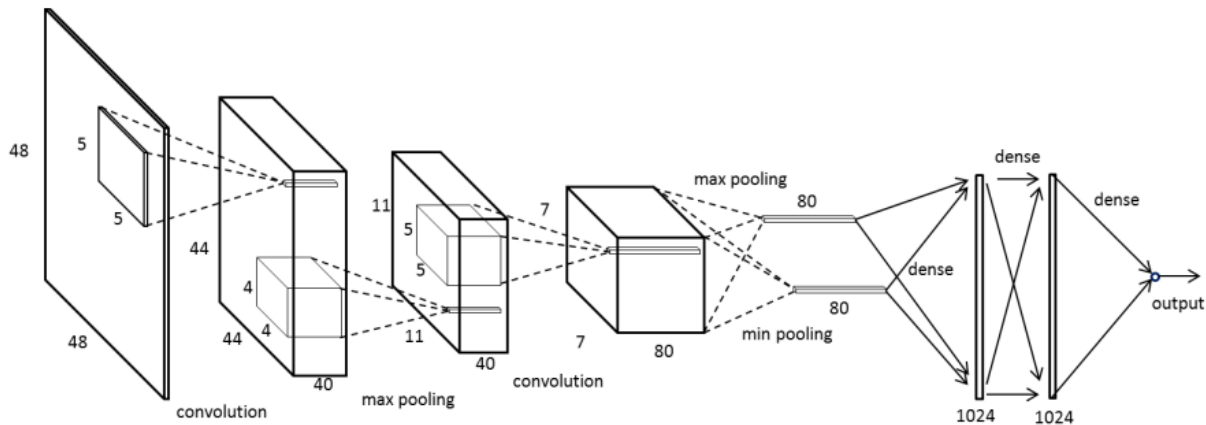


Figure 1 CNN model in [Kang et al. 2014] for OCR accuracy prediction.

MOS is widely used for subjective quality assessment [Ye and Doermann 2013]. However, there are challenges for subjective quality assessment. First, there is no existing human perception-based DIQA database to perform related experiments. Second, degradations could be present at different document levels, such as the character-level, the article-level, or the page-level. The appearance of multiple degradations increases the level of difficulty to design a global measurement. Third, a subjective quality assessment could be task-specific and might not be generalizable, as different tasks could command different values or emphases on how the quality of an image is judged or assessed.

Proposed Approach

Dataset Construction

Machine Learning, especially for deep learning, could require large amounts of labeled data to perform training. However, the lack of human perception-based DIQA database presents a challenge to investigations. We suggest **adding an interface to allow a user to describe the quality of the document images using five-level rating score, such as MOS (i.e., 5-Excellent, 4-Good, 3-Fair, 2-Poor, and 1-Bad), on aspects such as contrast, range-effect, background-cleanliness, and content density.** Over time, a human perception based DIQA database could be established to support studies and experiments, and could even be made publicly available for research competition for academia.

Integrating Existing Work

In the work of Image Analysis for Archival Discovery (Aida), an objective DIQA experiment was carried out to evaluate historical newspapers pages from 1834 to 1922 in the Chronicling America repository. The

objective DIQA aimed to evaluate four metrics for the newspaper page across different languages in different eras. These four metrics that could be automatically computed included (1) the skewness, (2) the contrast, (3) the range-effect, and (4) the bleed-through (Examples are shown in Appendix). The results of the experiment were numeric scores ranging based on algorithmically understanding. As these metrics were numeric, it would require human expertise to better interpret the results such as the range of values for an image to be considered of high contrast or low contrast. Furthermore, it would require application-specific needs to leverage these metrics; for example, how high the range-effect would have to be for an image to be rendered not usable or interpretable for a particular application such as natural language processing?

However, this is not necessarily suggesting the existing work is useless for subjective DIQA. With additional works, the existing objective DIQA results of Aida could be helpful. These works include: (1) pre-defining the range of the score that makes sense to human users; e.g., numerical scores on range-effect within 0 to 1 may be considered excellent, within 1 to 2 good, within 3 to 5 fair, within 5 to 6 poor, and finally, larger than 6 bad; and (2) normalizing numeric scores based on the pre-defined range for each metric for subjective DIQA experiments.

Deep Learning-Based Experiment

We propose an inference multi-output U-NeXt to perform a subjective DIQA using MOS. The main architecture of the model is a combination of ResNeXt [Xie et al. 2017] and U-Net [Ronneberger et al. 2015] that is attached by a min-max pooling and two dense layers, shown in Figure 2. Each output corresponds to one aspect of a five-level MOS. Note that the U-NeXt model will not be trained from scratch. A pre-trained model using ImageNet [Russakovsky et al. 2015] and ENP [Clausner et al. 2015] database will be adopted. By using transfer learning, a pre-trained model can help us to reduce numbers of training parameters and to make the training process faster.

In the current stage, we could perform experiments based on the normalized objective DIQA scores from the project Aida. Hence, for each newspaper page from 1834 to 1922 in the Chronicling America's repository, four quality metrics are included in the ground-truth on the skewness, contrast, range-effect, and bleed-through using MOS. Then, based on the ground-truth and data, we can train the U-NeXt model to rate the newspaper page subjectively using MOS. Such configuration would be able to show the strength of the model on subjective DIQA tasks. However, because the subjective score is a pseudo-score based on algorithmic score, they are not necessarily able to accurately represent the actual human perception. Hence, further experiments to evaluate the effectiveness of the subjective DIQA using the U-NeXt requires an actual human perception-based DIQA database would be helpful.

Reference

- [1] Clausner, C., Papadopoulos, C., Pletschacher, S., & Antonacopoulos, A. (2015, August). The ENP image and ground truth dataset of historical newspapers. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 931-935). IEEE.
- [2] Kang, L., Ye, P., Li, Y., & Doermann, D. (2014, October). A deep learning approach to document image quality assessment. In *2014 IEEE International Conference on Image Processing (ICIP)* (pp. 2570-2574). IEEE.
- [3] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
- [4] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211-252.
- [5] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500).
- [6] Ye, P., & Doermann, D. (2013, August). Document image quality assessment: A brief survey. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 723-727). IEEE.

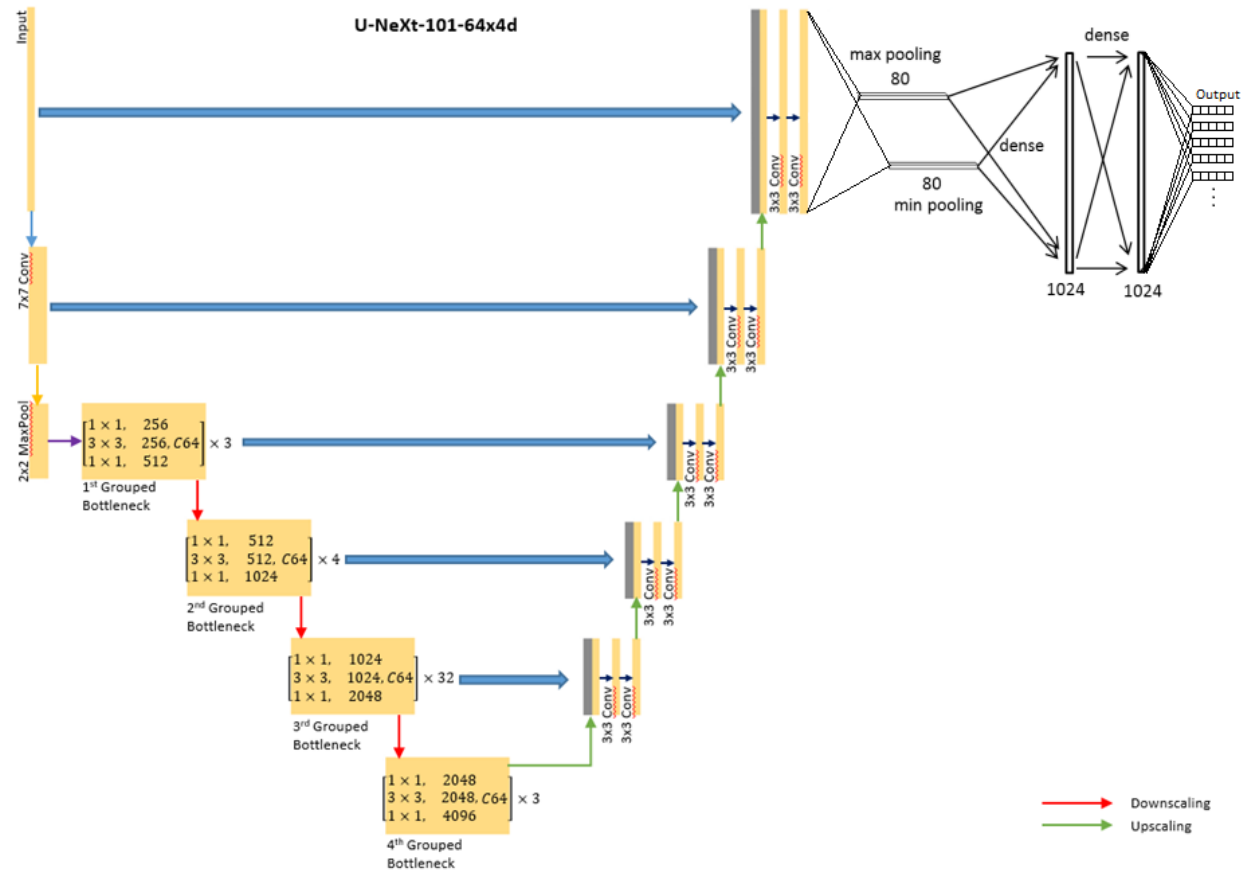


Figure 2 Inference multi-output U-NeXt model for subject DIQA.

Appendix

Table 1 Examples of newspaper pages having different levels of contrast, range-effect, and bleed-through.



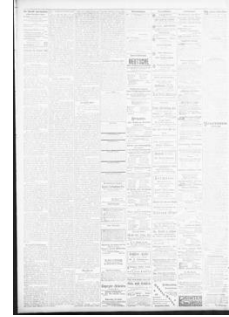











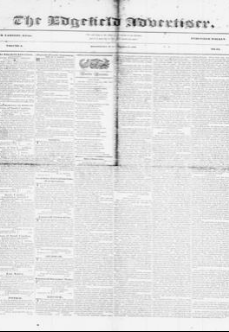


	High/Severe		Some		Low/None	
	Value	Image	Value	Image	Value	Image
Contrast	146.08	 1834-1922FullPages_20PerYr/1868_English/sn82014064/1868-07-18/ed-1/seq-4.jp2	55.2	 1834-1922FullPages_20PerYr/1848_English/sn83035366/1848-03-16/ed-1/seq-1.jp2	3.11	 1834-1922FullPages_20PerYr/1898_German/sn83045081/1898-12-29/ed-1/seq-2.jp2
Range-effect	14.11	 1834-1922FullPages_20PerYr/1896_English/sn88083938/1896-04-18/ed-1/seq-1.jp2	4.0	 1834-1922FullPages_20PerYr/1867_Spanish/2013201074/1867-02-09/ed-1/seq-4.jp2	0.0	 1834-1922FullPages_20PerYr/1904_Icelandic/sn90060662/1904-12-01/ed-1/seq-12.jp2
Bleed-through	0.129	 1834-1922FullPages_20PerYr/1861_Spanish/2013201074/1861-03-16/ed-1/seq-4.jp2	0.033	 1834-1922FullPages_20PerYr/1856_English/sn85026050/1856-08-15/ed-1/seq-4.jp2	0.001	 1834-1922FullPages_20PerYr/1907_Icelandic/sn90060662/1907-09-01/ed-1/seq-6.jp2

Table 2 Examples of newspaper pages having different levels of skewness.

Skewness			
Title	Value	Image	Note
/Archive/sn83016788_1840-05-26_ed-1_seq-2.jpg	0.0		No skewness
/Archive/sn83016788_1840-07-17_ed-1_seq-1.jpg	-0.5		Slightly tilting to left
/Archive/sn85025180_1837-10-14_ed-1_seq-1.jpg	0.5		Slightly tilting to right
/Archive/2013201074_1837-05-16_ed-1_seq-3.jpg	0.75		Slightly tilting to right

/Archive/2013201074_1837-01-24_ed-1_seq-3.jpg	-1.0		More tilting to left
/Archive/sn84026897_1838-09-27_ed-1_seq-1.jpg	1.0		More tilting to right
/Archive/sn84026897_1838-09-20_ed-1_seq-1.jpg	1.0		More tilting to right
/Archive/sn84026897_1840-05-28_ed-1_seq-3.jpg	2.0		More tilting to right