DEEP BLIND IMAGE QUALITY ASSESSMENT BY EMPLOYING FR-IQA

Jongyoo Kim and Sanghoon Lee

Yonsei University

ABSTRACT

In this paper, we propose a convolutional neural network (CNN)-based no-reference image quality assessment (NR-IQA). Though deep learning has yielded superior performance in a number of computer vision studies, applying the deep CNN to the NR-IQA framework is not straightforward, since we face a few critical problems: 1) lack of training data; 2) absence of local ground truth targets. To alleviate these problems, we employ the full-reference image quality assessment (FR-IQA) metrics as intermediate training targets of the CNN. In addition, we incorporate the pooling stage in the training stage, so that the whole parameters of the model can be optimized in an end-to-end framework. The proposed model, named as a blind image evaluator based on a convolutional neural network (BIECON), achieves state-of-the-art prediction accuracy that is comparable with that of FR-IQA methods.

Index Terms— Convolutional neural network, deep learning, image quality assessment, no-reference image quality assessment.

1. INTRODUCTION

The goal of no-reference image quality assessment (NR-IQA) is to predict the perceptual quality of a distorted image without referring to undistorted version of it. In contrast, full-reference image quality assessment (FR-IQA) uses both distorted and reference images. The primary advantage of FR-IQA is that it can directly use the difference between the distorted and reference images [1, 2, 3, 4]. In contrast, owing to a lack of information, it is more difficult for NR-IQA to predict visual quality than FR-IQA. NR-IQA relies on obtaining hand-crafted features effectively without using reference images [5]. Consequently, the accuracy of FR-IQA has been widely used as an upper bound of NR-IQA accuracy when performance is objectively evaluated. Reducing the performance gap between FR- and NR-IQA methods is a challenging issue.

In recent years, convolutional neural networks (CNNs) become frequently used in computer vision and image processing [6]. The fundamental difference between this and conventional machine learning is that, rather than using hand-crafted features, CNNs search for highly optimized features automatically. In this study, we adopted a CNN in NR-IQA to achieve better prediction accuracy. By optimizing the end-to-end framework of NR-IQA, we can expect highly optimized parameters in the deep model. However, it is difficult to apply deep learning to the NR-IQA framework seamlessly because we face new obstacles, which are described in the following sections.

1) Lack of training data Generally, an IQA training dataset is insufficient to train a deep neural network. For example, the LIVE IQA Database [7] contains about 200 images for each distortion type, while ImageNet datasets [8] contains 50 million labeled data. One solution would be to use data augmentation through rotation, cropping, reflection, and so on. However, it is unknown whether any transformation would change perceptual quality scores significantly.

2) Absence of local ground truth targets Rather than using raw images, a patch-based method, where the input image is divided into multiple patches, is more suitable for deep learning. However, for NR-IQA, the patch-based method is problematic because a ground truth target for each patch does not exist. Generally, only one scalar score is obtained for each image. It is well known that the error visibility varies dramatically depending on the surrounding signals [9], which results in various local scores across the image location.

To address these problems, we propose a CNN-based NR-IQA framework, called as a blind image evaluator based on a CNN (BIECON). BIECON employs the patch-based approach to avoid the first problem. To overcome the second problem, local quality maps are borrowed from existing FR-IQA methods. The substituted targets are reliable because the FR-IQA metrics provide the best correlation scores among all of the IQA metrics [10, 11]. Since each image patch is trained to the target in a supervised manner, meaningful features sufficient for IQA can be generated. In addition, the pooling stage is considered during the training process, which results in optimized parameters for the whole model.

1.1. Related work

A number of NR-IQA methods have been developed based on natural scene statistics (NSS) under the assumption that nat-

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B2014525).

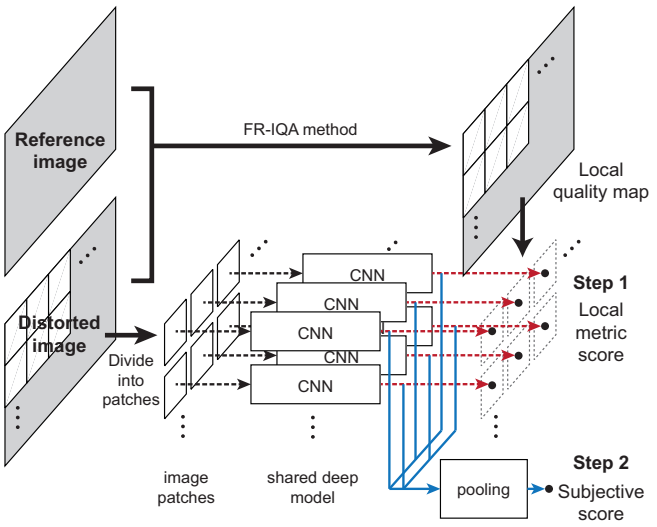

Fig. 1. Overall training process of proposed NR-IQA framework. Step 1: CNN model is regressed onto local metric score derived using FR-IQA metric. Step 2: pooled features are regressed onto subjective score.

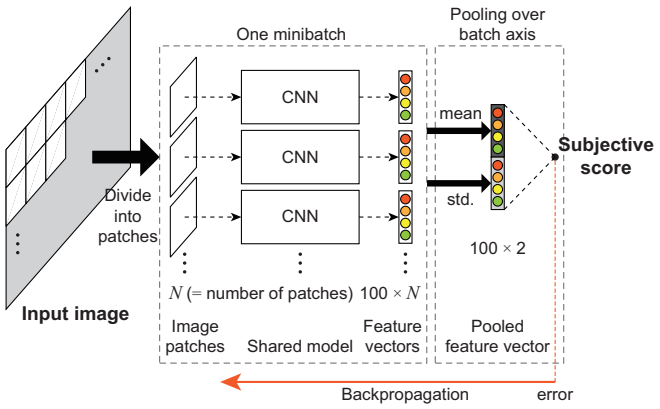

Fig. 2. Training flow of MOS regression via pooling.

ural scenes contain statistical regularity. NSS-based metrics were developed by analyzing statistical features of corrupted images in the transformed domain [12] or in the spatial domain [13, 11]. Differing from NSS-based approaches, learning-based metrics that utilize the power of machine learning have been proposed for NR-IQA [14, 15].

Relatively recently, a few NR-IQA metrics have adopted the deep learning technique to reinforce prediction accuracy [16, 17]. Although they enhanced the performances of the NR-IQA metrics, these focused on NR-IQA processes. Initially, hand-crafted features were developed, and then the deep model was used to replace a conventional regression machine. Kang et al. first applied a CNN to the NR-IQA framework by regressing images on the target subjective scores without hand-crafted features [10]. The employed method is closest to our framework. However, an equal mean opinion score (MOS), which cannot represent pixel-wise perceptual quality variation over the spatial domain, was used for all patches in an image.

2. FRAMEWORK OF BIECON

As shown in Fig. 1, the proposed NR-IQA framework is trained via two steps. In Step 1, the model is trained with respect to each patch. Each is regressed onto the target local metric score, which is derived from a FR-IQA metric. In Step 2, all of the model parameters are optimized simultaneously to minimize training loss. Throughout, let I_r be a reference image, I_d be the distorted image, and (i, j) be the pixel coordinates on the image.

The model is composed of a 5×5 convolutional layer with 2×2 maximum pooling, another 5×5 convolutional layer, four consecutive fully connected layers, and one fully connected regression layer. In the first and second convolutional layers, 48 and 64 kernels are used, respectively. The last hidden node with 100 dimensions is utilized as the feature vector and is fed into the pooling stage in Step 2.

2.1. Step 1: Local metric score regression

In Step 1, the input patches are trained independently without using the spatial correlations with the other patches of their original images. Here, four FR-IQA metrics are adopted to generate local quality maps: structural similarity index (SSIM) [1], gradient magnitude similarity deviation (GMSD) [2], feature similarity index (FSIM) [3], and visual saliency-based index (VSI) [4]. These metrics have demonstrated high prediction accuracy in the literature.

The input images are normalized by local normalization following [13]. Then the images are divided into a number of patches. For each image, the patches with low spatial variations are removed from $\mathbf{p}_d$ because the monotonous region does not contain sufficient information to train the CNN actively according to distortion [11]. Each patch is then regressed onto the corresponding local metric score obtained using FR-IQA, which provides more meaningful clues to train the CNN and achieve high FR-IQA performance by miniming the following loss function.

$$\mathcal{L}_1(\mathbf{p}_k, s_k; \theta, \phi_1) = \frac{1}{K}\sum_{k=1}^{K}\left\| g_{\phi_1}(f_\theta(\mathbf{p}_k)) - s_k \right\|_F^2. \quad (1)$$

where $\mathbf{p}_k$ is the k^{th} image patch, s_k is the corresponding local quality score, $f_\theta(\cdot)$ is the CNN model with parameters θ. and $g_{\phi_1}(\cdot)$ is the regression model with parameters ϕ_1.

2.2. Step 2: Subjective score regression

BIECON employs variable minibatch-based stochastic gradient descent, where the statistical moments of the features are considered. Furthermore, the pooling stage is incorporated in the complete optimization process to enhance the prediction accuracy, as shown in Fig. 2.

To extract a low-dimensional feature, the mean and standard deviation are selected. Taking an average is the most common pooling strategy adopted in many IQA algorithms [1, 11]. Standard-deviation (STD)-based pooling enables the global variation of local quality degradation to be determined [11, 2].

First, the image patches from each image are grouped into each minibatch and fed into the CNN in Step 1. From the obtained bundle of image feature vectors, the mean-pooled vector $\boldsymbol{\mu}$ and standard-deviation-pooled vector $\boldsymbol{\sigma}$ are derived. Then, the concatenated feature vector $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ is regressed onto the subjective score. Given the image I_d^j and the corresponding subjective score S_j, the objective function of Step 2 is defined as follows:

$$\mathcal{L}_2(I_d^j, S_j; \theta, \phi_2) = \left\| h_{\phi_2}(pool(\mathbf{f}_\theta(I_d^j))) - S_j \right\|_F^2 \quad (2)$$

where $\mathbf{f}_\theta(I_d^j) = [f_\theta(\mathbf{p}_1), ..., f_\theta(\mathbf{p}_{N_j})]^\intercal$ denotes the bundled output of the CNN, N_j is the number of patches of j^{th} image, $pool(\cdot)$ denotes the element-wise mean and STD pooling, and $h_{\phi_2}(\cdot)$ is the regression model with parameter ϕ_2. To enhance the training convergence rate, ADAM [18] is employed. The default hyperparameters suggested in the literature [18] were used in the experiment.

3. EXPERIMENT AND ANALYSIS

3.1. Database

3.2. Performances comparison

For evaluation of IQA algorithms, we used the LIVE IQA database [7], which contains 29 reference images and 982 distorted images with five distortion types: JPEG and JPEG2000 (JP2K), white Gaussian noise (WN), Gaussian blur (BLUR), and Rayleigh fast-fading channel distortion (FF).

We evaluated the performance of the IQA algorithm using two standard measures, i.e., Spearmans rank order correlation coefficient (SROCC) and Pearsons linear correlation coefficient (LCC). During the experiment, 80% of the database was chosen randomly for training, and the remaining 20% was used for testing, where there was no overlapped reference images in two sets.

The NR-IQA metric was trained in two steps, as described in Section 2. To evaluate the proposed method, we compared it to five FR-IQA metrics (PSNR, SSIM [1], GMSD [2], FSIMc [3], VSI [4]) and four NR-IQA metrics (DIIVINE [5], BRISQUE [13], NIQE [11] and CNN [10]). In Tables 1 and 2, the SROCC and LCC of the NR-IQA metrics are compared according to distortion type. The top model for each evaluation criterion are shown in boldface. BIECON achieves the best correlation scores among all of the NR-IQA algorithms. Furthermore, it is remarkable that the achieved score is very close to those of state-of-the-art FR-IQA methods, such as FSIMc.

Table 1. SROCC comparison on the LIVE IQA database.

Metrics	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.895	0.881	0.985	0.782	0.891	0.876
SSIM	0.961	0.976	0.969	0.952	**0.956**	0.948
GMSD	0.971	0.978	0.974	0.957	0.942	0.960
FSIMc	**0.972**	**0.979**	0.971	**0.968**	0.950	**0.963**
VSI	0.960	0.976	**0.984**	0.953	0.943	0.952
DIIVINE	0.912	0.921	**0.982**	0.937	0.869	0.925
BRISQUE	0.914	0.965	0.977	0.951	0.877	0.939
NIQE	0.917	0.938	0.967	0.934	0.859	0.914
CNN	0.952	**0.977**	0.978	0.962	0.908	0.956
BIECON	**0.957**	0.967	0.971	**0.965**	**0.963**	**0.960**

Table 2. LCC comparison on the LIVE IQA database.

Metrics	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.876	0.903	0.917	0.780	0.880	0.872
SSIM	0.941	0.946	**0.982**	0.900	**0.951**	0.945
GMSD	**0.963**	0.979	0.977	0.954	0.939	0.956
FSIMc	0.939	**0.985**	0.976	**0.979**	0.922	**0.960**
VSI	0.939	0.957	0.964	0.945	0.931	0.948
DIIVINE	0.923	0.935	**0.987**	0.937	0.891	0.923
BRISQUE	0.923	**0.973**	0.985	0.951	0.903	0.942
NIQE	0.937	0.956	0.977	**0.952**	0.913	0.915
CNN	0.917	0.938	0.967	0.934	0.859	0.953
BIECON	**0.950**	0.965	0.974	0.945	**0.951**	**0.958**

Table 3. SROCC and LCC comparison for each FR-IQA metric on the LIVE IQA database.

Metrics	SSIM	GMSD	FSIMc	VSI
SROCC	0.940	0.913	0.959	0.804
LCC	0.951	0.906	0.961	0.827

3.3. Dependency on FR-IQA metrics

We evaluated which FR-IQA metric harmonized with BIECON. The local metric scores derived from SSIM, GMS, FSIMc, and VSI were used as local regression targets for training in Step 1. To reduce training time, the parameters were optimized in the fully connected layers during Step 2.

The final SROCC and LCC values are given in Table 3, which also shows the same ranks as in the convergence speed results. Note that the obtained ranks are inconsistent with their prediction accuracies when they work as FR-IQA solely, which are given in Table 1. In particular, VSI is not well harmonized with BIECON since it ranks last. The power of VSI lies in the weighted pooling such that the generated local quality maps particularly emphasize salient objects. Generally, saliency is determined by analyzing the whole image, which inadequate for patch-wise training. In contrast, FSIMc fits well with BIECON. FSIMc adopts PC as the visual weight. Since PC acts as an edge detector, it can be inferred from the local information. Therefore, in the rest of our experiments, FSIMc was used as the default FR-IQA metric.

Table 4. SROCC comparison on the TID2008 database.

Metrics	JP2K	JPEG	WN	BLUR	ALL
SSIM	0.963	0.935	0.817	0.960	0.902
BRISQUE	0.832	0.924	0.829	0.881	0.896
BIECON	0.832	0.924	0.982	0.900	0.945

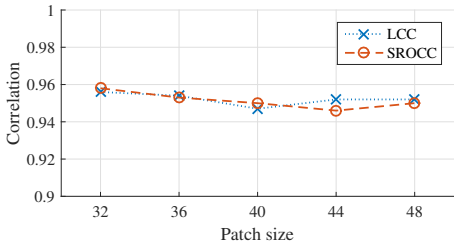

Fig. 3. Comparison of SROCC and LCC according to patch size.

3.4. Cross dataset test

To evaluate the generalization capability, we trained BIECON on the LIVE IQA database and tested on the TID2008 database [19]. In the TID2008 database, we chose only four common distortion types. The computer-generated image was excluded for the test because it had a different data distribution than those of the natural images [13]. Since the MOS has a different scale and meaning than DMOS, we added a logistic regression to match the predicted DMOS to MOS to evaluate LCC. The results of the cross-dataset test are shown in Table 4. It can be concluded that BIECON performs well in terms of the subjective score predictions and that the performance does not depend on the database

3.5. Effect of patch size

To investigate the effects of patch size on the final prediction accuracy, five patch sizes (32, 36, 40, 44, and 48) were tested. As shown in Fig. 3, the obtained SROCC and LCC scores are nearly the same regardless of patch size. Patch size 32 exhibits a slightly better result, which is caused by the larger training dataset. Since the patches are generated in a non-overlapping manner, a larger patch size leads to a decreased amount of data. In addition, the complexity of the model increases exponentially as the patch size increases. Therefore, a patch size of 32 was used as the template in the experiments.

3.6. Predicted local metric score visualization

As a consequence of Step 1, the model can predict the local metric score without using reference images. In Fig. 4, the predicted local quality maps and their ground truth versions are compared to validate Step 1 of BIECON. M_{FSIM_C} was used in the experiment for the local quality maps. Each row in Fig. 4 shows each distortion type, i.e., JP2K, WN, and BLUR. Generally, the learned model derives the local quality map fairly accurately with various distortion types. Note that our model does not consider patches with low standard deviations during the training stage because they rarely have meaningful information about image distortion. Therefore, the prediction of local metric scores on homogeneous patches is not trustworthy and should be ignored in the subjective score prediction stage.

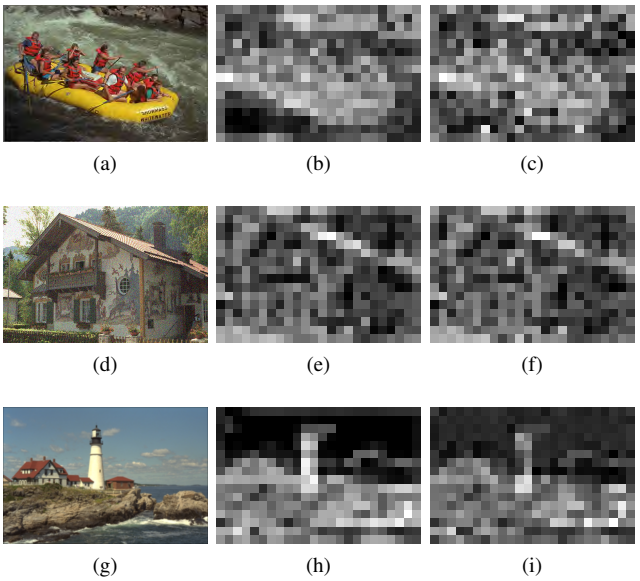

Fig. 4. Examples of predicted local metric score maps for each distortion type: (a), (d), and (g) show distorted images; (b), (e), and (h) are local quality maps M_{FSIM_C}; and (c), (f), and (i) are predicted local quality maps obtained by BIECON. Each row indicates three different distortion types (JP2K, WN, and BLUR).

4. CONCLUSION

Fully applying a deep model to the patch-based NR-IQA framework is problematic in that there is no ground truth target for each patch. We resolved this issue by employing local quality maps derived by FR-IQA metrics as intermediate regression targets. The proposed NR-IQA metric follows the behavior of FR-IQA, which enables BIECON to predict the subjective score with high accuracy. In addition, by incorporating the pooling stage into the final regression on the subjective score, the end-to-end optimization is conducted to derive highly optimized CNN parameters. As a result, BIECON outperformed all of the benchmarked NR-IQA methods, and demonstrated performance comparable to state-of-the-art FR-IQA.

5. REFERENCES

[1] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[2] Wufeng Xue, Lei Zhang, Xuanqin Mou, and A.C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.

[3] Lin Zhang, D. Zhang, Xuanqin Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[4] Lin Zhang, Ying Shen, and Hongyu Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, 2014.

[5] A.K. Moorthy and A.C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, 2011.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[7] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, 2006.

[8] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[9] Gordon E. Legge and John M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Am.*, vol. 70, no. 12, pp. 1458–1471, 1980.

[10] Le Kang, Peng Ye, Yi Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1733–1740.

[11] A. Mittal, R. Soundararajan, and A.C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[12] M.A. Saad, A.C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, 2012.

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

[14] Huixuan Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 305–312.

[15] Peng Ye, J. Kumar, Le Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1098–1105.

[16] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–1, 2014.

[17] Yuming Li, Lai-Man Po, Xuyuan Xu, Litong Feng, Fang Yuan, Chun-Ho Cheung, and Kwok-Wai Cheung, "No-reference image quality assessment with shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, 2015.

[18] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations (ICLR)*, 2015.

[19] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, M Carli, and F Battisti, "TID2008 - A database for evaluation of full-reference visual quality assessment metrics," *Adv. Mod. Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.