# Beyond OCRs for Document Blur Estimation

by

Pranjal Kumar Rai, Sajal Maheshwari, Ishit Mehta, Parikshit Sakurikar, Vineet Gandhi

in

*14th IAPR International Conference on Document Analysis and Recognition*
(*ICDAR-2017*)

Kyoto, Japan

Report No: IIIT/TR/2017/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2017

# Beyond OCRs for Document Blur Estimation

Pranjal Kumar Rai*, Sajal Maheshwari*, Ishit Mehta, Parikshit Sakurikar and Vineet Gandhi
KCIS, Center for Visual Information Technology, IIIT-Hyderabad
pranjal.kumarrai@research.iiit.ac.in, sajal.maheshwari@students.iiit.ac.in, ishit.mehta@research.iiit.ac.in,
parikshit.sakurikar@research.iiit.ac.in, vgandhi@iiit.ac.in

***Abstract*—The current document blur/quality estimation algorithms rely on the OCR accuracy to measure their success. A sharp document image, however, at times may yield lower OCR accuracy owing to factors independent of blur or quality of capture. The necessity to rely on OCR is mainly due to the difficulty in quantifying the quality otherwise. In this work, we overcome this limitation by proposing a novel dataset for document blur estimation, for which we physically quantify the blur using a capture set-up which computationally varies the focal distance of the camera. We also present a selective search mechanism to improve upon the recently successful patch-based learning approaches (using codebooks or convolutional neural networks). We present a thorough analysis of the improved blur estimation pipeline using correlation with OCR accuracy as well as the actual amount of blur. Our experiments demonstrate that our method outperforms the current state-of-the-art by a significant margin.**

***Keywords*—Image capture; Image quality; Image analysis;**

## I. Introduction

Using off-the-shelf camera-phones for sharing document images has become a common practice in today's digital workflow. This trend is driven by the availability of good quality cameras on smartphones and the ease of sharing at the users' end. The camera captured document images, however, lack controlled environment and stability of capture process in contrast to scanners. This instability often impacts the quality of camera captured images and may lead to failure in document processing algorithms like OCR, which in turn impacts the automation process. An automated camera captured *document image quality assessment* (DIQA) algorithm can resolve this problem by limiting the user intervention only to poorly captured images (filtered automatically) or by providing instantaneous feedback during the capture process itself, to avoid low-quality images.

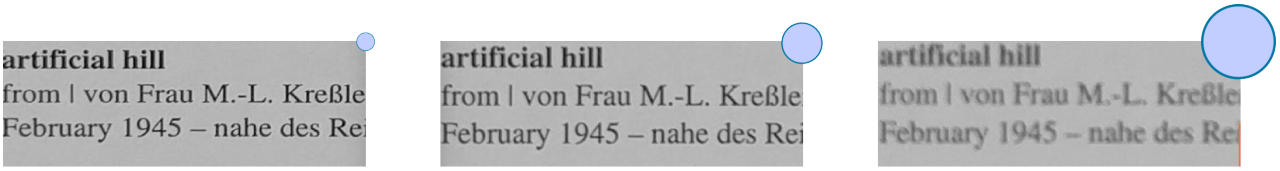
One major challenge for the DIQA problem has been to create ground truth data to physically quantify the quality. Considering the difficulty of this problem, most of the current state-of-the-art DIQA algorithms [1], [2], [3] rely on the obtained OCR values as a reflection of the quality of the document. It is true that there is a correlation between OCR accuracy and the quality of the document. However, this is not always the case (Fig. 2). For instance, multiple factors which are independent of image-quality like font type, language, layout of the document, spacing between the letters, background, presence of tables and figures may adversely effect OCR algorithms. Moreover, many OCR algorithms do some form of quality correction as preprocessing and hence, the output is not appropriate to measure the quality of capture. The results using different OCR algorithms also vary significantly, and therefore, using OCR accuracies may not be a reliable metric to reflect the quality of a document image.

Fig. 1: The capturing set-up used for creating the proposed dataset. The document is fixed on a surface parallel to the camera plane and the focal distance is computationally varied.

In this work, we overcome this limitation by proposing a novel dataset for document blur estimation, where we numerically measure the blur radius using a capture set-up (Fig. 1) which computationally varies the focal plane of the camera. Our approach is similar to focal stacking [4] (with object being fixed to a plane), however with an altogether different goal. The insight here is to decompose the DIQA problem into smaller problems, by looking at individual causes of quality degradations one at a time. We argue that only few types of quality degradations are prevalent in recent context i.e. focus blur, motion blur, skew and uneven illumination, and hence this approach is practically suitable. Our emphasis in this paper is on the aspect of focus blur.

Furthermore, as part of this work we rectify two of the main drawbacks of the current state-of-the-art learning based approaches [1], [2], [3]. The general pipeline of these methods is to divide image into patches and assign the OCR accuracy of the original document to each patch. These patches are then used to train a quality predictor. The testing constitutes select-

*Both authors contributed equally to this work.




OCR : 99.22, r=7.2mm OCR : 99.15, r=16.1mm OCR : 96.60, r=28.9mm

Fig. 2: The figure shows a patch from three different images from a focal stack of a document in the proposed dataset. The OCR accuracies obtained for the document and the computed radius of circle of confusion (blur radius) are written below the patches. The figure clearly shows that the OCR accuracies is not directly proportional to the amount of blur in the document (the blur radius increases four folds with only minor change in OCR accuracies). The circular markers (on the top right of each patch) indicate the size of the circle of confusion for the corresponding images.

ing random patches from the input image, using the trained regressor/network to predict the quality of each individual patch and using the consensus to predict the quality score for the document. The first major problem with this pipeline is to connect the overall document OCR accuracy with local patches. As previous datasets [5] have not been captured in tightly controlled settings, the quality often varies locally in the image itself (for instance some parts are visibly more blurred than other parts). The proposed dataset in this work avoids this problem right away, as the document is fixed on a plane (parallel to image plane) and the blur constantly varies in all parts.

The second problem is that the patch level analysis may not be consistent with varying font sizes, font type and amount of text present in the patch. Choosing the scale/size of patches is also crucial and is often dealt by testing at multiple scales (usually 5-10 discretized sizes), which in turn increases the computational burden. We resolve this problem by proposing a patch selection algorithm which takes into account the factors which are more important for identifying the quality of the image. We leverage upon the recently proposed non-learning based approach [6] for blur estimation, which shows that the transition between the non-text and text regions play a prominent role in estimating the blur quality. We use it as the proposal mechanism for a recently proposed patch based regression network [3]. We show with thorough experiments in Section V, that this patch selection mechanism significantly improves the quality prediction results in contrast to random selection. An analogy of the proposed pipeline can be drawn with recently successful selective search based object detection frameworks in computer vision [7].

Formally, the paper makes the following contributions:

1) We present a novel dataset for document blur estimation, where the ground truth is created by physically estimating the blur radius.
2) We propose a selective search algorithm for extracting the appropriate patches from the document image, which we argue are crucial for estimating the focus blur. We propose an improved pipeline for document blur estimation by combining this selective search algorithm with a recently proposed CNN based regression network [3].
3) We present extensive experiments over two different datasets to validate the proposed pipeline. The results demonstrate that it improves the learning procedure and brings about 4% improvement over the state-of-the-art in estimating the physical blur and over 8-10% improvement in cross dataset experiments using OCR accuracies as ground truth.

**Organization:** The remainder of the paper is structured as follows. First, we briefly review related work in the area of DIQA. Section III enumerates the details about the new proposed dataset. Section IV gives an overall view of the pipeline for blur estimation. Section V presents experimental results and evaluation of the proposed method and finally, the paper is concluded in Section VI.

## II. Related Work

The elementary approaches in the domain of image quality assessment (IQA) focused primarily on natural images. These methods have been analyzed in [8]. This analysis resulted in an enhanced IQA algorithm using the concept of Just Noticeable Blur. A parallel line of work exploited the Natural Scene Statistics for the quantification of the image quality [9], [10]. There have also been numerous approaches focusing on various low-level features in an image added with some post-processing over these features [11], [12]. However, document images are fundamentally different from natural images as they do not have a continuous foreground. Moreover, in a majority of cases, there is no colour constancy in the foreground which makes blur quantification in document images different from that in natural images.

The above stated reasons have led to development of algorithms for image quality assessment specifically for document images. One of the earliest work in this direction was by Blando et al. [13] which banked on low-level features like amount of white speckle, character fragments etc. and predicted the OCR accuracy of an image using these features. An attempt to address a similar problem using gradients was proposed in [14]. Some of the current state-of-the-art approaches for Document Image Quality Assessment (DIQA) based on low-level features also make use of gradients [15], [6] albeit combining the gradient information with other low

level features. The work by Rusinol et al. [15] selects multiple measures like gradient energy, histogram range etc. from a pool of low-level features. The features to be used are decided empirically. After the selection and normalization of the features, the worst performing metric for each image is taken as its image quality. The problem with this approach that the patch-size in this approach cannot change with the amount of text present, font size, etc. Also, some of the selected features are based on thresholds decided empirically which do not generalize well over multiple datasets.

A similar problem arises in the work by Kumar et al. [16]. This approach uses the ratio of number of sharp pixels with the total number of estimated edge pixels. However, the method does not generalize on the rescaling of images along with being susceptible to changes in thresholds for best performance over multiple datasets.

Approaches based on learning such as [17] have proven to surpass the approaches based on low-level features. The algorithms in [1], [2] extract raw patches from the input image randomly from a set of unlabelled images and then learn a dictionary from these patches in an unsupervised fashion. Thereafter, the codebook formed from the learning along with feature vector from a test image is used in a Support Vector Regression model to compute the quality of image. These approaches rely on heavy patch extraction making them unfeasible for images of smaller sizes as well as local quality assessment. Moreover, the computational load of these approaches make them unfit for camera captured images.

Deep learning based approaches have, in recent times, been enormously successful in the domain of feature learning with little or no manual intervention. This has been successful in core Computer Vision problems of object detection, classification [18] etc. Deep learning has therefore, also been explored in the domain of DIQA. Kang et al. [3] recently proposed a deep learning approach which was used for the task of DIQA. The network, however, was observed to be overfitting and therefore, to achieve reasonable generalization, multiple models over different data types had to be saved and searched during testing, making the approach non-viable.

Recently, another non-learning approach was proposed in [6], which argues that the transition region between the textual and non-textual parts in an image is sufficient to estimate the amount of blur, which is an interesting insight. However, the score prediction algorithm proposed in [6] building upon this insight is quite naive and does not completely utilize the information present in the transition region.

## III. Focal Stacks Dataset

The goal of the proposed dataset is to relate each camera captured document image with a blur value, which is computed by considering the optical aspects. In this section, we first explain the relationship between the radius of circle of confusion (which is directly related with amount of blur) and the varying focus distance of the camera. We then describe the capturing set-up configuration and other details of the resulting dataset.

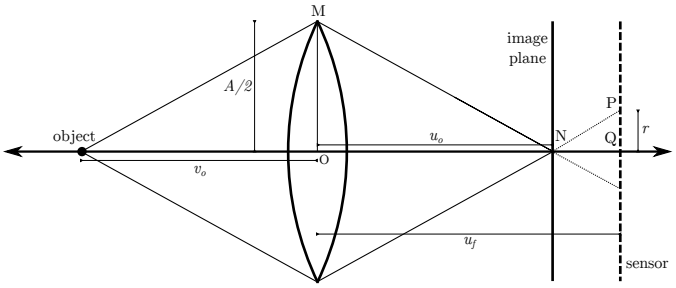



Fig. 3: A single lens camera assembly. The distance between the image plane and the sensor has a direct correlation with the blur in the documents. They are sized according to the focus (not to scale).

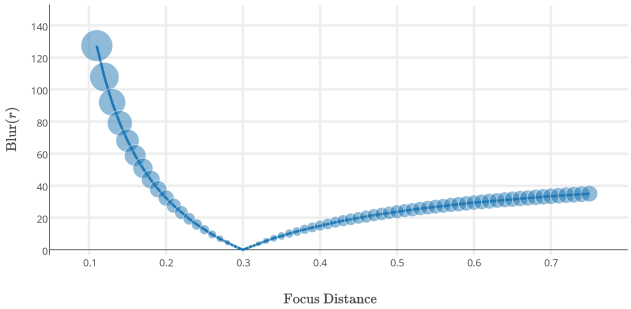



Fig. 4: The plot represents the relationship between the the focus distance and the blur radius. The circular markers represent the circle of confusion for the corresponding focal distances.

### A. The circle of confusion

Ideally, the light rays emerging from a point source intersect in the same point on the sensor after passing through the lens assembly, resulting in a sharp image. However, for an out-of-focus object the emerging rays may not intersect in the same point on the sensor. The area encompassing all the intersection points is known as the circle of confusion. The radius of the circle of confusion ($r$) corresponds to the amount of blur in the image.

Fig. 3 illustrates a standard imaging model of a point source, approximating the camera lens assembly as a single convex lens. The figure shows the ray diagram of a convex lens with focal length $f$ and the object (document) at a distance $v_0$ from the centre of the lens. The corresponding image formed on the image plane lies at a distance $u_0$ and the aperture of the lens is $A$. Initially, when image plane coincides with the sensor of the camera, document is in clear focus. The focus distance $v_f$ can be changed by changing the relative arrangement of the elements in the lens assembly. This is equivalent to the change in the apparent distance between the sensor and the lens. $u_f$ is the distance from the sensor to the lens when the focus distance is $v_f$. Using the law of similar triangles ($\Delta MON$

and $\Delta PQN$) in Fig. 3,

$$r = \frac{A}{2}\Big(\frac{u_f - u_0}{u_0}\Big) \tag{1}$$

According to the thin lens equation,

$$\frac{1}{f} = \frac{1}{u_f} + \frac{1}{v_f} \tag{2}$$

Using Eq. 1 and Eq. 2 we can obtain a relation between the radius of circle of confusion and the focus distance:

$$r = \frac{A}{2}\Big(\frac{f \cdot v_f}{(v_f - f)u_0} - 1\Big) \tag{3}$$

### B. Sampling

By capturing a focal stack with each image slice at a different focus distance, we can obtain a plot of the radius of circle of confusion against the focus distance of the camera. Using a Nexus 5x with a camera focal length ($f$) of 26mm and the object distance $v_o$ as 30cm, the value of $u_o$ turns out to be 2.846cm (using the thin lens equation). Now, by varying the focus distance $v_f$ in Equation 3 on a fixed value of aperture (A), we can obtain its relationship with the radius of circle of confusion.

Fig. 4 shows the plot obtained by using $A = f/2.0$ and by varying $v_f$ from 0.1m to 0.75m with an interval of 0.01m (a total of 66 samples for each document). With focus distance smaller than 0.1m, the details in the images become imperceptible and with focus distance larger than 0.75m, $v_f$ converges to hyperfocal distance. Using the graph, we finally sub sample 10 images (including the image in focus) from the 66 observations for each document, and note the corresponding values of radius of circle of confusion, which is treated as the ground truth for the corresponding image. Therefore, we get a set of alternate ground truths for all the images for this dataset accurately depicting the amount of blur.

## IV. Blur Estimation Pipeline

In this section we introduce the proposed framework for blur estimation. The main idea of the pipeline is to first perform the selective search to locate patches which we argue, precisely and sufficiently indicate the amount of blur present in the document image. We also briefly outline the deep learning framework as proposed in [3], which is used in combination with the patch selection algorithm.

### A. Preprocessing

The document images may include regions which are not a part of the concerned document and are eliminated using a preprocessing step. We use an approach similar to the work by Rusinol et al. [15], which segments out the bounding box of the largest connected component by exploiting the observation that the document occupies the largest portion of the image.

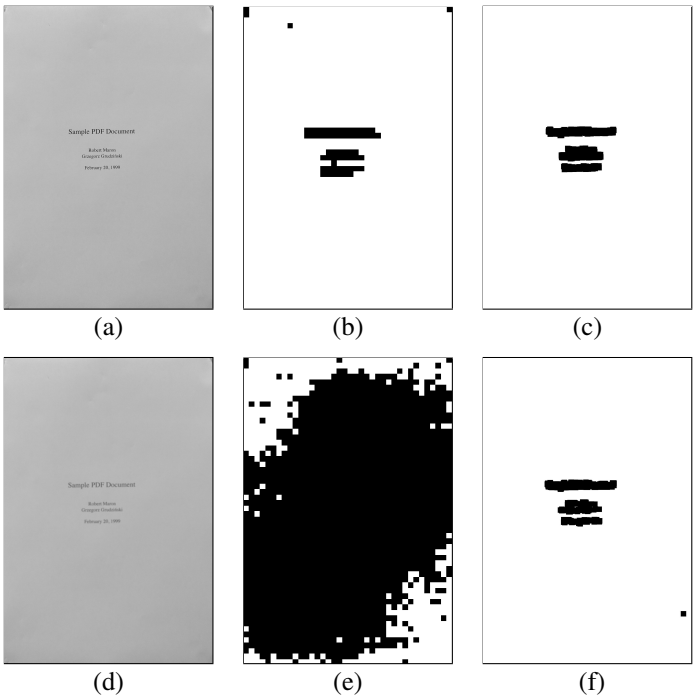



Fig. 5: Comparison of patches selected using the algorithm in DCNN [3] and the proposed approach for a sharp image (top) and a blurred image (bottom). Column 1 shows the two input document images. Column 2 shows the patches selected by the DCNN algorithm as a binary map ( black region shows the patches available for selection). Column 3 shows the binary map of the patches selected by the proposed algorithm. We can observe that the proposed method accurately selects the relevant patches both in the case of sharp and blurred image while the approach in [3] fails in the blurred case.

### B. Patch Selection Algorithm

Most of the recent learning based approaches [2], [3] for DIQA perform patch-level training and testing. The benefits of using a patch based algorithm are two-fold. First, it allows to perform efficient training even with smaller datasets (as several patches can be extracted from each document image). Second, it helps to keep the problem tractable, as learning directly on full size images (typically 4K resolution) is difficult due to practical reasons.

Moving to patch based learning poses the obvious question – which patches from the document image should be selected for the learning procedure? Earlier approaches [2] performed a simple random selection. A more recent approach [3] relies on binarization of the input image and subsequently the localization of textual regions. However, we observe that such binarization algorithms themselves tend to get affected with the amount of the blur in the image (Fig. 5). Furthermore, we argue that a good patch selection algorithm can bring significant improvements to the learning procedure.

We build upon the idea proposed in the recent work by Maheshwari et al. [6] on edge profile mining. They define the

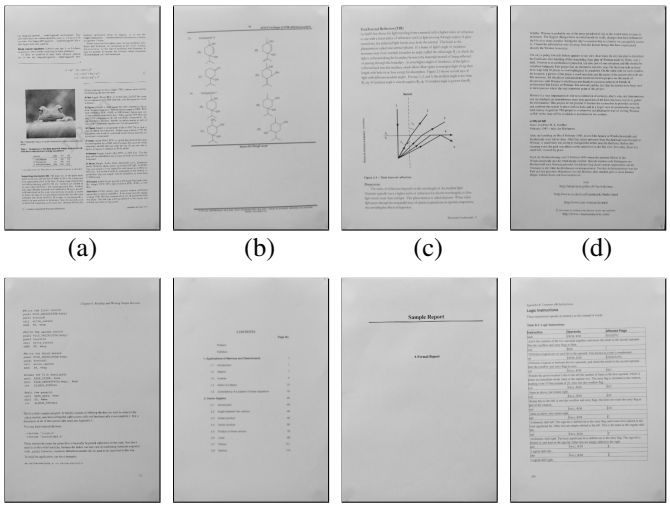


Fig. 6: The figure shows the variations available in the FS dataset, such as images and watermarking (a), layouts of textual regions (b and f), presence of figures and tables (c and h), changes in font types and font sizes (e) and varying amounts of text present in images (d and g).

edge profile mining (EPM) operator, which locates transition between text and background regions for a document image. They argue that the variation along the edge profiles are prominent indicators of the blur in the image (a blur area will tend to have longer transition regions than the sharper ones). We leverage upon this idea for the proposed patch selection algorithm. First, we compute all the edge profiles and remove the extremely short ones corresponding to speckles and other forms of noise. Then we extract a fixed-size patch centred at the mid-point of each edge profile (i.e. the zero-crossing point between the textual and non-textual region) for the learning process. In this work we use patches of size $48 \times 48$ pixels.

Fig. 5 compares the patches chosen by the thresholding algorithm in [3] with the proposed patch selection algorithm. We can observe that the patches are loosely selected from the similar regions in case of a sharp image. However, the exact location is also precisely guided in our case. On the other hand, in case of blurred images, it is evident that our algorithm outperforms the binarization approach.

Furthermore, the proposed patch selection algorithm is agnostic to varying font size, font types and amount of text present in the document. It also eliminates the need to perform testing at multiple scales, as the patches are centred around the zero-crossings, which invariably yields patches encompassing the transition region.

### C. CNN Architecture

The CNN architecture used in this work is the same as the one proposed by Kang et al. [3]. The input to the network are patches of size $(48 \times 48)$ and the output is a single floating point number. It is a regression network and the novelty lies in the use of max-min layers. We replaced the back-propagation algorithm from Stochastic Gradient Descent (SGD) in the original paper to Adaptive Gradient (AdaGrad) in our experiments, which resulted in faster and more consistent convergence of the loss function.

## V. Experimental Results

In this section we present the experimental results obtained using the proposed pipeline benefiting from a selective search mechanism and compare it with current state-of-the-art learning and non-learning based methods. We compare these results using two different metrics as ground truth representing the quality of the image - the conventionally used OCR accuracy and the blur radius/ radius of the circle of confusion.

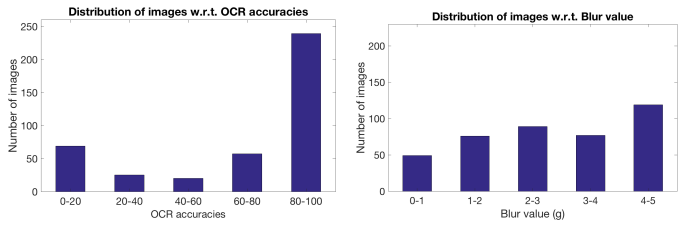
### A. Datasets

*1) SOC Dataset:* The SOC dataset [5] is a publicly available dataset with $175$ images collected over a set of $25$ documents. Each document has been captured $6 - 8$ times with varying amount of blur for each capture. The text in the images is retrieved using three OCR engines (ABBYY Finereader [19], Omnipage [20] and Tesseract [21]). The OCRed outputs are compared with the documentation of each image using the ISRI-OCR Evaluation Tool [22], and the percentage of the characters correctly recognised is considered as a quantitative measure of the image quality.

*2) Focal Stacks Dataset:* The documents in the SOC dataset are limited in terms of variety i.e. they include only few font types, there are no tables or figures/images etc. Moreover, only the OCR accuracies of the images are provided as the ground truth values, which as discussed earlier, may not be a true manifestation of the quality of an image. An additional problem with this dataset is that it has varying amount of blur within an image itself, which results in inconsistency while training, as the same ground truth value (OCR accuracies) is assigned to differently blurred patches. Therefore, we undertook the task of creation of a new generalized dataset covering much more variations (as illustrated in Fig. 6) and with a near uniform distribution of blur values (Fig. 7). Considering that in all images, the documents lie on a plane parallel to the focal plane, the quantity of blur for an image is constant throughout. In addition to the OCR ground truth, we propose an alternate ground truth which captures the amount of the blur considering the actual blur radii as discussed in Section III.

The proposed Focal Stacks (FS) Dataset contains 410 images for 41 different documents, viz. with 10 images of each document with varying amount of blur. In order to maintain consistency with the SOC dataset, we used ABBYY Finereader and Tesseract along with the ISRI-OCR evaluation tool to assign the ground truth values in terms of OCR accuracy.

As stated previously, we have also provided an alternate ground truth for this dataset. We use the blur radii of the images for this version of the ground truth due to the various limitations of having OCR accuracies as the ground truth. We assign each image a score in the range 0-5, with 5 being the




(a) (b)

Fig. 7: The distribution of images w.r.t. OCR accuracies and blur value as the ground truth on FS Dataset. The images in the dataset are selected randomly, therefore we expect uniformly distributed ground truth values. However, as can be seen in (a), OCR accuracies are highly biased towards higher values, with nominal decrement in the OCR accuracies even for significant increase in the amount of blur resulting in a skewed ground truth. The blur value, on the other hand, is distributed nearly uniformly, as it is a physical measure of the blur.

| | LCC | SROCC |
|---|---|---|
| $\Delta$DOM | 0.56 | 0.62 |
| Focus Measure | 0.65 | 0.84 |
| CORNIA | 0.88 | 0.85 |
| DCNN | 0.89 | 0.88 |
| EPM | 0.74 | 0.78 |
| **Proposed Approach** | **0.96** | **0.90** |

TABLE I: Comparison of different approaches on SOC dataset with OCR as ground truth.

sharpest image. This score is computed using the following formula:

$$g = 5\Big(1 - \frac{r}{r_m}\Big) \tag{4}$$

In Eq. 4, $r$ represents the blur radius of the image in consideration and $r_m$ represents the maximum blur radius at which any image is visually perceptible. As can be seen, the sharpest image has a score $g$ of 5 and this score decreases with the increase in the amount of blur.

### *B. Quantitative Evaluation*

Traditionally, the quantitative evaluation of DIQA approaches has been based on the correlation of the predicted quality with the ground-truth (usually the OCR accuracies). We also use the similar evaluation measures for comparing the presented pipeline with the state-of-the-art. Specifically, we use Linear Cross Correlation (LCC) and Spearman Rank Order Cross Correlation (SROCC) metrics. LCC measures the degree of linear dependency between two variables while SROCC is the correlation of rank values of the two variables, assessing the monotonic relationships between them.

| | LCC | SROCC |
|---|---|---|
| $\Delta$DOM | 0.38 | 0.45 |
| Focus Measure | 0.77 | 0.69 |
| CORNIA | 0.87 | 0.81 |
| DCNN | 0.86 | 0.84 |
| EPM | 0.65 | 0.66 |
| **Proposed Approach** | **0.92** | **0.85** |

TABLE II: Comparison of different approaches on Focal Stack dataset with OCR as ground truth.

| | LCC | SROCC |
|---|---|---|
| CORNIA | 0.81 | 0.75 |
| DCNN | 0.76 | 0.80 |
| **Proposed Approach** | **0.91** | **0.80** |

TABLE III: Comparison of different approaches with training on SOC Dataset and testing on Focal Stacks Dataset with OCR as ground truth.

| | LCC | SROCC |
|---|---|---|
| CORNIA | 0.86 | 0.85 |
| DCNN | 0.82 | 0.82 |
| **Proposed Approach** | **0.92** | **0.86** |

TABLE IV: Comparison of different approaches with training on Focal Stacks Dataset and testing on SOC Dataset with OCR as ground truth.

We compare our predicted quality scores with the scores of the following algorithms – $\Delta$DOM [16], Focus Measure (FM) [15], CORNIA [2], Deep Convolutional Neural Network (DCNN) [3] and Edge Profile Mining (EPM) [6]. Two of these approaches (DCNN and CORNIA) rely on learning filters automatically while the rest of the approaches employ hand-crafted low-key features. Firstly, we compare our algorithm with these approaches using the conventional method of treating OCR accuracy as the ground-truth. We present these results with OCR as the ground-truth on both the datasets individually and then present the cross-dataset results (training on one dataset and testing on another). Finally, we also demonstrate the correlation of the various algorithms, including ours, with the alternate ground-truth proposed in this work.

For all the non learning approaches, we calculate the LCC and the SROCC between the predicted scores and the ground-truth scores for the entire dataset (as there is no learning, the testing is performed without any partitioning of the dataset). However, for the learning based approaches, we partition the dataset into training (60%), validation (20%) and testing sets (20%). The division of the dataset is done on the group level rather than on individual images (a group is the set of images corresponding to the same document). We perform testing 100 times with random partitioning of the dataset, compute the correlation measure individually in each case and then report the median value as the final result. For cross-dataset experiments, the pre-trained models from one of the datasets are used to test on the other dataset. The testing in this case is performed on the entire dataset (not just 20%).

The individual dataset experiments for the SOC dataset and the Focal Stacks dataset are presented in Table I and Table II respectively. The comparison with CORNIA was done using the publicly available code whereas the code for DCNN was written by us. Generally the learning based approaches clearly

| | LCC | SROCC |
|---|---|---|
| $\Delta$DOM | 0.62 | 0.69 |
| Focus Measure | 0.74 | 0.86 |
| CORNIA | 0.92 | 0.93 |
| DCNN | 0.91 | 0.92 |
| EPM | 0.89 | 0.90 |
| **Proposed Approach** | **0.96** | **0.96** |

TABLE V: Comparison of different approaches on Focal Stacks Dataset with blur radii levels as ground truth.

seem to outperform the hand crafted ones. The proposed pipeline gives the best results over both the datasets i.e. the addition of selective search brings an improvement of about 7% in LCC measure over DCNN in SOC dataset and about 6% in the case of FS dataset. The SROCC measure improves over DCNN by 8% on SOC dataset, however the improvement is not significant on the FS dataset. In general the SROCC measure decreases over FS dataset for all the approaches, possibly due to the larger number of images in it.

The results for the cross-dataset experiments (using OCR accuracies as ground truth) are demonstrated in Table III and Table IV. Table III presents the results with training on SOC dataset and testing on FS dataset. Conversely, Table IV tabulates the results with training on FS dataset and testing on SOC dataset. These experiments evidently bring out the shortcomings in the training procedure of the previous approaches like DCNN and their lack of generalization. The results improve by 15% for the first case (Table III) and 10% for the second case (Table IV) over DCNN by using the proposed patch selection algorithm. The results demonstrate that choosing the appropriate patches generalizes the learning (irrespective of the dataset used for training, it learns what it is supposed to learn, avoiding irrelevant information). Another interesting observation is that CORNIA also fares better than DCNN on cross dataset experiments, this is possibly because of the underlying clustering of similar patches in it (clustering leads to broad form of patch categorization).

The results with the blur radii as the ground-truth have been presented in Table V. The proposed pipeline outperforms all other approaches in this case as well. Comparing with the case of OCR accuracies, the results of our approach further improve by 4% in LCC and 11% in SROCC measure. The results of all the other approaches also improve while using the new ground truth (compared with the OCR case). The uniform distribution of the blur radius values (Fig. 7) also bridges the gap between the LCC and SROCC values (reducing the bias towards high values while using OCR accuracies as ground truth). The improvement in results (in almost all methods) despite the larger variations in blur radius (compared with the OCR accuracies), indicates that the network is accurately learning the amount of blur present in the document.

## VI. Conclusion

In this paper we have proposed a novel dataset for document image blur estimation. The ground truth values in this dataset correspond to the actual circle of confusion during the capture process (achieved by modeling the underlying optics in a controlled capture process). This goes beyond the previous datasets, which are limited to the use of OCR accuracy as ground truth. We demonstrate that our new dataset leads to more accurate training by reducing ambiguities and bias exhibited while using OCR accuracy as ground truth. Another benefit of the proposed dataset, is that it provides patch level correspondences between the blurred and the sharp document images. These correspondences can be exploited for the deblurring problem, which is left for the future work. Furthermore, we have proposed a selective search algorithm which we demonstrate brings considerable improvements in patch based learning approaches. With the help of thorough experimentation, we show that the proposed pipeline outperforms the state of the art by a significant margin.

## References


[1] P. Ye and D. Doermann, "Learning features for predicting ocr accuracy," in *ICPR*, 2012.

[2] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *CVPR*, 2012.

[3] L. Kang, P. Ye, Y. Li, and D. Doermann, "A deep learning approach to document image quality assessment," in *ICIP*, 2014.

[4] S. F. Ray and W. Gates, "Applied photographic optics," 1994.

[5] J. Kumar, P. Ye, and D. 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.

[6] S. Maheshwari, P. K. Rai, G. Sharma, and V. 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.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[8] R. Ferzli and L. 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.

[9] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[10] A. K. Moorthy and A. C. 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.

[11] R. Hassen, Z. Wang, and M. Salama, "No-reference image sharpness assessment based on local phase coherence measurement," in *ICASSP*, 2010.

[12] J. Shi, L. Xu, and J. Jia, "Discriminative blur detection features," in *CVPR*, 2014.

[13] L. R. Blando, J. Kanai, and T. A. Nartker, "Prediction of ocr accuracy using simple image features," in *ICDAR*, 1995.

[14] X. Peng, H. Cao, K. Subramanian, R. Prasad, and P. Natarajan, "Automated image quality assessment for camera-captured ocr," in *ICIP*, 2011.

[15] M. Rusiñol, J. Chazalon, and J.-M. 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.

[16] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *ICPR*, 2012.

[17] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *CVPR*, 2013.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[19] "Abbyyfinereader," https://www.abbyy.com/finereader/.

[20] "Omnipage," http://www.nuance.com/.

[21] R. Smith, "An overview of the tesseract ocr engine," 2007.

[22] S. V. Rice, F. R. Jenkins, and T. A. Nartker, *The fifth annual test of OCR accuracy*. Information Science Research Institute, 1996.