https://doi.org/10.2352/ISSN.2470-1173.2021.16.COLOR-243

Deep Learning Approaches to Determining Optimal Resolution for Scanned Text Documents*

Litao Hu[a], Zhenhua Hu[a], Peter Bauer[b], Todd J. Harris[b], Jan P. Allebach[a]
[a]Purdue University; West Lafayette, IN 47906, U.S.A.
[b]HP Inc.; Boise, ID 84714, U.S.A.

Abstract

Image quality assessment has been a very active research area in the field of image processing, and there have been numerous methods proposed. However, most of the existing methods focus on digital images that only or mainly contain pictures or photos taken by digital cameras. Traditional approaches evaluate an input image as a whole and try to estimate a quality score for the image, in order to give viewers an idea of how "good" the image looks. In this paper, we mainly focus on the quality evaluation of contents of symbols like texts. Judging the quality for this kind of information can be based on whether or not it is readable by a human, or recognizable by a decoder such as an OCR engine. We mainly study the quality of scanned documents in terms of the detection accuracy of its OCR-transcribed version. For this purpose, we proposed a novel CNN based model to predict the quality level of scanned documents or regions in scanned documents. Experimental results evaluated on our testing dataset demonstrate the effectiveness and efficiency of our method both qualitatively and quantitatively.

Introduction

Nowadays, scanners on multi-functional printers (MFPs) are very commonly used both in offices and at home to digitize printed documents, drawings and hand-written documents, for convenient distribution. In most cases, these digitized documents will eventually be viewed on screens by human beings or fed into other software or algorithms for other purposes. With an MFP, the user usually needs to select a scanning resolution when scanning a document. With high resolution such as 600 DPI or 300 DPI, the user may end up with a file that is excessively large for distribution. On the other hand, when scanning at low resolution, such as 100 DPI or 75 DPI, the quality degradation may be very severe, resulting in loss of information. Therefore, choosing an appropriate resolution can sometimes be a very tricky task, since it depends on both the purpose of the scan and the content to be scanned.

When evaluating the quality of a photo, we usually consider various aspects, both aesthetically and perceptually, and the perceived quality can sometimes be very subjective and depend heavily on the preference of the viewer. On the other hand, evaluating the quality of contents such as texts, lines, bar-codes, QR-codes, and hand-writing is very direct and can be easily determined. For such content, we take readability or repurposability to be the only and necessary factors that determine the viewing quality. Therefore, a method that estimates readability or repurposability would be a better measure for quality of such contents in document images.

In this paper, as opposed to predicting minimum readable resolution (MAR), we seek the minimum scanning resolutions of scanned documents that ensure quality for decent OCR accuracy, i.e. minimum repurposable resolution (MRR). Here, we define repurposability as whether the contents in the document image can be detected and decoded by a computer algorithm (such as an OCR engine) with decent accuracy. For this purpose, we propose several models for document image quality assessment that can be used to estimate the MRR of document images. In addition, we design a compression system that first segments a document image into different regions of interest and then estimates the optimal scanning settings for these regions. Finally, the system outputs a compressed digital file (such as PDF) resampled based on the estimated optimal quality settings. The system diagram is shown in Figure 1.

There are three main contributions in this paper.

1. We propose a compression framework for scanned document images.
2. We propose to use a page segmentation algorithm to segment document images and apply region-specific algorithms to different regions.
3. We propose a CNN-based model for estimating minimum repurposable resolution of symbol regions in document images.

In the following sections, we introduce the proposed CNN-based approach for predicting MRR. Experimental results for our proposed models are presented and compared against a very popular model designed for mobile devices.

Related Work

There are numerous research papers and methods for image quality assessment. Based on whether a reference image is used when assessing a target image, the methods can be divided into three groups, namely full-reference (FR) image quality assessment, reduced-reference (RR) image quality assessment, and no-reference (NR) image quality assessment. Full-reference image quality assessment, or FR-IQA, estimates the quality score by comparing the target image with a reference image that, in most cases, has high quality. Representative works on FR-IQA include [1, 2, 3]. At the opposite extreme, no-reference image quality assessment, or NR-IQA, tries to estimate the quality score of the target image in the absence of a reference image. Based on the purposes, NR-IQA can be further divided into two categories,

*Research supported by HP Inc., Boise, ID 83714

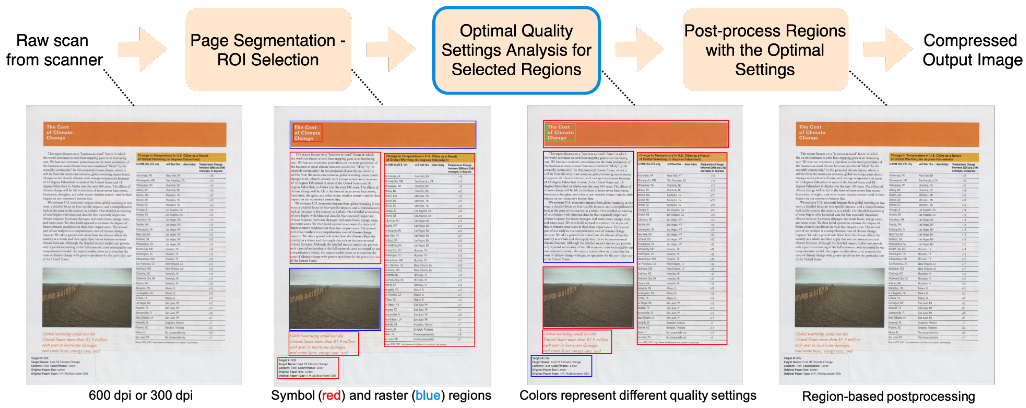

Figure 1: System Diagram.

distortion-specific NR-IQA (DS-NR) and general-purpose NR-IQA (GP-NR). Representative works on DS-NR include [4, 5, 6]; and representative works on GP-NR include [7, 8, 9]. Betweem FR-IQA and NR-IQA, RR-IQA tries to use less information from the reference image, or only uses part of the reference image, to achieve high accuracy in quality score estimation. Representative works include [10, 11]. In this paper, we propose a full-reference image quality assessment method based on CNN and designed specifically for low resolution degradation. To the best of our knowledge, no similar resolution-specific FR-IQA method like ours has been proposed before.

Deep learning has achieved success in many computer vision and image processing tasks in recent years, becoming the state-of-the-art in many areas. Thanks to the computational power provided by modern graphic processing units (GPUs) and the availability of large-scale datasets, people can now easily fit complex non-linear functions with great representation ability, constructed from architectures like convolutional layers or multilayer perceptrons.

However, many of the state-of-the-art methods based on deep learning rely heavily on the tremendous computational power of GPUs to work efficiently, which is not available in a lot of real-time applications on mobile devices. Therefore, it has been an active research area to design light-weight models with acceptable performance that can be easily implemented on mobile platforms where a GPU is absent. MobileNet[12], EfficientNet[13], GhostNet[14] are such examples.

In our project, algorithms will eventually be implemented on ARM-based CPUs, which are included on most HP MFPs. With the extremely limited computational power, it is therefore of great importance to design a light-weight and efficient model in order to make it fast enough to run in real-time.

Proposed Method

Optimal Resolution Prediction as a Classification Task

As shown in Figure 1, the system consists of three parts. In the first part, the image of a scanned document is segmented and multiple rectangular regions will be located, which will then be classified as symbol regions, raster regions, and vector regions. In the second part, trained predictive models will then be used to estimate the optimal scanning settings for different types of regions. Finally, in the last part, regions will be post-processed according the optimal quality settings to form an optimally resampled output file.

In this paper, we focus on the task of estimating the MRRs for symbol regions in document images in the second part of the compression framework. The first part of the framework, page segmentation, has been described in details in [15]. We adopt the same page segmentation algorithm in this paper. Since finding an exact optimal resolution is not necessary for us, similar to [15], we simplify the output for MRR to 4 tiers, as illustrated in Figure 2. The 4 tiers of MRR are as follows:

1. Tier 0: Minimum repurposable resolution is the base resolution (e.g. 600 DPI);
2. Tier 1: Minimum repurposable resolution is the base resolution divided by 2;
3. Tier 2: Minimum repurposable resolution is the base resolution divided by 4;
4. Tier 3: Minimum repurposable resolution is the base resolution divided by 8.

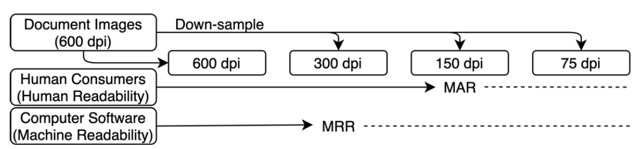

Figure 2: Predicting MAR and MRR as a Classification Task.

In our implementation, we choose 600 DPI (the upper limit of scanning resolution of MFPs) as our base resolution, so the following tiers are 300 DPI, 150 DPI, and 75 DPI (the lower limit of scanning resolution of MFPs), respectively. In this section, we will use 600 DPI as an equivalence to the base resolution without further mention.

Light-weight CNN for efficient Optimal Resolution Prediction

In this section, we will present a CNN-based approach to assess the quality of symbol regions in document images at various resolutions and predict the minimum repurposable resolution for the documents.

We start by building a simple baseline model using basic 2D convolutional layers and a fully-connected layer. The model structure for our baseline model is shown in Figure 3. We want to build an end-to-end model that takes an input image at the base resolution as input (in our case it is 600 DPI), and outputs the likelihood of the input image being in each tier of resolution. We follow a standard structure for a classification model such as LeNet when designing this model. It consists of 6 convolutional layers for feature extraction and 1 linear layer for classification. We apply ReLU non-linear activation functions between convolutional layers and a dropout layer before the linear layer. We also added batch normalization after every convolutional layer, in order to make the training process more stable and allow for higher learning rates.

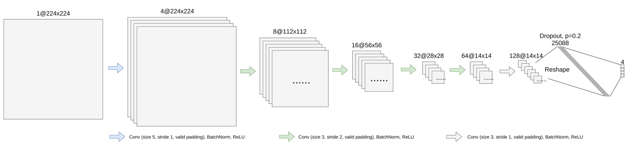

Figure 3: Baseline Model Structure for Predicting Optimal Repurposable Resolution.

Inspired by the method introduced in [15], which extracts MBF features from multiple scales of the input image, we also designed a light-weight neural network with structures to learn features from multiple scales of the input. The network structure is shown in Figure 4. The combined use of average pooling of different kernel sizes and dilated convolutions in the multiscale module allows for features learned from multiple scales of the input image and a significant reduction in the number of floating-point operations (FLOPs). For convenience, we name this model MultiScaleNet in this paper. We also propose another version of the model named MultisSaleNet-IRB, which utilizes inverted residual blocks [12], instead of regular convolutional layers after the multi-scale module, for performance comparison. The structure of this model is shown in Figure 5.

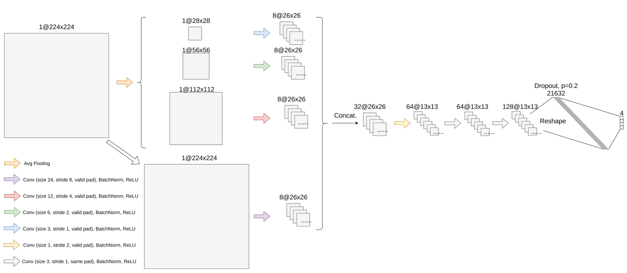

Figure 4: Multi-scale Model Structure with Regular Convolutional Layers for Predicting MRR.

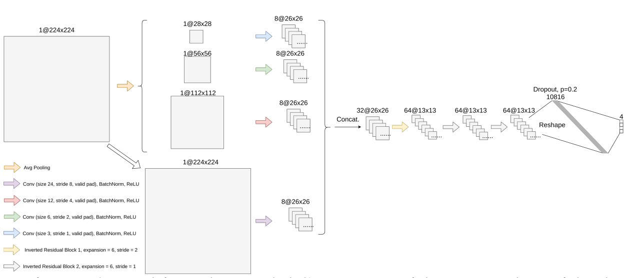

Figure 5: Multi-scale Model Structure with Inverted Residual Blocks for Predicting MRR.

As shown in the model structures in Figure 4 and Figure 5, the input to the model is a single channel fixed-size square image. The input size is chosen to be 224×224 as is used in many other popular deep learning models such as MobileNet. We follow the same standard for easier comparison between our models and those published models during evaluation. Following the multi-scale layers, we concatenate the feature maps learned from various scales and use 3 convolutional layers or inverted residual blocks to combine and further extract features. Finally, we use a single linear layer as a classifier to classify the learned features into 4 classes. The output is therefore a vector of size 4×1 where each entry represents the likelihood for the input being predicted to each class with the corresponding index. In this work, the 4 classes are the 4 possible minimum repurposable resolutions, namely 600 DPI, 300 DPI, 150 DPI, and 75 DPI. We use Softmax non-linear activation along with Cross Entropy Loss during training to optimize our model, as is commonly adopted in many other classification CNNs. The reasons for using Softmax function combined with Cross Entropy Loss can be found in [16].

Training Data Collection

For the training data, since the input is of size 224×224, while a scanned document page can have a size of 6600×5100 at 600 DPI, we need to partition the scanned document image into smaller tiles or patches that can be fed into our models. Note that it is not a good idea to directly down-sample the input images or regions of images to the required size, because that will decrease the quality of the input image. Similar to [15], we print 100 document pages containing English psudo-texts (Lorem Ipsum) in different sizes and fonts. We then scan these printed pages at 600 DPI resolution, without post-processing, and partition the scanned images into non-overlapping image patches of size 256×256.

To collect ground-truth MRRs for these image patches, we utilize an open-source OCR engine named Tesseract to first generate transcribed text files for the image patches and their down-sampled versions. We then compare the generated text files against their corresponding source texts, and obtain the corresponding OCR accuracy. From here, there are two ways of using these OCR accuracies. To formulate the task as a classification problem, we can apply a threshold to the OCR accuracy (for example, 85%) to find the minimum repurposable resolutions, and use them as labels for the image patches. The other way is to formulate the task as a regression problem, where we trained a model to directly predict the OCR accuracy associated with the input image patch and its down-sampled versions. In this project, we go with the latter approach and formulate the task as a classification problem. To be specific, by comparing these OCR accuracies to a preset threshold such as 0.85 (85%), we claim that the resolution is not repurposable for a given image if the OCR accuracy for that resolution is lower than 0.85, and vise versa. We then use the minimum repurposable resolution as the ground-truth label for that document image. After that, we partition these document images into image patches of size 256×256; and these image patches will inherit the ground-truth label from their corresponding parent document image. In this way, we managed to collect 30466 image patches in total.

We then split the dataset with 30466 image patches into three separate sets, namely training set, validation set, and testing set, each contains 21322, 6092, and 3050 samples, respectively. Because our dataset is unbalanced, we apply random duplication to randomly over-sample classes with fewer training samples. To be specific, we over-sampled the training set randomly, so that every class contains 7000 samples, which yields 28000 samples in the training set in total.

Experimental Results
Training Setup

As introduced in previous sections, we propose three different CNNs, as shown in Figure 3, Figure 4, and Figure 5. Besides the three models, we also fine-tuned MobileNetv2, a popular neural network designed for mobile platforms. In this section, we experiment with the four different architectures and compare their performances.

We trained the models to predict MRRs using the dataset we generated. During training, we augmented our training set dynamically by randomly cropping the input images to 224×224 and randomly flipping the images vertically.

We trained our models on a single NVIDIA Geforce 1080Ti GPU with 11Gb RAM. Note that for MobileNetv2, we loaded the pre-trained parameters and fine-tuned the whole network for only 50 epochs. The training pipeline for all models is shown in Figure 6. We used Cross Entropy Loss and Adam optimization to update the model parameters. In addition, all four models are trained with a fixed learning rate $1e-5$. Other training settings for our experiments are summarized in Table 1.

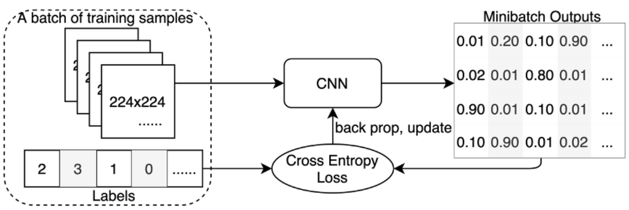

Figure 6: Training Pipeline.

Finally, we adopted model quantization technique in this project to further optimize the inference speed of our neural networks on CPU, making it suitable for implementation on an ARM CPU for real-time applications.

Performance Analysis

The training and validation loss curves for the four models are shown in Figure 7, Figure 8, Figure 9, and Figure 10. From these curves, we observed that the models tends to slightly overfit the training dataset. Therefore, we adopted an early-stopping technique and selected the models at earlier iterations where the validation loss curves and training loss curves intersect.

Since we are evaluating on an unbalanced testing dataset, we also generate the confusion matrices for the four models besides the overall accuracy, as shown in Figure 11, Figure 12, Figure 13, and Figure 14.

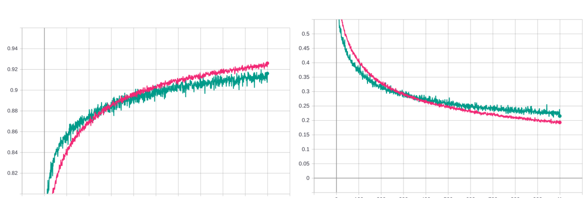

Figure 7: Training and Validation Curves for CNN baseline Model. Left: Accuracy vs. Epoch; Right: Loss vs. Epoch. Magenta: Training; Green: Validation.

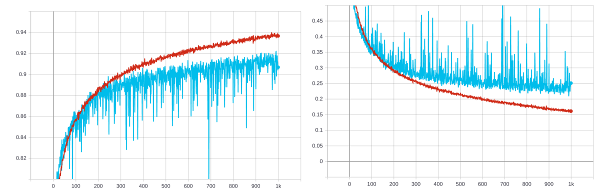

Figure 8: Training and Validation Curves for MultiScaleNet Model. Left: Accuracy vs. Epoch; Right: Loss vs. Epoch. Red: Training; Cyan: Validation.

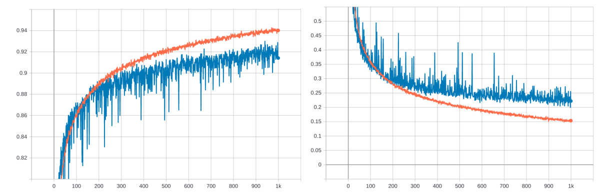

Figure 9: Training and Validation Curves for MultiScaleNet Model with Inverted Residual Blocks. Left: Accuracy vs. Epoch; Right: Loss vs. Epoch. Orange: Training; Blue: Validation.

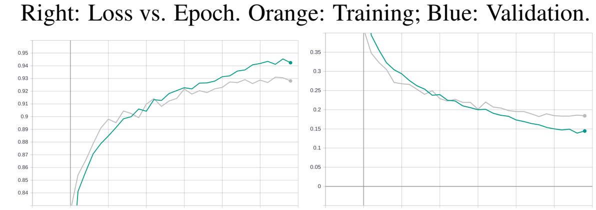

Figure 10: Training and Validation Curves for MobileNetv2 Model. Left: Accuracy vs. Epoch; Right: Loss vs. Epoch. Green: Training; Gray: Validation.

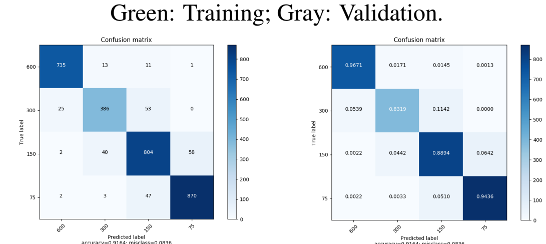

Figure 11: Confusion Matrices for CNN Baseline Model on Testing Set. Left: Not Normalized; Right: Normalized.

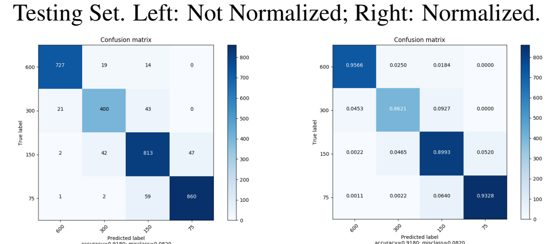

Figure 12: Confusion Matrices for MultiScaleNet Model on Testing Set. Left: Not Normalized; Right: Normalized.

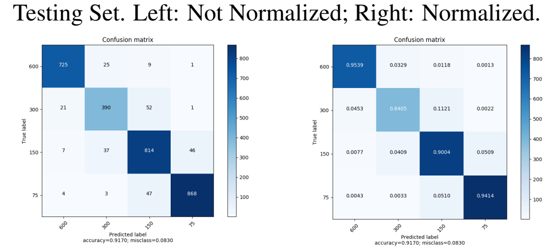

Figure 13: Confusion Matrices for MultiScaleNet Model with Inverted Residual Blocks on Testing Set. Left: Not Normalized; Right: Normalized.

Models	Epochs	Loss Function	Optimizer	Learning Rate	Pre-trained
Our Models	1000	Cross Entropy Loss	Adam	1e-5	No
MobileNetv2	50	Cross Entropy Loss	Adam	1e-5	Yes

Table 1: Training Setup for Our Experiments.

Models	Parameters	FLOPs	Latency/ms (Intel i7 8700)	Latency/ms (Raspberry Pi CPU)
CNN Baseline	0.199M	34.8M	1.47	32.90
MultiScaleNet	0.222M	26.2M	1.10	17.23
MultiScaleNet-IRB	0.179M	29.4M	2.18	19.40
MobileNetv2	3.51M	359.6M	18.56	102.0

Table 2: Model Sizes, Complexity, and Latencies.

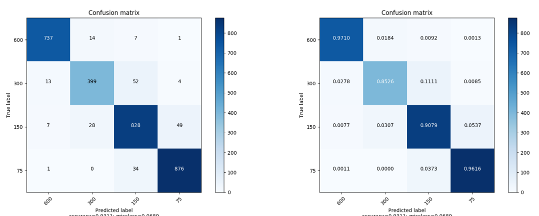

Figure 14: Confusion Matrices for MobileNetv2 Model on Testing Set. Left: Not Normalized; Right: Normalized.

The best validation accuracies and test accuracies (with early stopping) for the four models are summarized in Table 3. Based on the testing accuracy, MobileNetV2 achieved the best performance among the four models, followed by MultiscaleNet-IRB, MultiscaleNet, and CNN Baseline. Our proposed MultiScaleNet and MultiScaleNet-IRB achieved very close performance to the fine-tuned MobileNetV2.

Models	Best Val. Acc.	Test Acc.
CNN Baseline	92.01%	91.64%
MultiScaleNet	92.17%	91.80%
MultiScaleNet-IRB	93.03%	92.07%
MobileNetv2	93.11%	93.11%

Table 3: Validation Accuracy and Test Accuracy for Different Models.

Since we will eventually deploy our neural networks on an ARM-based CPU which possess limited computational power, we have to make our model light-weight. Therefore, we also measure the number of parameters and FLOPs for each model to compare their complexity. We ran the four models on an Intel i7 8700 CPU, as well as a Raspberry CPU, and measure their latencies. The measurements are summarized in Table 2. Latencies are computed by taking the average over runtime of all testing samples.

To deploy our model on a Raspberry Pi4, we adopted a model quantization technique to convert our models from floating point operations to integer operations, which is more efficient for Raspberry Pi's ARM-based CPU. There are a couple of ways to perform model quantization, such as dynamic quantization, static quantization, and quantization aware training (QAT). In this project, we adopt quantization aware training to fine-tune our trained models for 10 epochs to best preserve the performance after quantization.

Based on these tests on the performance and complexity, We can easily see that the MultiScaleNet and MultiScaleNet-IRB are able to achieve very close performance with MobileNetv2 with far fewer parameters and FLOPs, which allows them to run much faster on both the Intel CPU and the Raspberry CPU, and are thus more suitable for deployment. Comparing across models, we can see that the multi-scale structures can be a very effective module in our applications to extract multi-scale features efficiently, and the inverted residual blocks are also very effective in learning from the multi-scale features.

Simulated Document Automatic Compression based on Optimal Resolution Prediction

When testing the trained models on real document pages, we need to first partition the symbol regions into non-overlapping patches and pass all the patches to the model and pool all the predictions from the same region into one final prediction for that region. In our implementation, we used most frequent pooling, i.e. select the most frequent prediction from all the predictions as the final prediction for the region of interest. The inference pipeline is shown in Figure 15.

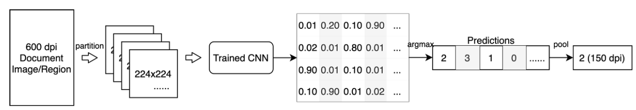

Figure 15: Inference Pipeline.

To demonstrate the effectiveness of our trained MultiScaleNet-IRB in real applications and to evaluate it qualitatively, we run a few test pages with our algorithm and the visualization of their outputs are shown in Figure 16. In the output test pages, we color-coded the bounding boxes of detected regions of interest to indicate the optimal repurposable resolutions. To be specific, we use red, yellow, blue, and green to indicate 600 DPI, 300 DPI, 150 DPI, and 75 DPI, respectively.

To show the amount of compression our model provide, we also measure the actual sizes of the compressed outputs for these test pages, as well as their original sizes without compression by our model, in different file formats. The measured sizes are summarized in Table 4. We can see that our output files are significantly smaller compared to their original counterparts.

Our models have successfully generated satisfying results; and the system can properly produce compressed scans of document pages according to the estimated optimal resolutions.

Conclusion

In conclusion, we proposed a novel document image quality assessment method to estimate the minimum repurposable resolution of scanned documents or regions in scanned documents. Our experiments successfully demonstrated the effectiveness and efficiency of our proposed methods. For now, our system is only tested with English text, but the same idea can be easily expanded and applied to other languages, as well as other symbol contents

	TIFF(original)	TIFF(ours)	JPEG(original)	JPEG(ours)	JB2(original)	JB2(ours)
Test Page 1	48918.18	13199.97	3479.18	1201.95	192.33	103.29
Test Page 2	49236.45	3123.83	3374.40	362.49	97.28	26.46
Test Page 3	47523.64	35006.44	3393.65	2513.77	115.87	82.80

Table 4: Output File Sizes (kB).

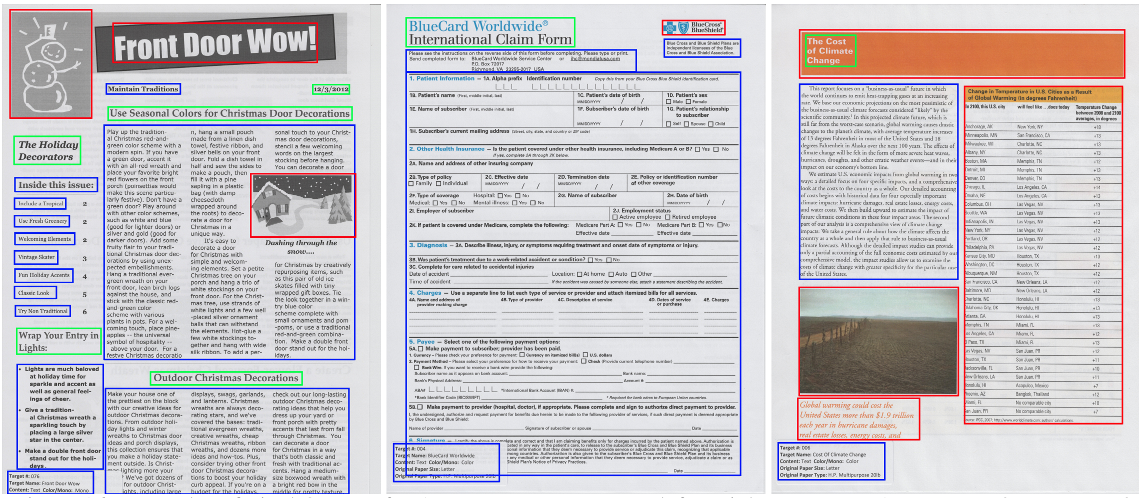

Figure 16: Examples of Final Outputs for A Few Test Pages. From left to right: Test Page 1, Test Page 2, Test Page 3. Red: 600 DPI; Yellow: 300 DPI; Blue: 150 DPI; Green: 75 DPI.

that share characteristics similar to text.

References

[1] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, Nov 2003, pp. 1398–1402 Vol.2.

[3] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug 2011.

[4] Zhou Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proceedings. International Conference on Image Processing*, vol. 1, Sep. 2002.

[5] T. Brandão and M. P. Queluz, "No-reference image quality assessment based on DCT domain statistics," *Signal Processing*, vol. 88, no. 4, pp. 822 – 833, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168407003337

[6] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1918–1927, Nov 2005.

[7] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, May 2010.

[8] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583–586, June 2010.

[9] W. Lu, K. Zeng, D. Tao, Y. Yuan, and X. Gao, "No-reference image quality assessment in contourlet domain," *Neurocomputing*, vol. 73, no. 4, pp. 784 – 794, 2010, Bayesian Networks / Design and Application of Neural Networks and Intelligent Learning Systems (KES 2008 / Bio-inspired Computing: Theories and Applications (BIC-TA 2007). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231209003890

[10] X. Gao, W. Lu, D. Tao, and X. Li, "Image quality assessment based on multiscale geometric analysis," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1409–1423, July 2009.

[11] D. Tao, X. Li, W. Lu, and X. Gao, "Reduced-reference IQA in contourlet domain," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 6, pp. 1623–1627, Dec 2009.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[13] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[14] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," 2019.

[15] L. Hu, Z. Hu, P. Bauer, T. Harris, and J. Allebach, "Document image quality assessment with relaying reference to determine minimum readable resolution for compression," *Electronic Imaging*, vol. 2020, no. 9, pp. 323–1–323–8, 2020. [Online]. Available: https://www.ingentaconnect.com/content/ist/ei/2020/00002020/00000009/art00027

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

Author Biography

Litao Hu received his BS in Electronic Engineering from the Hong Kong University of Science and Technology (2017) and is

currently a PhD candidate in Electrical and Computer Engineering at Purdue University. As a research assistant in the Electronic Imaging System Laboratory, his research interests include image processing, machine learning, and deep learning.