A Deep Transfer Learning Approach to Document Image Quality Assessment

Tan Lu
Department of Mathematics
Vrije Universiteit Brussel
Brussels, Belgium
Tan.Lu@vub.be

Ann Dooms
Department of Mathematics
Vrije Universiteit Brussel
Brussels, Belgium
Ann.Dooms@vub.be

***Abstract*—Document image quality assessment (DIQA) is an important process for various applications such as optical character recognition (OCR) and document restoration. In this paper we propose a no-reference DIQA model based on a deep convolutional neural network (DCNN), where the rich knowledge of natural scene image characterization of a previously-trained DCNN is exploited towards OCR accuracy oriented document image quality assessment. Following a two-stage deep transfer learning procedure, we fine-tune the knowledge base of the DCNN in the first phase and bring in a task-specific segment consisting of three fully connected (FC) layers in the second phase. Based on the fine-tuned knowledge base, the task-specific segment is trained from scratch to facilitate the application of the transferred knowledge on the new task of document quality assessment. Testing results on a benchmark dataset demonstrate that the proposed model achieves state-of-the-art performance.**

***Keywords*-document image quality assessment, OCR accuracy, deep convolutional neural network, deep transfer learning**

I. Introduction

Despite the prevalence of digital media, paper documents (such as printed books and manuscripts) continue to play an irreplaceable role in modern society. To better preserve these documents, which constitute an important part of our cultural heritage [1], digitization of especially historical documents is widely engaged in heritage sectors such as libraries and archives [2]. The overwhelming volume of document images produced has prompted research in the area of document image processing, where the optical character recognition (OCR) plays a central role. Unfortunately, performance of OCR engines is sensitive to image noise and unproductive results can be obtained when information on the quality of input images is missing from the recognition process [3]. Proper assessment of the quality of document images is therefore beneficial for the OCR process. A similar note could also be drawn in other document processing applications such as document restoration [4] and document image enhancement [5], where document image quality assessment (DIQA) proves to be relevant.

In the field of image quality assessment (IQA), primary attention has been given to the development of general purpose IQA models for natural scene images [6]. However, these models may not be applicable to document images given not only the structural differences between natural scene and document images, but also the deviation in terms of metrics formulation. The IQA of natural scene images is linked to human perceptual scores, while that of document images is mostly related to OCR accuracy [7]. Given that most of the document processing applications are related to the OCR process, in this paper we adopt OCR accuracy as the quality descriptor. A demonstration of the OCR accuracy oriented document image quality assessment is depicted in Fig. 1, where the generic building blocks of a DIQA system are highlighted.

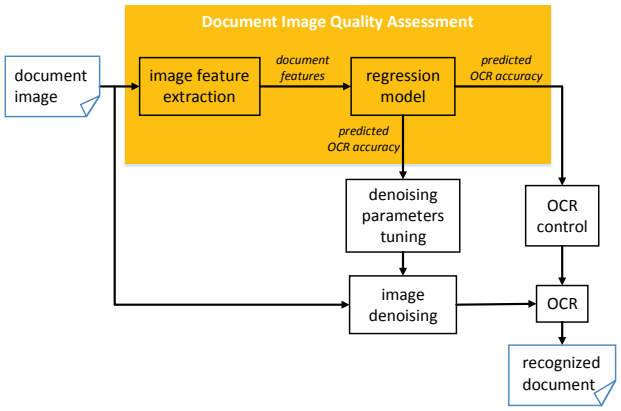

Figure 1. OCR accuracy oriented document image quality assessment.

As shown in Fig. 1, a DIQA system mainly consists of a feature extractor and a regressor. In particular, the extraction of effective document features for OCR accuracy prediction has proven to be a challenging task, as is evident from the diversity of hand-crafted features proposed in the literature [7]. However the progress in the field of convolutional neural networks (CNNs) gives incentives for the development of CNN based DIQA models where the extraction of quality descriptive features is automated through the use of convolutional layers together with back propagation. In this paper we propose a deep CNN (DCNN) based DIQA model using a transfer learning approach, where the rich natural scene image characterization knowledge of a previously-trained DCNN is exploited towards document image quality

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00221

assessment.

The rest of the paper is organized as follows: we briefly review related work in Section 2 and present our approach in Section 3. Experimental results are discussed in Section 4, while Section 5 concludes the paper.

II. Related Work

Divided approaches have been taken in DIQA where both human perceptual scores [1], [8], [9] and OCR accuracy [10], [11], [12], [13], [14], [15], [16] are used as quality metrics. While the perception oriented metrics relate IQA of document images to that of natural scene images, OCR accuracy may be preferred in the general context of document processing, where a large number of applications either rely on or link to the OCR process. Meanwhile, as pristine images are often unavailable in practical scenarios, much effort has been devoted to the development of no-reference (NR) DIQA models where the extraction of descriptive image features is a critical step. Comparing to hand-crafted features [4], [17], learning based feature extraction has attracted more attention.

Ye *et al.* [16] proposed an unsupervised learning framework for feature extraction, where label-free raw image patches are extracted as local features. Based on *K*-means clustering of these local features, a visual codebook is constructed and image level quality descriptors are subsequently crafted for OCR accuracy regression. Peng *et al.* [12] followed a similar bag-of-visual-words (BoW) approach where, instead of using conventional methods such as max-pooling, a latent Dirichlet allocation process together with Gibbs sampling is applied to learn higher level image quality 'topics' from the codebook. The same group of researchers later proposed a semi-supervised learning method for document feature extraction, where a sparse representation model is applied to learn a more discriminative codebook [11]. Xu *et al.* [10] also proposed a NR DIQA model using BoW framework where a comprehensive codebook consisting not only the mean but also the covariance and coskewness of the clusters which are constructed from normalized image patches using *K*-means clustering. The introduction of higher order statistics enabled the formulation of more detailed document quality descriptors.

On the other hand, CNN based DIQA models attracted more attention in recent studies. Kang *et al.* [14] proposed a NR DIQA framework using a 6-layer CNN model, where raw image patches are first extracted and then sifted so that constant patches are removed to avoid confusion in the training or testing process. Quality descriptive features are automatically extracted from the image patches using two convolutional (CONV) layers consisting of 40 and 80 filter kernels respectively. To reduce the dimension of the extracted features, a max-min pooling layer is installed after the CONV layers. The network is terminated with two fully connected (FC) layers with 1024 neurons which are trained to map the extracted features to OCR accuracy. All patches extracted from a document are labeled with the same OCR accuracy of the document, and the OCR prediction of all patches of an image are averaged to produce a single quality score. Li *et al.* [15] proposed an attention based DIQA model where CNN and RNN (recurrent neural network) are integrated to form an interactive glimpse-action mechanism which helps to automatically direct the attention of the model to the text regions of a document. To simulate a retina-like representation, a sequence of glimpse patches with increasing area and decreasing resolution are extracted once a focus point is determined. These glimpse patches are analyzed collectively for quality assessment.

Despite recent attempts, the capability of especially deep convolutional neural networks is yet to be fully exploited for document image quality assessment. However the absence of large datasets with sufficient training samples hinders the development of DCNN based DIQA models. To address this problem, we propose a DCNN based DIQA model using a deep transfer learning approach where, instead of training a DCNN from scratch, the knowledge base of a previously-trained DCNN is exploited in the context of document image quality assessment.

III. A Deep Transfer Learning Approach to Document Image Quality Assessment

Deep convolutional neural networks have proven to be powerful when it comes to elaborated image characterization and the knowledge preserved by a previously-trained DCNN is a valuable asset especially in the context of deep transfer learning where networks are trained to apply their original knowledge on new tasks [18], [19]. We explore in this section the idea of training a DCNN for document image quality assessment while exploiting its knowledge of image feature extraction learned from a large image classification dataset.

A. Model Structure

The problem of NR DIQA can be abstracted as one with domain $\mathcal{D}_{doc}$ and task $\mathcal{T}_{doc}$ [26]. The domain can be characterized as $\mathcal{D}_{doc} = \{\mathcal{X}_{doc}, P(X_{doc})\}$, where $\mathcal{X}_{doc}$ is the feature space for document images, $X_{doc} = \{x_1, x_2, ..., x_m\} \subseteq \mathcal{X}_{doc}$ represents a set of m training samples, and $P(X_{doc})$ is the marginal distribution of the samples. On the other hand the task can be described as $\mathcal{T}_{doc} = \{\mathcal{Y}_{doc}, f_{doc}(\cdot), \tilde{f}_{doc}(\cdot)\}$, where $\mathcal{Y}_{doc}$ is the OCR accuracy space corresponding to the feature space $\mathcal{X}_{doc}$, $f_{doc}(\cdot)$ is a mapping from $\mathcal{X}_{doc}$ to $\mathcal{Y}_{doc}$, and $\tilde{f}_{doc}(\cdot)$ is an approximation of $f_{doc}(\cdot)$. In this paper we follow a transfer learning approach and obtain $\tilde{f}_{doc}(\cdot)$ from $\tilde{f}_{ns}(\cdot)$ which corresponds to a natural scene image classification task.

In particular, we construct a NR DIQA model based on the AlexNet [20] which is a well-known deep convolutional

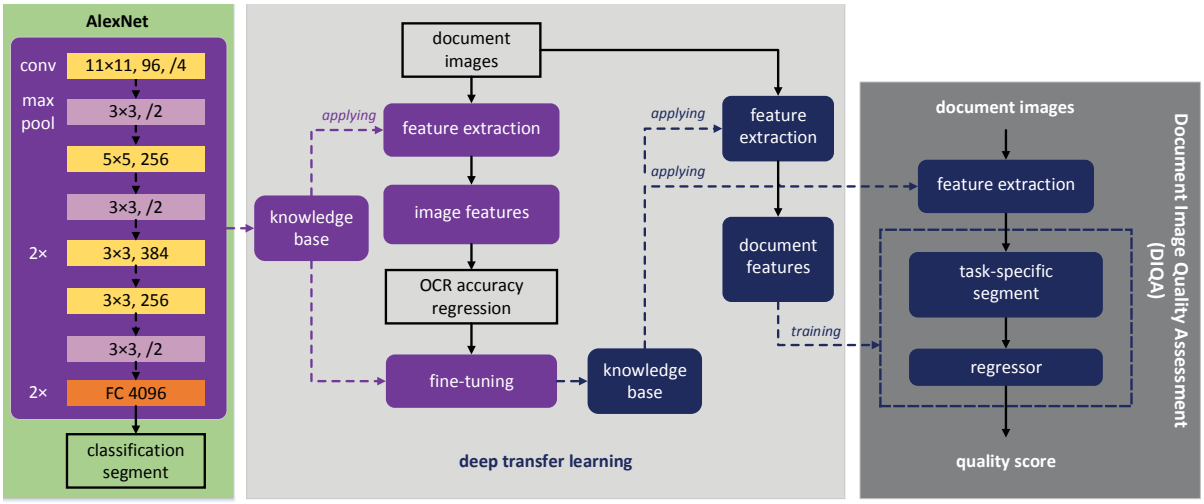

Figure 2. Structure of the proposed DIQA model using deep transfer learning.

neural network. Considering that DCNNs easily get over-fitted especially when the number of learnable parameters of the networks greatly exceeds that of the training samples, we chose AlexNet as the baseline structure such that our DIQA model is less demanding of training samples. Nevertheless, with 23 layers, AlexNet is capable of elaborated image characterization. Specifically, 5 CONV layers containing over a thousand filter kernels are installed in AlexNet for image feature extraction. Meanwhile 3 FC layers are used to map the extracted image features onto categorical outputs. To enhance the non-linearity modeling, rectified linear units (ReLUs) are inserted in between these CONV and FC layers. Having been trained on the very large ImageNet dataset [20], the learnable (especially the CONV) layers of AlexNet have preserved rich knowledge of extracting general features for image description.

To exploit the image characterization knowledge of AlexNet towards document quality assessment, we re-configure AlexNet for regression, where the classification segment consists of the last FC layer with 1000 neurons (FC-1000), the softmax layer and the classifier is replaced with a regression segment consisting of one FC-1 layer and a L_2-loss based regressor. Under this new configuration we fine-tune the kernel knowledge of the original AlexNet in the context of DIQA, where the learnable layers are re-trained to extract image quality descriptors for OCR accuracy regression.

After the fine-tuning process, a task-specific segment containing mainly three FC layers is introduced after the FC-4096 layer to facilitate the application of the transferred knowledge base of the network on the new task of quality assessment. This task-specific segment is trained from scratch to learn a mapping from the quality sensitive features extracted using the fine-tuned knowledge base to document OCR accuracy, while the knowledge base itself is frozen (i.e. excluded from training) in this second training phase. The structure of the proposed DIQA model using deep transfer learning is demonstrated in Fig. 2 while the task-specific segment is depicted in Fig. 3.

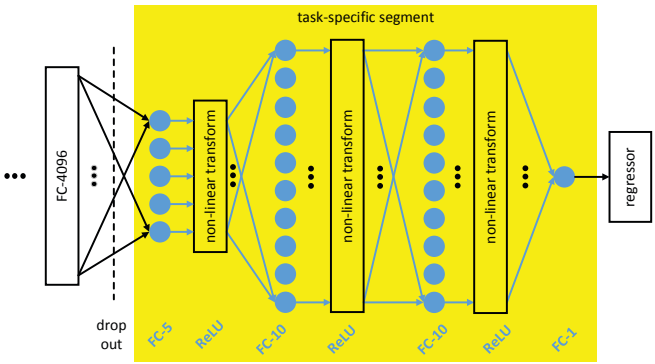

Figure 3. Task-specific segment of the DIQA model.

B. Learning Procedure

To adapt to the AlexNet configuration, images are first divided into small patches before being fed through the network. All patches extracted from an image are labeled with the same OCR accuracy of the image. The predicted quality scores of all patches of an image are averaged to obtain a single OCR accuracy prediction. Take note that null patches (a.k.a patches with constant pixel values) are excluded from the entire training-testing process as they do not contain useful information for OCR quality assessment. To sift out these null patches, we follow the same procedure proposed in [14].

When fine-tuning the knowledge base, a L_2-loss based regressor is employed where the loss function is formulated as:

$$\mathcal{L}(p_i, y_i, w) = \frac{1}{2N}\sum_{i=1}^{N}(f(p_i) - y_i)^2 + \lambda\Omega(w)$$

where N is the number of patches, $f(p_i)$ is the network response corresponding to patch p_i, y_i is the target score and $\Omega(w)$ is the L_2 regularization of the weights w of the network. A regularization coefficient λ is used to adjust the severity of the penalization on large weights. A relatively large regularization coefficient of 0.0005 is imposed during the training process to alleviate the overfitting problem. Stochastic gradient descent with adaptive momentum estimation is adopted to minimize the cost function during the training. For each experiment the model is trained on 60 epochs. Within one epoch, the model parameters are updated by following a minimum batch scheme where the size of a minimum batch is set to 128.

A similar procedure is followed when training the task-specific segment, except that smaller regularization coefficient (0.0001) and minimum batch size (64) are used as the number of trainable parameters contained in the task-specific segment is much smaller than that of the original network.

IV. Experiments

We test the performance of the proposed DIQA model on the Sharpness-OCR-Correlation (SOC) dataset [21] which is frequently used as the benchmark dataset for document quality assessment. Performance of the proposed model is being compared with the state-of-the-art in this section.

A. Dataset and Protocol

The SOC dataset contains 175 document images with resolution of 1840×3264. These images are grouped into 25 sets, where each set contains 6 to 8 pictures taken from the same document. By varying the focal distance of the capture camera, different levels of focal-blur were introduced to the images when creating the dataset. Three different types of OCR engines (namely ABBYY FineReader, Tesseract and Omnipage) were engaged during the OCR accuracy evaluation process, where FineReader demonstrated the best performance [21]. Therefore in our experiment the OCR accuracy obtained based on the FineReader output is used as the quality score for all images. To increase the number of training samples, we introduce an overlap between neighboring patches during the image partitioning process. Specifically the patch extraction stride is set to 113 while each patch has a dimension of 227×227.

During the experiments, 80% samples are used for training while the remaining 20% samples are used for testing. This random split is conducted on the set level so that images taken from the same document are not shared across training and testing. Following the convention, the performance of the proposed DIQA model is measured by calculating the Pearson's linear correlation coefficient (PLCC) and the Spearman's rank correlation coefficient (SRCC) between the predicted and the target OCR accuracy of the testing images. The experiment is repeated 10 times while the PLCC and SRCC performances are recorded at each time. We report the median of these values in the following discussions.

Table I
Performance of learning based DIQA models

DIQA Model	PLCC	SRCC
CORNIA [22]	0.937	0.862
CNN [14]	0.950	0.898
LDA [12]	-	0.913
HOS [10]	0.960	0.909
Sparse Model [11]	0.935	0.928
RNN [15]	0.956	0.916
proposed method	**0.965**	**0.931**

B. Results Discussion

Much effort has been devoted to the development of learning based DIQA models. We obtain encouraging results as can be seen in Table I, where we compare the proposed model to six other learning based approaches proposed in recent studies. Take note that the performance of CORNIA on the SOC dataset was reported in [14], where the authors compared CORNIA, which was originally developed for natural scene IQA, to a CNN based DIQA model.

Our method outperforms the other learning based models under both PLCC and SRCC criteria. The PLCC performance of HOS [10] is competitive, however the relation between the predicted and target OCR accuracy may not be well described in a monotonic way in their model, as indicated by the discrepancy between their PLCC and SRCC scores. On the other hand, the sparse model [11] demonstrates strong SRCC performance, but their PLCC score is lower than the other learning based approaches.

We attribute the performance of our model to the exploitation of the rich knowledge base of the original DCNN, which has been trained extensively for elaborated image characterization. To investigate the effect of the fine-tuning of the knowledge kernels, we compare in Fig. 4 a middle section (filter kernels 46 to 51) of the first CONV layer of the original DCNN, to its corresponding section of the transferred DCNN. The transfer kernels $T_i(46 \leq i \leq 51)$ depicted in the middle of Fig. 4 measure the normalized absolute differences between the original and fine-tuned kernels:

$$T_i = \frac{|X_i' - X_i|}{\beta_i}$$

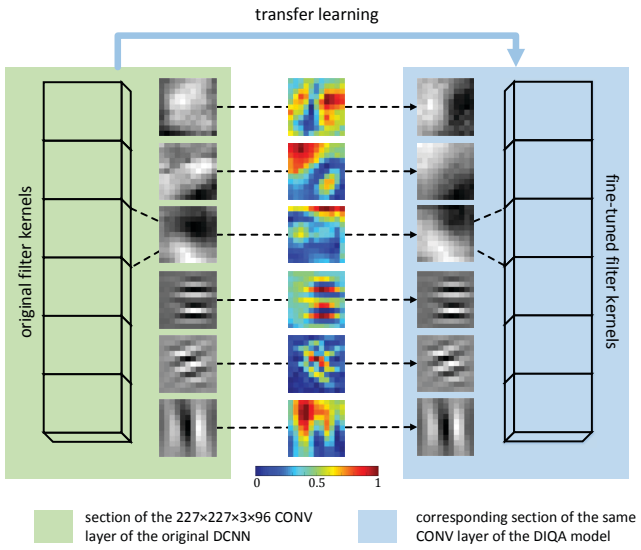

Figure 4. Comparison between original and transferred filter kernels.

where $X_i' \in R^{227\times227\times3}$ and $X_i \in R^{227\times227\times3}$ represent the original filter kernel trained for image classification and the transferred filter kernel learned for quality assessment respectively. The denominator β_i is incorporated for normalization such that values in T_i are within the range of 0 to 1.

As can be observed from Fig. 4, some filter kernels received major structural changes, while others have preserved their original kernel patterns during the fine-tuning process in the particular performance of PLCC = 0.9809 and SRCC = 0.9467 within the repetition. This observation suggests that the knowledge base of a previously-trained classification oriented DCNN can be effectively engaged under the paradigm of deep transfer learning for the new task of document quality assessment.

We also compare our method to metric based DIQA models which employ hand-crafted features, as shown in Table II.

Document-wise evaluation is commonly employed in metric based DIQA models, where PLCC and SRCC are calculated independently for each document such that a total of 25 PLCCs/SRCCs are obtained on the SOC dataset. The median of these 25 evaluations is then used as the performance indicator. In consideration of the bias towards good results among the experiments in document-wise evaluation [23], [24], metric based models were often evaluated by following a general protocol where all documents are considered altogether for PLCC and SRCC calculation. Take note that the document-wise performance of CORNIA was reported in [21] where a 25-fold cross-validation scheme was used such that 24 sets of images are used for training and 1 set of images are used for testing during the experiments. We follow the same procedure to evaluate the performance of our model under the document-wise scenario.

Table II
PERFORMANCE OF METRIC BASED DIQA MODELS

DIQA Model	Document-wise		General	
	PLCC	SRCC	PLCC	SRCC
CORNIA [22]	0.9747	0.9286	0.9370	0.8620
Focus [23]	0.9378	**0.9643**	0.6467	-
MetricNR [24]	0.9750	0.9107	0.8867	0.8207
CG-DIQA [25]	0.9523	0.9429	0.9063	0.8565
proposed method	**0.9763**	0.9550	**0.9651**	**0.9312**

As can be observed from Table II, learning based methods outperform state-of-the-art metric based models under the general evaluation criteria where our method yields the best results under both PLCC and SRCC measurements. However it should be noted that normally only a subset of the entire image dataset is used when testing learning based models. The performance of the proposed model is still, under both PLCC and SRCC measurements, competitive to the other methods using document-wise evaluation, where the proposed model achieves best and second-best scores under PLCC and SRCC respectively.

V. CONCLUSION

This paper proposes a no-reference document image quality assessment model based on a deep convolutional neural network, where the knowledge of natural scene image characterization captured within a previously-trained DCNN is exploited for document quality assessment using a two-staged deep transfer learning approach. In particular, the knowledge base of the original network is fine-tuned in the first stage and a task-specific segment is introduced and trained from scratch using the transferred knowledge base in the second phase. Testing results on a benchmark dataset demonstrate that the proposed model yields competitive performance when compared to the state-of-the-art learning and metric-based models. The promising performance of the proposed model not only motivates future exploration of DCNN based DIQA models using transfer learning, but also encourages future study on the generalization across natural scene and document images in the area of image quality assessment.

ACKNOWLEDGMENT

This research is supported by the Auditing Digitisation Outputs in the Cultural Heritage Sector (ADOCHS) project (Contract No. BR/154/A6/ADOCHS), financed by the Belgian Science Policy (Belspo) within the scope of the BRAIN programme.

REFERENCES

[1] A. Shahkolaei, H. Z. Nafchi, S. Al-Maadeed, and M. Cheriet, "Subjective and objective quality assessment of degraded document images," *J. Cult. Herit.*, vol. 30, pp. 199–209, 2018.

[2] A. Antonacopoulos and A.C. Downton, "Special issue on the analysis of historical documents," *Int. J. Doc. Anal. Recognit.*, vol. 9, pp. 75–77, 2007.

[3] D. Berchmans and S. S. Kumar "Optical character recognition: an overview and an insight," in *Proceedings of International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 1361–1365, 2014.

[4] M. Cannon, J. Hochberg, and P. Kelly, "Quality assessment and restoration of typewritten document," *Int. J. Doc. Anal. Recognit.*, vol. 2, pp. 80–89, 1999.

[5] R. Garg and S. Chaudhury, "Automatic selection of parameters for document image enhancement using image quality assessment," in *Proceedings of 12th IAPR Workshop on Document Analysis Systems*, pp. 422–427, 2016.

[6] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction," *IEEE Signal Processing Magazine*, pp. 130–141, 2017.

[7] P. Ye and D. Doermann, "Document image quality assessment - a brief survey," in *Proceedings of 12th International Conference on Document Analysis and Recognition*, pp. 723–727, 2013.

[8] A. Alaei, D. Conte, M. Blumenstein, and R. Raveaux, "Document image quality assessment based on texture similarity index," in *Proceedings of 12th IAPR Workshop on Document Analysis Systems*, pp. 132–137, 2016.

[9] A. Alaei, D. Conte, and R. Raveaux, "Document image quality assessment based on improved gradient magnitude similarity deviation," in *Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 176–180, 2015.

[10] J. Xu, P. Ye, Q. Li, Y. Liu, and D. Doermann, "No-reference document image quality assessment based on high order image statistics," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 3289–3293, 2016.

[11] X. Peng, H. Cao, and P. Natarajan, "Document image quality assessment using discriminative sparse representation," in *Proceedings of 12th IAPR Workshop on Document Analysis Systems*, pp. 227–232, 2016.

[12] X. Peng, H. Cao, and P. Natarajan, "Document image OCR accuracy prediction via latent Dirichlet allocation," in *Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 771–775, 2015.

[13] X. Peng, H. Cao, K. Subramanian, R. Prasad, and P. Natarajan, "Automated image quality assessment for camera-captured OCR," in *Proceedings of 18th IEEE International Conference on Image Processing*, pp. 2621–2624, 2011.

[14] L. Kang, P. Ye, Y. Li, and D. Doermann, "A deep learning approach to document image quality assessment," in *Proceedings of IEEE International Conference on Image Processing*, pp. 2570–2574, 2014.

[15] P. Li, L. Peng, J. Cai, X. Ding, and S. Ge, "Attention based RNN model for document image quality assessment," in *Proceedings of 2017 14th International Conference on Document Analysis and Recognition*, pp. 819–825, 2017.

[16] P. Ye and D. Doermann, "Learning features for predicting OCR accuracy," in *Proceedings of 21st International Conference on Pattern Recognition (ICPR)*, pp. 3204–3207, 2012.

[17] A. Souza, M. Cheriet, S. Naoi, and C. Y. Suen, "Automatic filter selection using image quality assessment," in *Proceedings. of The Seventh International Conference on Document Analysis and Recognition (ICDAR)*, pp. 508–512, 2003.

[18] E.R.S. de Rezende, G.C.S. Ruppert, A. Theóphilo, and E.K. Tokuda, "Exposing computer generated images by using deep convolutional neural networks," *Signal Processing: Image Communication*, vol. 66, pp. 113–126, 2018.

[19] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full Training or Fine Tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no.5, pp. 1299–1312, 2016.

[20] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems Conference*, pp. 1097–1105, 2012.

[21] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in *Proceedings of International Workshop on Camera-Based Document Analysis and Recognition*, pp. 39–44, 2013.

[22] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1098–1105, 2012.

[23] M. Rusiñol, J. Chazalon, and J.-M. Ogier, "Combining focus measure operators to predict OCR accuracy in mobile-captured document images," in *Proceedings of 2014 11th IAPR Workshop on Document Analysis Systems*, pp. 181–185, 2014.

[24] N. Nayef and J.-M. Ogier, "Metric-based no-reference quality assessment of heterogeneous document images," in *Proceedings of SPIE 9402 Document Recognition and Retrieval (DRR) XXII*, 2015.

[25] H. Li, F. Zhu, and J. Qiu, "CG-DIQA: no-reference document image quality assessment based on character gradient," in *Proceedings of 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3622–3626, 2018.

[26] S. J. Pan, Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Engineer*, vol. 22, no.10, pp. 1345–1359, 2010