A Document Image Quality Assessment Method Based on Feature Fusion

Weisheng Wang, Zhiyang Yan, and Hongli Lin

Abstract Document image quality assessment (DIQA) is an essential step in the development of optical character recognition (OCR) products. Due to the complex and diverse distortion types in the real captured document images, DIQA is still a challenging problem. In this paper, we propose a new DIQA model, which is based on the feature fusion in convolutional neural network (CNN). In our network, shallow network part is used to extract low-level local features of document images to represent local non-uniform distortions. And deep network part is used to learn global features to represent global uniform distortions in document images. In addition, a quality regression network is used to predict the document image quality score by using the fusion of the low-level and deep-level features. Experimental results demonstrate that our model outperforms the state-of-the-art methods on complex distortion datasets.

Keywords Document image quality assessment · Document image · DIQA · Feature fusion

1 Introduction

As the popularity of smart devices grows, document image recognition is not just for traditional scanned text, but more for real document images captured by smart device cameras. In recent years, many Internet companies have developed document image recognition services, of which OCR services occupy the mainstream position. The performance of the OCR engine is closely related to the quality of the document image, however, due to the defects of the shooting equipment or photography skills, the document image will be distorted during the capture process, resulting in different degrees of image quality problems [18] and lower OCR accuracy. In this case, the important information in the document image is recognized incorrectly or lost, causing immeasurable costs. Therefore, it makes sense to apply DIQA before

W. Wang · Z. Yan · H. Lin (✉)
College of Computer Science and Electronic Engineering, Hunan University, Hunan, China
e-mail: hllin@hnu.edu.cn

J. Yao et al. (eds.), *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, Lecture Notes in Electrical Engineering 813,
https://doi.org/10.1007/978-981-16-6963-7_78

recognizing the document image, so as to remove the low-quality image, or to further restore [15] or enhance [16] the document image.

Generally, the field of image quality assessment is divided into the quality assessment of natural scene images and document images. In recent years, the field of natural scene image quality assessment has developed rapidly [17]. However, there are not only significant difference between document images and natural scene images in terms of structure and measurement formulas [6], but the goals of the two are also very different. Unlike natural images quality assessment, which can be evaluated based on human perception, DIQA can be evaluated based on OCR accuracy. Consequently, the natural scene image quality evaluation model may not be directly applied to document images [18].

In the past few years, many people have made great efforts to assess the quality of document images through different methods. Although some progresses have been achieved, there are still huge challenges to the evaluation of document image quality with complex multiple distortions. There are various types of document image distortions, and multiple distortions may be concentrated on one image in unexpected ways [13]. Because of the diversity of distortion types, different document images have different types of distortion, but they may have similar OCR accuracy, as shown in Fig. 1. Nevertheless, In the existing document image quality assessment methods, whether based on traditional manual features [5, 7, 14], or learning based document image quality assessment methods [4, 8, 9, 11], they either only tend to extract low-level or global features of the image, ignoring deep-seated features and local features, yet, in the document image captured by smart device camera, there may be global distortion caused by defocus or illumination, or local distortion caused by lens jitter or shadow. As a result, the algorithms which only tend to extract low level features and the algorithms which only tend to learn global features with deep models still have not worked well. Therefore, aggregating both local distortion features and global distortion features, and then predicting document image quality upon this multi-scale representation is an efficient approach.

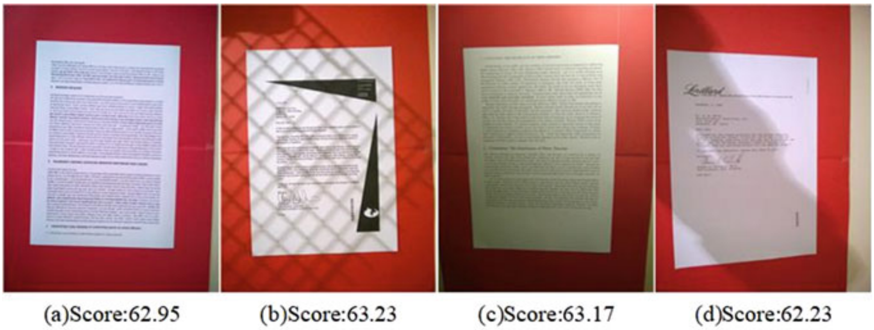

Fig. 1 **a**, **b**, **c** and **d** are four document images with different degrees of distortion on the SmartDoc-QA [13] data set. Different distortion features are mapped to similar OCR accuracy in these four images

In this paper, we develop a DIQA model for the complex distortion of real document images. We extract low-level local features and deep-level global features from multi-scale feature maps, which are fused to deal with the distortion of diversity.

Our model uses two network structures to extract low-level and deep-level features, and then merges local distortions features which are captured by a local feature extractor with global quality features. A final quality score is predicted through a quality regression network, which is trained by fusing low-level and deep-level features. Conducted a series of experiments demonstrate that our model achieves significant effects on complex datasets in terms of DIQA and precedes the latest DIQA methods [4, 8, 9, 11] reported in the literature.

The following chapters of this paper will describe the related work, our specific methods, the analysis of experimental results and the summary of this paper.

2 Related Work

In the process of OCR, the distorted document image may lose key information, resulting in incorrect recognition results. As a consequence, it is very significant to add DIQA in OCR process. Given that DIQA is bound up with OCR accuracy, OCR accuracy is adopted as the quality descriptor in most DIQA methods. The current latest DIQA methods are usually divided into two categories: metric-based methods and learning-based methods.

2.1 The Metric-Based DIQA Methods

The metric-based DIQA method generally extracts different manual features to generate a quality map to the quality score of the document image. Kumar et al. [5] used the grayscale change of image after median filtering to calculate the sharpness information to assess the image quality. In [14], Nayef et al. developed an DIQA method based on OCR accuracy. This method calculates the quality score through proportional weighted summation based on the dependence between different distortions of the document image and combined with a specific distortion measure. In Kumar et al. [7], the quality score is calculated by character gradient. However, these techniques focus only on the specific characteristics of the image, and the effect is not obvious for document images with complex and diverse distortion types.

2.2 The Learning-Based DIQA Methods

The learning-based document image quality assessment model generally includes two steps: feature extraction and quality score regression. In recent years, Kang et al.

[4] proposed a CNN model to assess document image quality. Li et al. [9] implemented an attention-based recurrent neural network (RNN) for DIQA. Different from the traditional DIQA method based on OCR accuracy, this framework integrates CNN and RNN to form a glimpse-RNN-Action combined network. Lu et al. [11] applied the deep transfer learning method to DIQA and put forward a deep CNN model. In [8], Li et al. proposed a DIQA framework where the overall quality score was weighted by the quality score of each text block. Nevertheless, when the document image quality is low, these methods will lose some key text information or text lines that affect the accuracy of OCR, resulting in inaccurate quality prediction.

Although these methods have made great progress, these methods are more inclined to extract the low-level features of the text area, ignoring the deeper and complex semantic features contained in the uniform distorted document image. In practical applications, a truly distorted document image may be merged by multiple distortions in a complicated manner, and low-level features cannot fully represent the document image with diverse distortions. In addition, diverse distortions may exist locally or globally, and the sensitivity of the OCR engine is determined by these two conditions. In this paper, inspired by [17], we built a new DIQA framework. Our framework combines low-level features and deep features while fusing local non-uniform distortions and global uniform distortions, and the results are obtained through a quality regression module. The experimental results prove that our quality evaluation model advantage over ones reported in the literature.

3 Method

In this section, we develop a DIQA network, and the network architecture is shown in Fig. 2, including shallow feature extraction module, deep feature extraction module and quality regression module. Each module is described in detail below.

3.1 Shallow Feature Extraction Module

The main distortion problems of document images include: illumination, blur, scene background, stains, and color degradation, resolution, etc., resulting in image quality problems [13]. Therefore, we try to keep the quality information of the original image in the low-level feature extraction stage, and perform preliminary extraction of the quality information. In this part, we are inspired by [12] and combine 3 × 3 convolution and generalized divisive normalization (GDN) [2] as the backbone, where GDN is highly non-linear [1] and has spatial adaptability. In order to better capture the local distortion information, we use 1 × 1 convolution and global average pooling (GAP) to convert multi-scale features into local feature vectors. It was proved in [17] that this structure can be considered as an attention-based local feature extractor,

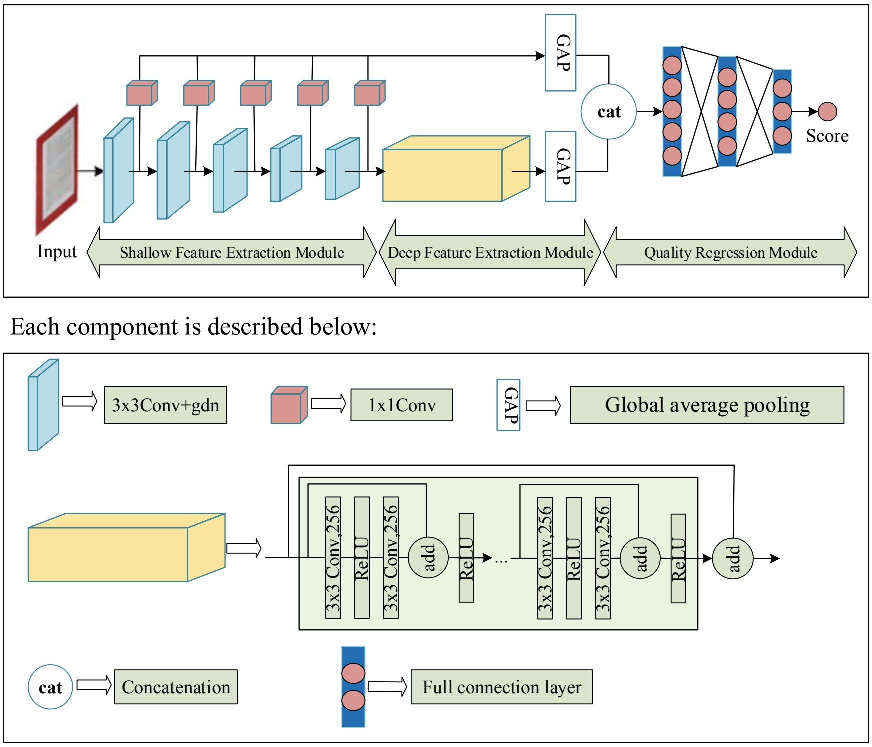

Fig. 2 The DIQA model framework proposed in this paper contains three modules: shallow feature extraction module, deep feature extraction module and quality regression module

which can perceive the regional features corresponding to the local distortion, so as to better capture its quality.

3.2 Deep Feature Extraction Module

Because of the diversity and complexity of document image distortion types, it is necessary to extract deeper quality features from document images. For fear of preventing the degradation caused by the increase of the network depth, we use the structure shown in Fig. 3 to increase the network depth. In Fig. 3c, we removed the batch normalization (BN) layer in ResNet [3], which not only reduces network consumption, but also increases the flexibility of the network. Among them, we set the size of the convolution kernel to 3 and the number of channels to 256. After removal, we can stack more network layers, and each layer can extract more features [10]. We extract deep semantic features by stacking 18 structures in Fig. 3c, and merge the shallow semantic features extracted in Sect. 3.1 with the output of the

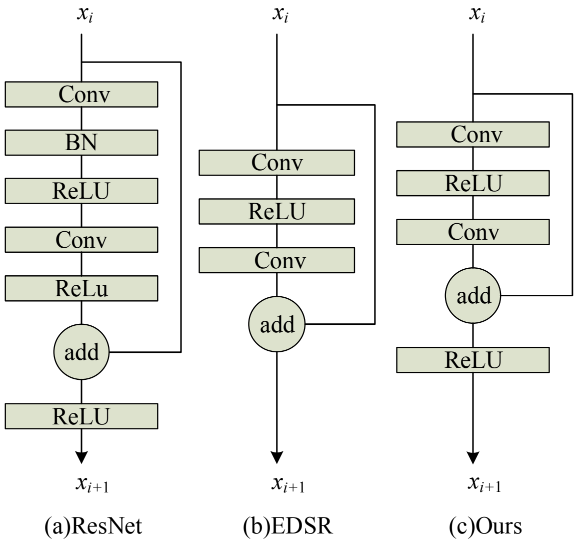

Fig. 3 Comparison of the residual structure of ResNet [3], EDSR [10] and this paper

deep semantic feature module. Finally, GAP is also fused with local features to feed the quality regression network.

3.3 Quality Regression Module

In this part, our goal is to map the previously extracted image features to the quality score, so we build a quality regression network with three fully connected layers. As shown in Fig. 2, the multi-scale feature fusion vector is used as input and propagated through three fully connected layers where the ReLU function serves as the activation function, and finally the document image quality score is obtained. Our model can be described as

$$s = \varphi(L(x), D(x), \gamma) \tag{1}$$

where φ represents the network model, x and $L(x)$ are the input image and the output of the shallow feature extraction network respectively, $D(x)$ is recorded as the result of the deep feature extraction network, and γ is the model parameter.

3.4 Implementation Details

In our experiment, the dataset is spilt into a training set and a test set in a ratio of 8–2, and set the batch size to 16. Adam optimizer is selected to optimize the prediction network, where the learning rate and the betas are set to 1e-4 and (0.9, 0.999) respectively, meanwhile, eps is adjusted to 1e-8, and weight decay is 0. In the training phase, we randomly cut the input image into 224 × 224 × 3 to form 16 patches, and the ground truth of each patch is the same as the input image. For the entire training process, the loss function is l_1-norm.

$$\ell = \left\| s - \hat{s} \right\|_1 = \sum_{i=1}^{n} \left| s_i - \hat{s}_i \right| \tag{2}$$

where s_i and $\hat{s}_i$ represent the ground truth and predicted quality score of the i-th patch, n is the total number of patches. In the testing phase, the sample is also randomly divided into 16 patches, and the quality scores of the 16 patches are averaged to obtain the final quality score.

4 Experiment

Our model is evaluated on two public datasets Sharpness-OCR-Correlation (SOC) [6] and SmartDoc-QA [13] and compared them with the state-of-the-art approaches.

4.1 Datasets and Evaluation Metrics

The SOC dataset is made up of 175 document images with a resolution of 1840 × 3264 and is composed of 25 English documents taken with a smartphone, and each takes 6–8 images with different focal lengths to produce varying degrees of distortion. The SOC dataset uses three OCR engines (ABBY FineReader, Tesseract, and Omnipage) to evaluate the OCR accuracy of each image. In our experiments, we use the mean results of the three OCR engines as the ground truth. The SmartDoc-QA dataset is a more complex data set with more distortion types. The dataset contains 4260 document images, which were taken from 30 documents by two different mobile phones. The 30 document images are mainly composed of three types of official documents, old official documents and receipts. The OCR accuracy of this dataset is the recognition results of the FineReader and Tesseract OCR engines. Similarly, we calculate the mean of the two OCR recognition results as the ground truth.

We choose two conventional evaluation indicators: Spearman Rank Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC) to evaluate the performance of the model. SROCC indicates the monotony of the predicted results and is defined as

$$SROCC = 1 - \frac{6\sum_i^n d_i^2}{n\left(n^2 - 1\right)} \tag{3}$$

where d_i is the rank difference between the prediction result of the i-th test image and the ground truth, n is the number of test set. PLCC is commonly used to describe the accuracy of prediction results and is defined as

$$PLCC = \frac{\sum_i^n (s_i - s_m)\left(\hat{s}_i - \hat{s}_m\right)}{\sqrt{\sum_i^n (s_i - s_m)^2 \sum_i^n \left(\hat{s}_i - \hat{s}_m\right)^2}} \tag{4}$$

where s_i and $\hat{s}_i$ are the ground truth and prediction result of the i-th test image separately,s_m and $\hat{s}_m$ are the mean values of all ground truth and predictions, n is the number of test images. The larger the value of these two indicators, the better the performance, and the range is between 0 and 1.

4.2 Comparison with the State-Of-The-Art Methods

We have compared seven latest DIQA methods, including three metric-based methods: Sharpness [5], MetricNR [14] and CG-DIQA [7], four learning-based methods: CNN [4], RNN [9], TL [11] and DTL [8]. As shown in Table 1, on the SOC dataset, our method is significantly better than other methods. Our PLCC results are

Table 1 Comparison of PLCC and SROCC results on SOC and SmartDoc-QA datasets with the latest methods

Methods	SOC		SmartDoc-QA	
	PLCC	SROCC	PLCC	SROCC
Sharpness [5]	N/A	N/A	0.624	0.596
MetricNR [14]	0.887	0.820	N/A	N/A
CG-DIQA [7]	0.906	0.856	0.625	0.631
CNN [4]	0.950	0.898	N/A	N/A
RNN [9]	0.956	0.916	0.814	**0.865**
TL [11]	0.914	0.872	0.743	0.757
DTL [8]	0.965	0.931	N/A	N/A
Ours	**0.991**	**0.968**	**0.956**	0.854

The bold values represent the optimal results of all the DIQA methods that were compared leading in the other four methods and the SROCC are only slightly lower than RNN [9] for SmartDoc-QA dataset. Among them, the result of DTL on the data set SOC is better than the other methods. Smartdoc-QA dataset is more complex and has more types of distortion. our method performs on this dataset is slightly lower than that on the SOC dataset. This is because 40% of the 2160 document images scored by the Tesseract OCR engine on the Smartdoc-QA dataset have a result of 0%, which means that the OCR accuracy distribution of this dataset is unbalanced. From the results of PLCC, our method is still superior to the four most advanced methods on Smartdoc-QA dataset. In the results of SROCC, the attention mechanism-based RNN model [9] is better than the other four approaches, and also slightly exceed our method, which shows the attention-based RNN model [9] has better results on the monotonicity of prediction. In addition, our method is greatly superior to the other three methods in SROCC results. From the discussion above, it is obvious that our approach has an excellent performance for DIQA.

4.3 Ablation Study

We performed ablation experiments on the SOC datasets and the SmartDoc-QA dataset to assess the effectiveness of each components in our DIQA framework. We first proved the effectiveness of low-level feature extraction network (LC) and deep- level feature extraction network (DC). The results are shown in Table 2. Both indicators are superior to all current technologies on SOC datasets, and PLCC results are significantly superior to other methods on SmartDoc-QA dataset. Then we verify the effectiveness of the local distortion feature extraction module (MS). When LC is added to the local distortion feature extraction module, LC improves on both datasets. It is significantly improved by 1.3% on the SmartDoc-QA dataset in SROCC. And when we Combining LC, MS and DC, our model has been further improved in SROCC and PLCC, which reached 96.8% and 99.1% on the SOC dataset, and 85.4% and 95.6% on the SmartDoc-QA dataset.

Table 2 Results of ablation experiment on SOC and SmartDoc-QA datasets

Components	SOC		SmartDoc-QA	
	PLCC	SROCC	PLCC	SROCC
LC	0.985	0.964	0.944	0.835
LC + MS	0.986	0.967	0.946	0.848
DC	0.969	0.955	0.952	0.837
LC + MS + DC	**0.991**	**0.968**	**0.956**	**0.854**

5 Conclusion

This paper proposes a new CNN model based on feature fusion to evaluate document image quality. Our model takes account of the diverse, local and global distortions of real document images by feature fusion, rather than the distortions in single aspect. In order to better predict the quality of real distorted document images, our shallow feature extraction module extracts low-level quality information and local distortion features, and try to preserve the original quality of image. Then we use the deep feature extraction module to acquire the high-level information of the distortion features, and finally combine the two features while fusing the local distortion features with the global semantics, and feed them to the quality regression module to get the final quality score. The experimental results prove that our model shows strong robustness to both simple distortion and complex multiple distortion document images.

In addition, this method explores the DIQA method through feature fusion, and also provides a prospect for multiple distortion document image quality evaluation in the field of document image quality evaluation in the future.

References

1. Ball´e, J., Laparra, V., Simoncelli, E.P.: End-to-End Optimized Image Compression. ArXiv abs/1611.01704 (2017)
2. Ballé, J., Laparra, V., Simoncelli, E.P.: Density modeling of images using a generalized normalization transformation. Int. Conf. Learning Represent. (2015)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90, https://doi.org/10.1109/CVPR..90
4. Kang, L., Ye, P., Li, Y., Doermann, D.: A deep learning approach to document image quality assessment. In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 2570–2574 (2014)
5. Kumar, J., Chen, F., Doermann, D.: Sharpness estimation for document and scene images. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 3292–3295 (2012)
6. Kumar, J., Ye, P., Doermann, D.S.: A dataset for quality assessment of cam-era captured document images. In: Iwamura, M., Shafait, F. (eds.) Camera-Based Document Analysis and Recognition - 5th International Workshop, CB-DAR 2013, Washington, DC, USA, August 23, 2013, Revised Selected Papers. Lecture Notes in Computer Science, vol. 8357, pp. 113–125. Springer (2013)
7. Li, H., Zhu, F., Qiu, J.: CG-DIQA: No-reference Document Image Quality Assessment Based on Character Gradient. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3622–3626 (2018).
8. Li, H., Zhu, F., Qiu, J.: Towards Document Image Quality Assessment: A Text Line Based Framework and a Synthetic Text Line Image Dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 551–558 (2019)
9. Li, P., Peng, L., Cai, J., Ding, X., Ge, S.: Attention based RNN Model for Document Image Quality Assessment. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 819–825 (2017)

10. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
11. Lu, T., Dooms, A.: A Deep Transfer Learning Approach to Document Image Quality Assessment. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1372–1377 (2019)
12. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-End blind image quality assessment using deep neural networks. IEEE Trans. Image Process. **27**(3), 1202–1213 (2018)
13. Nayef, N., Luqman, M.M., Prum, S., Eskenazi, S., Chazalon, J., Ogier, J.: SmartDoc-QA: A dataset for quality assessment of smartphone captured document images - single and multiple distortions. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1231–1235 (2015)
14. Nayef, N., Ogier, J.: Metric-based no-reference quality assessment of heterogeneous document images. In: Ringger, E.K., Lamiroy, B. (eds.) Document Recognition and Retrieval XXII, San Francisco, California, USA, February 11–12, 2015. SPIE Proceedings, vol. 9402, pp. 94020L. SPIE (2015)
15. Ouafek, N., Kholladi, M.: A binarization method for degraded document image using artificial neural network and interpolation inpainting. In: 2018 4th International Conference on Optimization and Applications (ICOA). pp. 1–5 (2018)
16. Sharma, P., Sharma, S.: An analysis of vision based techniques for quality assessment and enhancement of camera captured document images. In: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence). pp. 425–428 (2016)
17. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3664–3673 (2020)
18. Ye, P., Doermann, D.: Document Image Quality Assessment: A Brief Survey. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 723–727 (Aug 2013)