NO-REFERENCE DOCUMENT IMAGE QUALITY ASSESSMENT BASED ON HIGH ORDER IMAGE STATISTICS

Jingtao Xu[1], *Peng Ye*[2], *Qiaohong Li*[3], *Yong Liu*[1] *and David Doermann*[4]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Airbnb, San Francisco, CA, USA
[3]Nanyang Technological University, Singapore
[4]University of Maryland, College Park, MD, USA

ABSTRACT

Document image quality assessment (DIQA) aims to predict the visual quality of degraded document images. Although the definition of "visual quality" can change based on the specific applications, in this paper, we use OCR accuracy as a metric for quality and develop a novel no-reference DIQA method based on high order image statistics for OCR accuracy prediction. The proposed method consists of three steps. First, normalized local image patches are extracted with regular grid and a comprehensive document image codebook is constructed by K-means clustering. Second, local features are softly assigned to several nearest codewords, and the direct differences between high order statistics of local features and codewords are calculated as global quality aware features. Finally, support vector regression (SVR) is utilized to learn the mapping between extracted image features and OCR accuracies. Experimental results on two document image databases show that the proposed method can accurately predict OCR accuracy and outperforms previous algorithms.

Index Terms— Document image quality assessment, high order statistics, OCR accuracy, no-reference

1. INTRODUCTION

Previous image quality assessment (IQA) research has mainly focused on natural scene images. In addition to natural scene images, other types of images, such as document images [1], screen content images, medical images and so on, also play important roles in our daily life. However, algorithms developed for natural scene images may not be directly applied to other types of images. In this paper, a novel no-reference document image quality assessment (DIQA) method is proposed. The quality of natural scene images is usually evaluated with respect to human perception. However, for DIQA, based on the specific applications, the definition of quality can be different. We may evaluate document image quality with respect to both human or machine [2]. We define the document image quality as its OCR accuracy in this paper. Under this problem formulation, our algorithm can be used for automatic OCR accuracy prediction. Since there is no reference document image under most real-world scenarios, DIQA algorithms are generally no-reference. A good DIQA algorithm can be utilized in many applications, such as document image selection for OCR software and parameter optimization for document analysis systems.

Distortions in document images can be classified into two categories according to the level at which they affect the document. The first one is character level distortion, such as touching characters, broken characters, noise around characters and so on. Early DIQA research has focused on this type of distortion and attempted to design handcrafted features for specific distortions. Several quality aware factors for typewritten document images are investigated in [3–5]. These quality aware factors include: Font Size (FS), Stroke Thickness (ST), Small Speckle Factor (SSF), Touching Character Factor (TCF), White Speckle Factor (WSF) and Broken Character Factor (BCF). These factors are typically designed based on connected components and have been used to predict OCR accuracy. These character related features rely on the font size of document images, however, document images may contain different font sizes of characters in practice.

The second one is page level distortion, such as out of focus blur, global noise from camera sensors, contrast change and so on. Several methods are proposed to evaluate specific distortion in document images. Kumar *et al.* [6] propose a sharpness metric to evaluate the blurriness in camera-captured images. Rusiñol *et al.* [7] combine several simple focus measure operators to evaluate document image blur. However, these distortion specific methods cannot work with all types of distortions. Fortunately, feature learning relevant approaches have been successfully introduced to build general purpose DIQA method. Ye *et al.* [8] propose the first general purpose method for DIQA which is an extension of CORNIA method [9]. However, when the codebook size decreases, the performance drops significantly, and the 10K codeword codebook limit its computation speed. Later they improve this model with supervised filter learning [10]. With a 100 codeword codebook, the performance is acceptable but still

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

inferior to [8]. Kang *et al.* [11] establish a convolution neural network (CNN) based method to combine feature learning and regression procedures. The training stage for CNN is quite time consuming. All the feature learning based methods only employ 0-th order statistics and overlook image high order statistics which actually benefit OCR accuracy prediction.

In this paper, high (1-st, 2-nd and 3-rd) order image statistics are incorporated to build efficient quality aware feature for DIQA. Our method is referred as HOS-DIQA and it differs from previous codebook based DIQA algorithms in the following ways. First, a more comprehensive document image codebook is generated by K-means clustering including not only the mean of each cluster but also the dimension wise variance and skew. Second, high order statistics differences between local features and codewords are calculated as global aware quality feature. With high order statistics, we are able to achieve high prediction accuracy using a very compact codebook with 100 codewords, which runs much faster and requires much less training time. It also outperforms previous DIQA methods on two document image databases.

The remainder of this paper is organized as follows. Section 2 describes HOS-DIQA model. Experimental results and discussions are presented in Section 3 and Section 4 concludes the paper.

2. METHOD

We first introduce how local features are extracted and used to form the codebook with high order statistics information. We then describe how the codebook is used to generate powerful features. Finally, the mapping between features and OCR accuracies are learned by a regression model.

2.1. Local Feature Extraction

In the proposed method, normalized raw patches are extracted from document images as local features. Given a document image, the local feature $\boldsymbol{x}(i,j)$ is extracted from $B \times B$ patch $\boldsymbol{I}(i,j)$, where i, j are spatial indices sampled on a regular grid. The local contrast normalization scheme which has been widely used in IQA domain [9,12] is applied to each patch as follows:

$$\boldsymbol{x}(i,j) = \frac{\boldsymbol{I}(i,j) - \mu}{\sigma + 10}, \tag{1}$$

where μ and σ are the local mean and standard deviation of patch $\boldsymbol{I}(i,j)$, and the constant 10 prevents instability when the denominator tends to zero. Finally N local features are extracted for each image: $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \boldsymbol{R}^D$ ($D = B \times B$).

2.2. Codebook Construction

Unlike codebooks generated in [8,9,11] which only contain mean values of codewords, we build a more comprehensive codebook from normalized image patches. In addition to the mean of each cluster provided by traditional K-means clustering [13], we further calculate the covariance ($\boldsymbol{\sigma}^2$) and coskewness (γ) for each cluster. These matrices are assumed diagonal as the computational cost of diagonal matrices is much lower than the cost involved for full matrices. Therefore, the dimension wise variance and skew are provided. For the kth cluster, two additional D-dimensional vectors, $\boldsymbol{\sigma}_k^2$ and $\boldsymbol{\gamma}_k$, are calculated. Finally the generated codebook is described by a set of parameters $\lambda = \{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2, \boldsymbol{\gamma}_k, k = 1, 2, \ldots, K\}$. Each cluster represents one codeword.

Since document images are usually grayscale and contain some flat patches which are useless for codebook construction, we employ Otsu's binarization [14] on raw images to select image patches containing text. The constant patches (all ones or all zeros) on the binary map are discarded. 219 text zone images which contain less than 30 characters from grayscale newspaper images [15] are used to build the codebook. The $\boldsymbol{\mu}$ of 100 codewords from K-means clustering is shown in Fig. 1. We observe that the constructed 100 codewords (mean) contain various types of patterns which are similar to different parts of strokes. Patches with different levels and types of distortions will have distinctive "distances" to these codewords.

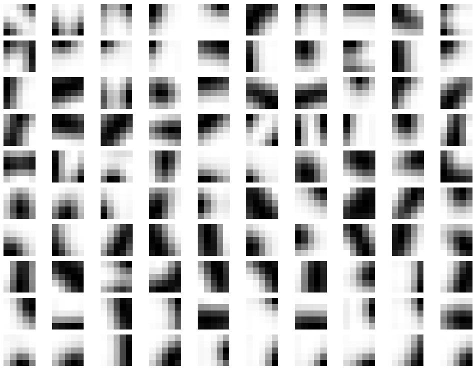


Fig. 1. Mean values of 100 codewords

2.3. High Order Statistics based Feature

Previous feature learning based DIQA methods [8,10] consider only 0-th order statistics from images, i.e., only codeword counting statistics is employed to represent the relationship between local features and the simple codebook. By contrast, we use high order statistics differences between local features and clusters to depict the relationship more comprehensively with a much smaller codebook. This relationship describes the approximate location of local image features in each cluster (relatively to the mean, variance and skewness). Different clusters represent diverse stroke patterns with different distortion levels, therefore this relative relationship will vary with image quality (OCR accuracy).

For every local feature $\boldsymbol{x}_i$, the r nearest codewords $rNN(\boldsymbol{x}_i)$ are selected and a Gaussian kernel weight ω_{ik} is regraded as the "distance" from local feature $\boldsymbol{x}_i$ to codeword k. First we calculate the differences between the soft weighted mean of local features assigned to cluster k and the mean of cluster k:

$$\begin{cases} \boldsymbol{m}_k^d = \hat{\boldsymbol{\mu}}_k^d - \boldsymbol{\mu}_k^d = \sum\limits_{i:k\in rNN(\boldsymbol{x_i})} \left[\omega_{ik}\boldsymbol{x}_i^d\right] - \boldsymbol{\mu}_k^d \\ \omega_{ik} = \dfrac{e^{-\beta\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2}}{\sum_{j:k\in rNN(\boldsymbol{x_j})} e^{-\beta\|\boldsymbol{x}_j-\boldsymbol{\mu}_k\|^2}} \end{cases} \tag{2}$$

where $\hat{\boldsymbol{\mu}}_k$ is the mean of the local features assigned to codeword k, the superscript d denotes the d-th dimension of a vector and ω_{ik} is the Gaussian kernel similarity weight between local feature $\boldsymbol{x}_i$ and codeword k. The sum of the weights for each codeword is 1.

Then we formulate the 2-nd order statistics as follows:

$$\boldsymbol{v}_k^d = \hat{\boldsymbol{\sigma}}_k^{2\,d} - \boldsymbol{\sigma}_k^{2\,d} = \sum_{i:k\in rNN(\boldsymbol{x_i})} \left[\omega_{ik}(\boldsymbol{x}_i^d - \hat{\boldsymbol{\mu}}_k^d)^2\right] - \boldsymbol{\sigma}_k^{2\,d}, \tag{3}$$

where $\hat{\boldsymbol{\sigma}}_k^2$ is the dimension wise variance of local features assigned to codeword k. Therefore $\hat{\boldsymbol{\sigma}}_k^{2\,d}$ is the variance of dth dimension in cluster k.

For a standard Gaussian distribution, the 1-st and 2-nd order statistics are sufficient. However, low level image features do not usually follow the Gaussian distribution in practice [16]. Therefore, we also employ the 3-rd order statistics to exploit complementary information for quality evaluation. The 3-rd order statistics can be written as follows:

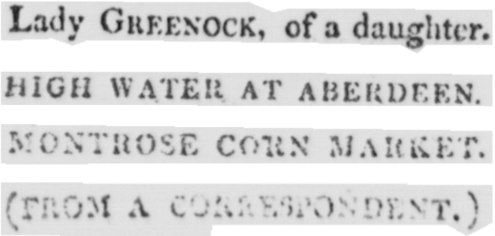
$$\boldsymbol{s}_k^d = \hat{\boldsymbol{\gamma}}_k^d - \boldsymbol{\gamma}_k^d = \sum_{i:k\in rNN(\boldsymbol{x_i})} \left[\frac{\omega_{ik}(\boldsymbol{x}_i^d - \hat{\boldsymbol{\mu}}_k^d)^3}{(\hat{\boldsymbol{\sigma}}_k^{2\,d})^{\frac{3}{2}}}\right] - \boldsymbol{\gamma}_k^d, \tag{4}$$

where $\hat{\boldsymbol{\gamma}}_k$ is the dimension wise skewness of local features assigned to codeword k. Thus $\hat{\boldsymbol{\gamma}}_k^d$ is the skew of dth dimension for cluster k.

Finally all the three types of statistics differences are concatenated to a single long quality aware feature $\boldsymbol{V} = [\boldsymbol{m}_k^\top, \boldsymbol{v}_k^\top, \boldsymbol{s}_k^\top], k = 1, 2, \ldots, K$. The new image quality aware representation contains 1-st, 2-nd and 3-rd order statistics (c.f. equations (2)-(4)). With a given codebook of size K, full quality aware representation provides a vector of dimensionality $3DK$.

2.4. Regression Model

Given quality aware feature vectors, we need to learn a mapping from the feature space to OCR accuracy. We use support vector regression (SVR) to learn this mapping. Since the feature dimensionality is high, we use linear SVR [17].



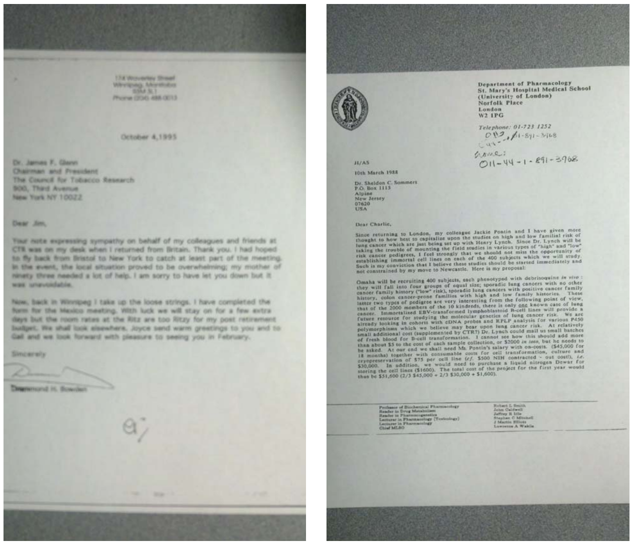
Lady GREENOCK, of a daughter.

HIGH WATER AT ABERDEEN.

MONTROSE CORN MARKET.

(FROM A CORRESPONDENT.)

Fig. 2. Sample images from Newspaper database for codebook construction



(a)

(b)

Fig. 3. Sample images from SOC database, OCR accuracy: (a) 0.1286, (b) 0.9905

3. EXPERIMENTS

3.1. Database and Protocol

We test competing methods on two document image databases.

(1) Newspaper database: This database contains 521 grayscale text zone images with different resolutions. These images are a subset of historical document images with machine printed English and Greek. And every image has more than 30 characters. A commercial OCR software (ABBYY Fine Reader) is used to obtain OCR results and ISRI-OCR evaluation tool [18] is used to generate OCR accuracies in the range $[0, 1]$. This database contains character level distortions. Some sample images for codebook construction are shown in Fig. 2.

(2) Sharpness-OCR-Correlation(SOC) database [19]: It contains camera-captured document images with blur. A 8 mega-pixel cell phone camera was used to generate blurred versions of 25 non-distorted document images. 6-8 photos with varying levels of blurriness were taken for every im-

age. Finally a total of 175 color images with fixed resolution (3264×1840) were created. The OCR results were obtained by a commercial OCR software (ABBYY Fine Reader) and OCR accuracies were also generated by the ISRI-OCR evaluation tool. Some sample images are shown in Fig. 3.

There are some parameters of HOS-DIQA need to be set. To do this, 10000 5×5 patches are extracted via a regular grid for each image. The codebook size K is set to 100, 5 nearest codewords are selected for each local feature, and the parameter β in the Gaussian kernel weight function is set to 0.05.

We compare HOS-DIQA with four general purpose DIQA methods, BRISQUE-L [12], CORNIA [9], SFL [10] and CNN [11]. Two commonly used criteria, Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) are employed to evaluate the performance of competing DIQA methods. Before calculating PLCC, a logistic nonlinear fitting procedure is applied to map the algorithm scores to OCR accuracies. Each database is randomly split into two non-overlapping subsets 100 times. 80% of images are used for training and the remaining 20% for testing. The median values of SROCC and PLCC over 100 train-test loops are reported.

3.2. Evaluation

We evaluate all competing methods on two databases and show the results in Table 1. HOS-DIQA achieves better SROCC and PLCC on both databases. It is not surprising that all methods obtain a higher performance on the SOC database. Images with global distortions, such as blurriness and noiseness, are included, and these degradations have obvious influence on statistics of low level image features. Conversely, the distortions in the Newspaper database are primarily localized broken strokes. Thus all methods show a decrease of performance on this database. However, with a comprehensive codebook containing high order statistics of various parts of strokes, not only global distortions but also character level distortions can be well addressed. Therefore HOS-DIQA produces relatively better result on this database.

Table 1. Results on Newspaper and SOC Databases

Methods	Newspaper		SOC	
	SROCC	PLCC	SROCC	PLCC
BRISQUE-L [12]	0.709	0.722	0.836	0.904
CORNIA [9]	0.725	0.751	0.862	0.937
SFL [10]	0.708	0.735	0.854	0.927
CNN [11]	0.726	0.731	0.898	0.950
HOS-DIQA	**0.766**	**0.800**	**0.909**	**0.960**

In order to further evaluate the proposed method, we conduct the following experiments. First, the performance variation with codebook size K is presented in Fig. 4. It can be observed that even with only 25 codewords, HOS-DIQA still performs relatively well on both databases. Second, we show the boxplot of SROCC with different types of codebooks (from natural scene images [20] and document images) in Fig. 5. Since images in the SOC database are distorted by common distortions in natural images, codebook from natural scene images works well. In contrast, it decreases HOS-DIQA's performance on the Newspaper database apparently because characteristics of distortions in local features and codebooks are different. Finally, the average computational time for feature extraction on images with fixed resolution from SOC database is reported. It is only 0.507 seconds/image on a laptop with Intel i5 CPU at 2.30 GHz with un-optimized MATLAB code.

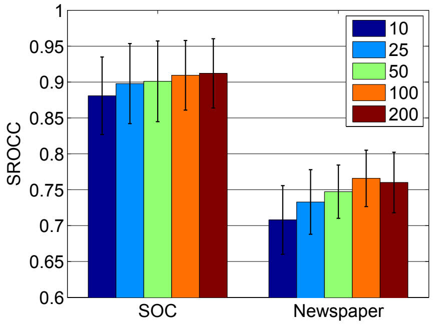


Fig. 4. Performance of HOS-DIQA with different codebook sizes (SROCC)

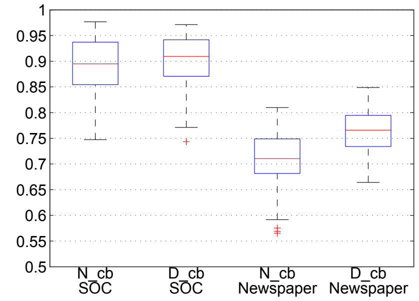


Fig. 5. Boxplot of SROCC with different types of codebooks: N_cb, codebook from document images; D_cb, codebook from natural images

4. CONCLUSION

In this paper, we design a novel no-reference document image quality assessment method. With a comprehensive 100 codeword codebook, differences between image high order statistics and codewords are computed as quality aware features. The extracted features have high correlation with OCR accuracies. Results on two document image databases show HOS-DIQA achieves state-of-the-art performance.

5. REFERENCES

[1] J. Kumar, R. Bala, H. Ding, and P. Emmett, "Mobile video capture of multi-page documents," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 35–40.

[2] P. Ye and D. Doermann, "Document image quality assessment: A brief survey," in *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 723–727.

[3] L. R. Blando, J. Kanai, T. Nartker, et al., "Prediction of ocr accuracy using simple image features," in *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR)*, 1995, vol. 1, pp. 319–322.

[4] M. Cannon, J. Hochberg, and P. Kelly, "Quality assessment and restoration of typewritten document images," *International Journal on Document Analysis and Recognition*, vol. 2, no. 2-3, pp. 80–89, 1999.

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

[6] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *2012 21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 3292–3295.

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

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

[9] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1098–1105.

[10] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 987–994.

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

[12] 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, Dec. 2012.

[13] S. Lloyd, "Least square quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[14] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[15] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Historical document layout analysis competition," in *2011 International Conference on Document Analysis and Recognition (ICDAR)*, 2011, pp. 1516–1520.

[16] Y. Jia and T. Darrell, "Heavy-tailed distances for gradient based image descriptors," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 397–405.

[17] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[18] R. Smith, "Isri-ocr evaluation tool," `http://code.google.com/p/isri-ocr-evaluation-tools/`.

[19] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in *Camera-Based Document Analysis and Recognition (CBDAR)*, 2013, pp. 113–125.

[20] J. Xu, Q. Li, P. Ye, H. Du, and Y. Liu, "Local feature aggregation for blind image quality assessment," in *Proc. IEEE Conf. Visual Communication and Image Processing (VCIP)*, 2015, pp. 1–4.