DOCUMENT QUALITY ESTIMATION USING SPATIAL FREQUENCY RESPONSE

Pranjal Kumar Rai[1] *Sajal Maheshwari*[2*] *Vineet Gandhi*[1]

[1]CVIT, IIIT-Hyderabad, [2]Qualcomm India Pvt. Ltd.
pranjal.kumarrai@research.iiit.ac.in, sajal.maheshwari@alumni.iiit.ac.in, vgandhi@iiit.ac.in

ABSTRACT

The current Document Image Quality Assessment (DIQA) algorithms directly relate the Optical Character Recognition (OCR) accuracies with the quality of the document to build supervised learning frameworks. This direct correlation has two major limitations: (a) OCR may be affected by factors independent of the quality of the capture and (b) it cannot account for blur variations within an image. An alternate possibility is to quantify the quality of capture using human judgement, however, it is subjective and prone to error. In this work, we build upon the idea of Spatial Frequency Response (SFR) to reliably quantify the quality of a document image. We present through quantitative and qualitative experiments that the proposed metric leads to significant improvement in document quality prediction in contrast to using OCR as ground truth.

***Index Terms*—** Document image; Image quality; Image capture; Image analysis; Spatial resolution

1. INTRODUCTION

Phone cameras are becoming the default choice for document image capture due to: (a) the improvement in the quality of phone cameras and (b) the portability and ease of sharing. However, the casual nature of capture using phone cameras (as compared to controlled capture using scanners) may lead to instability and in turn degrade the image quality. These degradations may lead to challenges in automated document workflows, often employing OCR algorithms. DIQA algorithms can resolve this problem by either reducing manual intervention (by identifying poorly captured images) or by giving real-time feedback during the capture.

A major obstacle in DIQA is to generate a ground truth value to quantify the image quality. Most state-of-the-art algorithms make use of OCR accuracies as ground truth for document images. It is true, that there is a broad correlation between OCR accuracies and quality of the capture i.e. poorly captured document lead to low OCR accuracy and sharp capture results in higher OCR accuracy. However, a careful observation unearths a number of issues associated with using OCR accuracies as ground truth for document images. For example, OCR accuracies may degrade with factors independent of capture quality like presence of figures/graphs, varying layouts and fonts, spacing between letters, language etc. It fails to address the issue of intra-image variations (i.e. part of the image is blurred, and part is sharp). It is also sensitive to the direction of motion of the hand during capture (equal motion blur in different directions may lead to different OCR accuracies). Additionally, a considerable difference in accuracies is observed while using different OCR algorithms, which makes the metric algorithm dependent.

We propose a solution to this problem by exploiting signal processing concept of SFR, which has been commonly used for measuring the sharpness of a photographic imaging system[1, 2]. SFR is often computed by capturing an image of a slanted edge, which is included in specialized imaging charts. Our idea is to capture four extra slanted edges around a document (as illustrated in Fig. 1), which are then used to compute the sharpness at each spatial position (patch-wise) in the document. This provides an accurate measure of quality (assuming blur as the main degrading factor) which is agnostic to the content of the document (type of font, layout etc.) and also takes into account the intra-document variations. Moreover, creating such ground truth does not require any specialized setup and can be captured in natural settings.

We then train a patch-wise neural network to regress the corresponding SFR values i.e. a network which takes image patch as input and predicts its quality as output. Once the network is trained, testing can be performed without using slanted edges. We perform extensive quantitative and qualitative experiments to demonstrate that using SFR brings significant improvement in predicting the quality of a document image, compared to using OCR as ground truth. The work also exhibits how signal processing concepts can help in creating accurate datasets (which is difficult otherwise), and in turn be used for learning more accurate prediction networks.

2. RELATED WORK

The problem of image quality estimation was initially focused on natural images [3, 4, 5, 6]. Early methods proposed low-level image processing concepts like Just Noticeable Blur [3] and phase coherence [4]. Learning based approaches applying inference on natural scene statistics [6, 5] were then ex-

[*] author contributed when he was a student at IIIT Hyderabad

plored. These algorithms, however, do not generalize well over document images, which led to specific efforts towards DIQA.

Early efforts in DIQA were also based on low-level feature extraction and analysis. Kumar et al. [7] quantified the sharpness of an image by taking a ratio of sharp and non-sharp pixels. Rusinol et al. [8] combined several features like gradient energy, histograms etc. on individual image patches to quantify their quality. A more recent effort exploits the edge profile statistics [9] highlighting the aspect that the transition from text to non-text areas are strong indicators of amount of blur.
Recently, learning based approaches have proven to be more successful for DIQA. The algorithms proposed by Ye et al.[10, 11] extract raw patches randomly from images to build a codebook. The codebook and the features are then used to estimate the quality using a Support Vector Regression. The computational load and inability to handle intra-document variations, however, limits the applicability of this approach.

Recently, deep learning based approaches have shown further improvements. Kang et al. [12] suggested a novel max-min pooling based CNN architecture for patch-wise prediction, using OCR accuracies as ground truth. Since OCR accuracy is computed over the entire document, it fails to address the intra-document variations which adversely affects the training. Recent efforts have been made to address this issue, however, it is limited to the case of out of focus blur [13]. We propose a more generalized approach which benefits from the concept of SFR and overcomes the limitations of either using OCR accuracies or computationally controlled camera setup.

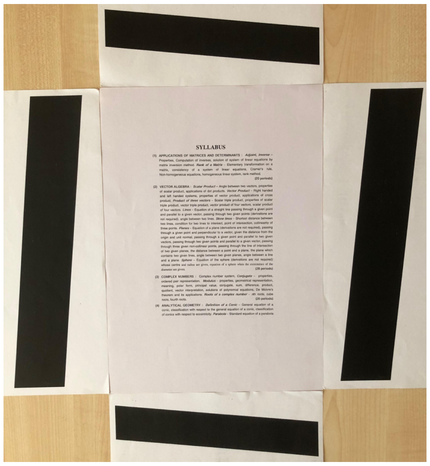
3. SFR AS GROUND TRUTH

The motivation behind proposal of a new dataset is to accurately quantify the quality of an image and assign patch level ground truth to account for intra-image variations. We use the concept of SFR for this task. In this section, we first provide a brief overview of standard SFR calculation technique. We then give a detailed explanation as to how we combine multiple SFR values to generate a ground truth value for patches at varying spatial locations in a document image.

3.1. SFR and its Calculation

An ideal camera, which generates an image of the object without any distortion, must have an 'infinite' resolution in terms of sharpness. Such cameras, obviously do not exist. Therefore, the notion of SFR is used to calculate the 'resolution power' of a camera.

The idea is to measure maximum input frequency the system is able to discern. The ideal 'one-shot' algorithm for this requires an input with all frequencies. This input can be a step-function which has all the frequencies uniformly distributed. The analog for a step-function in the spatial domain is a sharp edge between a dark and a bright region. However, we also want the frequencies from the sensor array beyond it's pitch which is not possible with a vertical edge. Therefore, we utilize the notion of slanted edges [14].



SYLLABUS

(1) APPLICATIONS OF MATRICES AND DETERMINANTS : *Adjoint, Inverse* – Properties, Computation of inverses, solution of system of linear equations by matrix inversion method. *Rank of a Matrix* – Elementary transformation on a matrix, consistency of a system of linear equations, Cramer's rule, Non-homogeneous equations, homogeneous linear system, rank method. (20 periods)

(2) VECTOR ALGEBRA : *Scalar Product* – Angle between two vectors, properties of scalar product, applications of dot products. *Vector Product* – Right handed and left handed systems, properties of vector product, applications of cross product. *Product of three vectors* – Scalar triple product, properties of scalar triple product, vector triple product, vector product of four vectors, scalar product of four vectors. *Lines* – Equation of a straight line passing through a given point and parallel to a given vector, passing through two given points (derivations are not required). angle between two lines. *Skew lines* – Shortest distance between two lines, condition for two lines to intersect, point of intersection, collinearity of three points. *Planes* – Equation of a plane (derivations are not required), passing through a given point and perpendicular to a vector, given the distance from the origin and unit normal, passing through a given point and parallel to two given vectors, passing through two given points and parallel to a given vector, passing through three given non-collinear points, passing through the line of intersection of two given planes, the distance between a point and a plane, the plane which contains two given lines, angle between two given planes, angle between a line and a plane. *Sphere* – Equation of the sphere (derivations are not required) whose centre and radius are given, equation of a sphere when the extremities of the diameter are given. (28 periods)

(3) COMPLEX NUMBERS : Complex number system, *Conjugate* – properties, ordered pair representation. *Modulus* – properties, geometrical representation, meaning, polar form, principal value, conjugate, sum, difference, product, quotient, vector interpretation, solutions of polynomial equations, De Moivre's theorem and its applications. *Roots of a complex number* – *n*th roots, cube roots, fourth roots. (20 periods)

(4) ANALYTICAL GEOMETRY : *Definition of a Conic* – General equation of a conic, classification with respect to the general equation of a conic, classification of conics with respect to eccentricity. *Parabola* – Standard equation of a parabola

Fig. 1. The figure shows a typical image and the slanted edge arrangement used to create the SFR dataset. The slanted edges are placed on all sides of the document. We use four slanted edges to capture the motion and out-of-focus effects robustly. The degradations in slanted edges due to these factors lead to a low SFR score, which in turn results in a lower ground truth value for all the affected patches.

The intensities in an image corresponding to line segments parallel to edges tend to be equal. However, the intensities change as we move away from the zero-crossing of the edge. The slanted edge is beneficial in capturing these intensity variations as the difference is effectively lesser then sensor dimensions. Therefore, if we take Fourier Transform of the intensity values along the gradient axis, we get the spatial frequencies well-above the Nyquist limit of the sensor array. The fraction of frequencies of the image obtained with its DC component gives us the resolution ratio at a particular frequency. The frequency corresponding to a certain pre-defined resolution ratio (typically being equal to 0.5) can now be taken as a measure of the resolution power of a camera assembly.

The conventional use case of SFR is to give the camera a score using a sharp slanted edge. The transition between dark and bright regions, in this case, is abrupt. We propose an alternate use case of the SFR in this work. Our main insight is that if the frequency value corresponding to the pre-defined resolution ratio changes for each image for identical slanted edge input and the same camera, then other factors such as motion or improper focus during the capture process is the reason for this change. Therefore, we conclude that the SFR can be used as a metric representing the image quality, with a higher value corresponding to a sharper image.

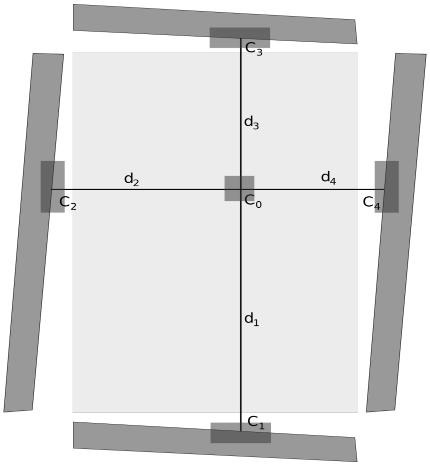


Fig. 2. The figure is an illustrative representation of the capture process. The document is shown in light-gray color, while the slanted edges placed are represented by a darker shade. A selected patch (C_0) and its coresponding RoI's for each slanted edge (C_1, C_2, C_3, C_4) are also visualized.

3.2. SFR based Ground Truth Calculation

We select a set of documents of varying types (in terms of fonts, layout, presence of figures and tables, amount of text present etc). For each document, we place four slanted edges around its sides as seen in Fig. 1 and capture images using a hand-held phone camera. We deliberately shake the hand during capture process to introduce a combination of both out-of-focus and motion blur of varying degrees.

The SFR values around slanted edges are then used to quantify the spatially varying blur in a document image. A score indicating its quality is assigned to each patch using the following procedure (illustrated in (Fig. 2):

1. For every patch in consideration, we move in four directions from the centre of the patch (C_0). The two lines l_1 and l_2 pass through the centre of the patch and are parallel to the directions of the document boundary.

2. The lines l_1 and l_2 intersect the slanted edges at the points C_1, C_2, C_3 and C_4 at distances d_1, d_2, d_3 and d_4. We now consider four Region of Interest (RoI)s at the four slanted edges, centred around the corresponding intersection points.

3. For each RoI, we calculate frequency at which the resolution ratio equals 0.5. We denote this value by s. Thus, we get four scores s_1, s_2, s_3 and s_4 corresponding to the RoIs.

4. We pick two RoIs with minimum scores s_i, s_j for a conservative assignment of ground truth values to the patches, i.e. we consider those values that indicate the highest degradation, which has regularly been used for DIQA [8]. The final ground truth score (g) is a weighted mean of s_i and s_j, weighed according to the distance, as defined in Eqn. 1.

$$g = \frac{\frac{s_i}{d_i} + \frac{s_j}{d_j}}{\frac{1}{d_i} + \frac{1}{d_j}} \tag{1}$$

4. BLUR ESTIMATION PIPELINE

We now provide a brief overview of the pipeline used for the quality estimation in this section. We first use a simple preprocessing step to exclude regions which are not a part of the document using the concept of connected components as in [8]. After this, instead of a simple binarization, which is itself affected by blur, we select patches from the transition regions between textual and non-textual regions [9, 13]. This step avoids selecting non-informative homogeneous patches and helps in achieving scale invariance. The patches of size 48×48 centred are then used to train a CNN based regression network [12] with ground truth being the score calculated for each patch. We modify the back-propagation from Stochastic Gradient Descent (SGD) used in [12] to Adaptive Gradient (AdaGrad) in our implementation for faster and better convergence.

5. EXPERIMENTS

In this section, we present the results obtained for various algorithms against the novel ground truth proposed in this work and the typically used OCR ground truth.

5.1. SFR Dataset Details

The previously proposed datasets [15] were limited in many ways such as lack of inclusion of all types of degradations (both focus and motion blur), images taken in a controlled environment etc. However, the biggest inconsistency affecting performance of various DIQA algorithms is a single and incorrect ground truth in form of OCR accuracies. We have therefore proposed a new dataset which is a true manifestation of the local and global variation of image quality. Our dataset contains 8 images each from 25 documents, i.e. a total of 200 images, varying in blur (motion and focus), scale, orientation etc.

In order to maintain consistency with the previous dataset, we also provide OCR accuracy for each image. The OCR is computed using ABBYY Finereader and ISRI-OCR evaluation tool [16] to compute OCR ground truth for the images.

5.2. Metric Evaluation

We evaluate the predicted score by various algorithms against the two ground truths described above. The quantitative evaluation is traditionally done by computing the Linear Cross

	LCC	SROCC
ΔDOM	0.64	0.65
Focus Measure	0.69	0.80
CORNIA	0.89	0.87
EPM	0.79	0.82
DCNN	0.85	0.82

Table 1. Results of different approaches on SFR Dataset with OCR accuracies as ground truth.

Correlation (LCC) and Spearman Rank Order Cross Correlation (SROCC) between predicted and ground truth values. Additionally, we also present qualitative examples comparing the proposed pipeline trained over SFR based ground truths and state-of-the-art deep networks.

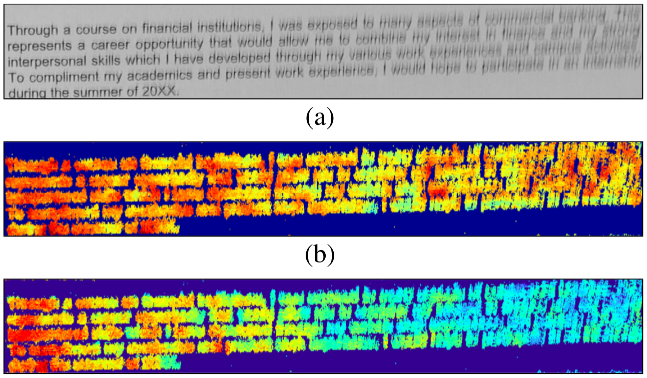
For all the non-learning approaches [7, 8, 9], we evaluate the scores on the entire dataset. For learning based approaches [10, 12], we partition the dataset into training (60%), validation (20%) and testing (20%) sets. The correlations are, therefore, computed only on 20% of the dataset. We run 100 such iterations for testing images, selecting the three sets randomly each time and report the median of all the scores as the final correlation values.

The quantitative results using traditional ground truth of OCR values are presented in Table 1. The learning based approaches clearly outperform the non-learning ones. The correlation values using the SFR generated ground truth values is presented in Table 2. We can see a significant improvement for all the algorithms on using these ground truth values. The results over all non-learning metrics improved by at least 10% for LCC and 7% for SROCC. The scores on learning-based approaches have also increased by 7% on CORNIA and around 10% for DCNN.

We also provide a quantitative measure of local estimation for the CNN based regression network in form of patch-level correlations. We have calculated LCC and SROCC between the predicted and the respective ground truth values. The median LCC and SROCC values of the predicted scores with the SFR based scores comes out to be 0.90 and 0.83 respectively. These values are substantially lower (0.74 and 0.73) when the network is trained using OCR accuracies as ground truth.

5.3. Qualitative Evaluation

In this section, we present a qualitative comparison of the DCNN pipeline [12] with OCR accuracies as ground truth and the proposed pipeline with SFR based ground truths. Fig. 3 shows an image with both focus and motion blurs and corresponding patch-level 'blur maps'. We observe that the DCNN network using OCR accuracies is biased towards a sharper score. Conversely, training the network with the proposed metric as ground truth results in a more accurate blur map. This demonstrates that the proposed ground truth handles intra-document variations more accurately than OCR accuracies. This improved local estimation can be useful for important applications such as selective denoising.



Through a course on financial institutions, I was exposed to many aspects of commercial banking. This
represents a career opportunity that would allow me to combine my interest in finance and my strong
interpersonal skills which I have developed through my various work experiences and campus activities.
To compliment my academics and present work experience, I would hope to participate in an internship
during the summer of 20XX.

(a)

(b)

(c)

Fig. 3. Blur maps obtained using DCNN network and proposed pipeline. Red indicates a high score i.e. high quality while blue indicates a low score, i.e. low quality for a patch. Part (a) shows an image with both motion and focus blur. Part (b) shows the obtained blur map with DCNN network using OCR accuracies as ground truth. Part (c) shows the blur map generated using the pipeline and ground truth values proposed in this work.

	LCC	SROCC
ΔDOM	0.80	0.74
Focus Measure	0.70	0.89
CORNIA	0.96	0.87
EPM	0.94	0.89
Proposed Approach	**0.97**	**0.89**

Table 2. Comparison of different approaches on SFR Dataset with proposed ground truth.

6. CONCLUSIONS

In this work we use the concept of Spatial Frequency Response of a slanted edge to quantify the quality of an image with patch-wise accuracy. This is a step ahead of previous datasets which either use OCR accuracies as ground truth or capture images using computationally controlled setup. We also perform extensive experiments over multiple DIQA algorithms and demonstrate that the proposed ground truth leads to a more accurate training of deep neural networks. In future work, we plan to explore this dataset for the application of deblurring document images. As the current algorithm can generate localized (patch-wise) degradation maps, it can help achieve an adaptive deblurring kernel estimation, in contrast to currently prevalent single kernel estimation approaches.

7. REFERENCES

[1] Stephen E Reichenbach, Stephen K Park, and Ramkumar Narayanswamy, "Characterizing digital image acquisition devices," *Optical Engineering*, vol. 30, no. 2, pp. 170–178, 1991.

[2] Don Williams, "Benchmarking of the iso 12233 slanted-edge spatial frequency response plug-in," in *PICS*, 1998, pp. 133–136.

[3] Rony Ferzli and Lina J Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.

[4] Rania Hassen, Zhou Wang, and Magdy Salama, "No-reference image sharpness assessment based on local phase coherence measurement," in *ICASSP*, 2010.

[5] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[6] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[7] Jayant Kumar, Francine Chen, and David Doermann, "Sharpness estimation for document and scene images," in *ICPR*, 2012.

[8] Marçal Rusiñol, Joseph Chazalon, and Jean-Marc Ogier, "Combining focus measure operators to predict ocr accuracy in mobile-captured document images," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, 2014, pp. 181–185.

[9] Sajal Maheshwari, Pranjal Kumar Rai, Gopal Sharma, and Vineet Gandhi, "Document blur detection using edge profile mining," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*. ACM, 2016, p. 23.

[10] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *CVPR*, 2012.

[11] Xujun Peng, Huaigu Cao, Krishna Subramanian, Rohit Prasad, and Prem Natarajan, "Automated image quality assessment for camera-captured ocr," in *ICIP*, 2011.

[12] Le Kang, Peng Ye, Yi Li, and David Doermann, "A deep learning approach to document image quality assessment," in *ICIP*, 2014.

[13] Pranjal Kumar Rai, Sajal Maheshwari, Ishit Mehta, Parikshit Sakurikar, and Vineet Gandhi, "Beyond ocrs for document blur estimation," 2017.

[14] Peter D Burns, "Slanted-edge mtf for digital camera and scanner analysis," in *Proc. PICS Conf. IS&T*, 2000, p. 135.

[15] Jayant Kumar, Peng Ye, and David Doermann, "A dataset for quality assessment of camera captured document images," in *International Workshop on Camera-Based Document Analysis and Recognition*, 2013, pp. 113–125.

[16] Stephen V Rice, Frank R Jenkins, and Thomas A Nartker, *The fifth annual test of OCR accuracy*, Information Science Research Institute, 1996.