A DEEP LEARNING BASED NO-REFERENCE IMAGE QUALITY ASSESSMENT MODEL FOR SINGLE-IMAGE SUPER-RESOLUTION

*Bahetiyaer Bare, Ke Li, Bo Yan**

Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, China

Bailan Feng, Chunfeng Yao

Noah's Ark Laboratory, 2012Labs Huawei Technologies Co., Ltd., Beijing, China

ABSTRACT

Single-image super-resolution (SISR) is a very important and classic problem of the computer vision community. Although a lot of SISR methods have been proposed, few studies have been conducted to address the quality assessment of SISR methods. In this paper, we proposed a deep learning based no-reference image quality assessment (NR-IQA) model for SISR. We took small patches from images to form our training set and labeled them with different scores. With the aid of well-designed architecture and training strategy, our method achieved a performance leap than state-of-the-art methods. Experimental results proved the generalizability and the effectiveness of the proposed model.

***Index Terms*—** Quality assessment, Super-resolution, Convolutional neural network, Deep learning, No-reference

1. INTRODUCTION

Single-image super-resolution (SISR) algorithms aim to restore a high-resolution (HR) image from a low-resolution (LR) one. Until now, there have been numerous SISR methods proposed in the literature. But they always use peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [1] as the evaluation metric, which have been designed for image degradation. In SISR benchmark study [2], such as PSNR and SSIM are proved to have lower consistency with human visual system (HVS). So, it is very important to develop a specific quality assessment method for SISR.

Information fidelity criterion (IFC) [3] is proved to have better consistency with HVS in the SISR benchmark study [2]. However, since it is a full-reference metric, reference images is not always available for the images to be tested in the practical application. Therefore, developing a no-reference image quality assessment (NR-IQA) for SISR has a strong practical significance. Although there have been many NR-IQA methods [4, 5, 6, 7, 8, 9], they are not specially designed for SISR. Generic NR-IQA methods are designed based on image signal and noise, while quality assessment of SISR should be designed based on visual perception.

In [10], Ma $et\ al.$ firstly addressed the NR-IQA of SISR methods with a two-stage regression model. More importantly, they built a database for quality assessment of SISR methods. They chose 30 images from Berkeley segmentation dataset (BSD) [11] and processed them with nine different SISR methods [12, 13, 14, 15, 16, 17, 18] at six different settings. Hence, their database has 1620 images with subjective scores. With the aid of this database and well-designed 138 hand-crafted features, their two-stage regression method achieved state-of-the-art performance. In spite of its good performance, it is time costing because of the 138 hand-crafted features. In recent years, deep learning technique plays dominant role in the computer vision field, especially convolutional neural networks (CNNs). CNNs based methods have better performance without hand-crafted features. To the best of our knowledge, there is no CNNs-based NR-IQA method for SISR proposed yet. Therefore, these reasons motivate us to develop a CNNs-based NR-IQA method for SISR.

In this paper, we proposed a deep learning based NR-IQA for SISR. The main contributions of our work lie in two aspects:

1) We proposed a deep CNNs model for no-reference quality assessment of SISR metrics. Inspired by the successful models used for image classification, we designed a specific NR-IQA model for SISR.

2) Since our proposed model is a patch-wise model, we designed a label distribution method. So, each of the patches in the training set has different labels. This method improves the performance of the proposed model.

The rest of the paper is organized as follows. In Section 2, we describe our proposed model. Then we present experimental results and discuss the properties of proposed deep CNNs model in Section 3. Finally, we draw conclusions in Section 4.

*This work was supported in part by NSFC (Grant No.: 61522202; 61772137) and Huawei Technologies Co., Ltd (Contract No.: YBN2017050058).

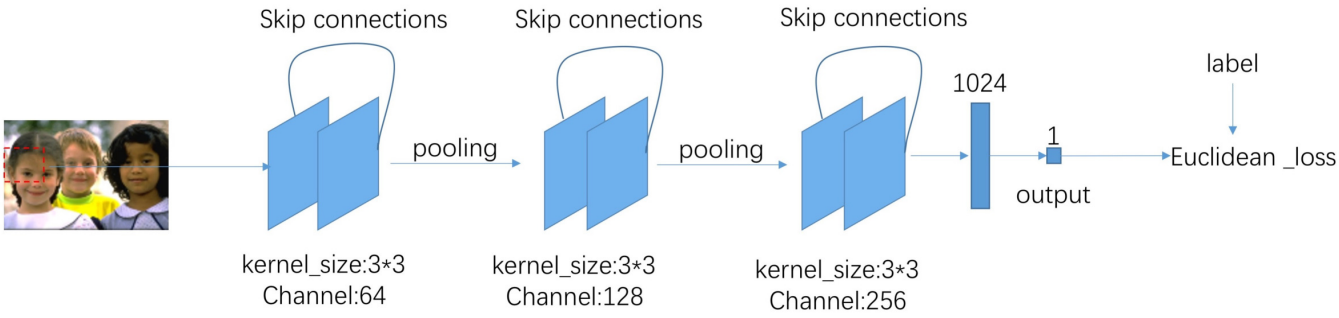

Fig. 1. Network architecture of the proposed model. Our network consists of six convolutional layers, two max pooling layer, three skip connections, and two fully connected (FC) layers. The output of the second FC layer is the predicted quality score of the input image patch. The overall quality score of the input image is the average sum of the predicted score of small patches.

2. PROPOSED METHOD

In this section, we first introduce the proposed model. Then, we introduce the network configuration and each of the used layers. Finally, we introduce the data preparing method of our proposed model, especially the proposed label distribution method.

2.1. Proposed model

We present an accurate deep CNNs model for NR-IQA of SISR. For an input image, we split it to multiple 32×32 small patches without overlaps at first. Then our proposed model predict quality scores for each small patch. At the end, the final quality score of the input image is the average sum of the quality score of small patches. The network architecture of our model is shown in Fig.1. As shown in this figure, our model consists of six convolutional layers with taking rectified linear unit (ReLU) as activation function, two fully connected layers and two max pooling layers. It is worth noting that we add three skip connections to our network inspired by deep residual network [19].

2.2. Layers

The network configuration is listed in Table 1. Our model has six convolutional layers to extract local features. The filter size of convolutional layers is fixed to 3×3. When the feature map size is reduced to half, we increase the filter number to twice of the previous one. Each convolutional layer takes ReLU as activation function. Denote C_j as the feature map of the j^{th} layer, W_j and B_j as the weight and bias of the j^{th} layer, then the local information is extracted into deeper layers by Eq.(1), where * denotes the calculation of convolution. We set the bias B_j to zero in our model.

$$C_{j+1} = \max(0, W_j * C_j + B_j) \tag{1}$$

In order to reduce the complexity and computation cost, we add pooling layers to our model. We employed max pooling with 2×2 window size. So after the pooling layer, the feature map size is reduced to half. The max pooling is applied as Eq.(2), where R is the pooling region of corresponding feature map.

$$C_{j+1} = \max_{R} C_j \tag{2}$$

In order to easily converge and prevent gradient descent, we add three skip connections to our method. As revealed in [19], skip connection is a very effective way to train deep CNNs model. The first and second skip connections add the outputs of the previous convolutional layers and send them to the max pooling layer. Similarly, the third skip connection adds the outputs of the 5^{th} and 6^{th} convolutional layers and send them to the first fully connected layer. The second fully connected layer outputs the predicted value. The image score is predicted by minimizing the following Euclidean loss,

$$\min_{W} ||f(X; W) - Y|| \tag{3}$$

where X and Y denote the input image patch and its label respectively and $f(X; W)$ be the predicted score of X with network weights W.

Stochastic gradient decent (SGD) and back-propagation are used to solve the parameters W in Eq.(3) that minimizes the distance between predicted quality score and ground truth.

$$\begin{aligned} \Delta_{i+1} &= m \cdot \Delta_i - \eta \frac{\partial L}{\partial W_i^j} \\ W_{i+1}^j &= W_i^j + \Delta_{i+1} - \lambda \eta W_i^j \end{aligned} \tag{4}$$

where m is the momentum factor, η is the learning rate, j is the index of the layer, Δ_{i+1} is the gradient increment for training iteration i, and λ is the weight decay factor. Momentum factor and weight decay factor were fixed to 0.9 and

0.0005 respectively in our model training. Learning rate is set to different values ranges from 0.01 to 0.00001 at different epoches.

2.3. Data preparing

For the images in training and testing sets, we pre-process them using the same method in previous works [4, 6]. For an image, we compute locally normalized luminances via local mean subtraction and divisive normalization. This is beacause applying a local non-linear operation to log-contrast luminances to remove local mean displacements from zero log-contrast and to normalize the local variance of the log-contrast has a decorrelating effect [4]. This pre-processing operation can be formulated as:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C}, \tag{5}$$

where $\hat{I}(i,j)$ denotes the normalized image, $I(i,j)$ denotes the input image, $\mu(i,j)$ denotes the mean value, $\sigma(i,j)$ denotes the contrast value, and C is a constant value preventing the denominator to be zero. In our method, we set $C = 1$. The mean value $\mu(i,j)$ and the contrast value $\sigma(i,j)$ can be formulated as:

$$\mu(i,j) = \sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} I(i+p, j+q), \tag{6}$$

$$\sigma(i,j) = \sqrt{\sum_{p=-P}^{p=P} \sum_{q=-Q}^{q=Q} (I(i+p, j+q) - \mu(i,j))^2}, \tag{7}$$

where $w = wp, q|p = -P, ..., P, q = -Q, ...Q$ is a 2D circularly-symmetric Gaussian weighting function sampled out to $7/6$ standard deviations and rescaled to unit volume. In our implementation, $P = Q = 7$.

After pre-processing, we split an image to small patches. We took 32×32 sized patches from images with stride 32. So, the small patches taken from an image have no overlapping region. In training, due to the database just has the whole image's label, we design a label distribution method for the small patches, which are taken from a same image. The label distribution method can be formulated as:

$$S_p = S_{image} + \frac{MSE_p - MSE_{average}}{|MSE_p - MSE_{average}|_{max}} * S_{image}, \tag{8}$$

where S_p denotes the label score of small patches; S_{image} denotes the label score of a whole image; MSE_p denotes the mean squared error between small patches and the corresponding small patches from original image, and $MSE_{average}$ denotes the average value of MSE_p from an image. Then, we normalize all the labels between 0 and 1. After label distribution procedure, we can begin our training procedure. This will detailed in the next section.

Table 1. Configuration of our deep CNNs model

Layer name	Padding	Filter size	Stride	Output size
input				$32 \times 32 \times 3$
conv1 / ReLU	1	3×3	1	$32 \times 32 \times 64$
conv2 / ReLU	1	3×3	1	$32 \times 32 \times 64$
skip connections 1				$32 \times 32 \times 64$
max pooling 1		2×2	2	$16 \times 16 \times 64$
conv3 / ReLU	1	3×3	1	$16 \times 16 \times 128$
conv4 / ReLU	1	3×3	1	$16 \times 16 \times 128$
skip connections 2				$16 \times 16 \times 128$
max pooling 2		2×2	2	$8 \times 8 \times 128$
conv5 / ReLU	1	3×3	1	$8 \times 8 \times 256$
conv6 / ReLU	1	3×3	1	$8 \times 8 \times 256$
skip connections 3				$8 \times 8 \times 256$
fc1				1024
fc2				1

3. EXPERIMENT

In this section, we demonstrate experiments to validate the effectiveness of the proposed model. We first introduce the dataset and the evaluation method. Then, we introduce the training details. Moreover, we demonstrate the experimental results. In the end, we discuss the experimental results.

3.1. Dataset

In [10], Ma *et al*. built a dataset for quality assessment of SISR. We used this dataset to train and test our model. This dataset took 30 images from Berkeley segmentation dataset and processed them with 9 different methods at 6 different settings. These 9 different methods are : bicubic interpolation (Bicubic), back projection (BP) [20], Shan08 [12], Glasner09 [13], Yang10 [14], Dong11 [15], Yang13 [16], Timofte13 [17], and SRCNN [18]. And 6 different settings are downsample factors $s \in \{2, 3, 4, 5, 6, 8\}$ with corresponding kernel width factors $\sigma \in \{0.8, 1.0, 1.2, 1.6, 1.8, 2.0\}$. Thus this dataset has 1620 images with subjective perceptual scores.

Following the experimental settings in [10], we used spearman rank correlation coefficients (SROCC) value to measure the correlation between subjective scores and the predicted objective scores. SROCC measures how well one quantity can be described as a monotonic function of another quantity.

3.2. Training details

We apply deep learning toolbox Matconvnet [21] to train the deep CNNs model for NR-IQA of SISR methods. The SISR

Table 2. Mean SROCC value comparison with state-of-the-arts.

	Bicubic	Bp	Shan08	Glasner09	Yang10	Dong11	Yang13	Timofte13	SRCNN	Overall
PSNR	0.572	0.620	0.564	0.605	0.625	0.634	0.631	0.620	0.645	0.604
FSIM	0.706	0.770	0.648	0.778	0.757	0.765	0.768	0.756	0.780	0.747
SSIM	0.588	0.657	0.560	0.648	0.649	0.649	0.652	0.656	0.660	0.635
IFC	0.884	0.880	0.934	0.890	0.866	0.865	0.870	0.881	0.885	0.810
BIQI	0.770	0.740	0.254	0.523	0.556	0.236	0.646	0.563	0.617	0.482
DIVINE	0.784	0.842	0.653	0.426	0.525	0.763	0.537	0.122	0.625	0.589
CNNIQA	0.926	0.956	0.832	0.914	0.943	0.921	0.927	0.924	0.908	0.904
CORNIA	0.889	0.932	0.907	0.918	0.908	0.912	0.923	0.911	0.898	0.919
BLIINDS	0.886	0.931	0.664	0.862	0.901	0.811	0.864	0.903	0.843	0.853
BRISQUE	0.850	0.917	0.667	0.738	0.886	0.783	0.784	0.843	0.812	0.802
Ma *et al.*	0.933	0.966	0.891	0.931	0.968	0.954	0.958	0.930	0.949	0.931
Ours without label distribution	0.961	0.965	0.919	0.946	0.960	0.934	0.954	0.920	0.941	0.947
Ours with label distribution	**0.973**	**0.977**	**0.926**	**0.950**	**0.971**	**0.955**	**0.971**	**0.934**	**0.953**	**0.958**

quality assessment dataset [10] is used to train and test our model. At the training stage, we first extract 32×32 patches with stride 32 from the images in the SISR quality assessment dataset. Since different image patches have different quality values, we used the proposed label distribution method to label them with different values. Then we train our model with different learning rates. Learning rate is changed to $1/10$ of the previous one at the interval of ten epoch. It varies from 0.01 to 0.00001. In order to control the gradient at specified range, we set gradient clip value to a fixed value 0.1. Finally, our deep CNNs model is obtained by training after 40 epoches. It took 40 minutes to train a model with GTX1070 GPU. The experimental results are demonstrated in the next subsections.

3.3. Experimental results

Following experimental settings in [10], we adopt 5-fold validation to test our model. In the test phase, we randomly split the dataset to 5 folds. Then, we select 4 folds to form training set and the remaining one fold be the testing set. We continue this process until each fold is selected as a testing set for one time. After 5 iterations, we can get the predicted quality score of each image in dataset. For fair comparison with other methods, we run 5-fold test 50 times and demonstrate mean value of 50 tests' results.

We compare our methods with some generic full-reference image quality assessment methods: PSNR, SSIM [1], IFC [3], and feature similariy index (FSIM) [22]; and some generic NR-IQA methods: BIQI [8], DIVINE [9], CNNIQA [6], CORNIA [5], BLIINDS [7], and BRISQUE [4]; and one specific SISR quality assessment methods: Ma *et al.* [10]. Other methods' results are reported in [10]. For fair comparison, we took the results of NR-IQA methods with training on the same dataset. We compare our results with other methods on overall SROCC value and the seperate SROCC value on each different SISR methods in the dataset.

The experimental results are displayed in Table 2. In order to highlight the proposed label distribution method, we also present the proposed model results without it. Namely, label each patches from same image with the image's perceptual score.

3.4. Discussion

As we can see from the results in Table 2, our proposed model outperformed other methods in overall comparison. This proved the superiority of the proposed model. Our model possesses strong representation ability with the wider filter number at small scales. Even though training with same patch score, our model outperformed other compared methods on overall SROCC value comparison. With the aid of the proposed label distribution method, performance of our method is further improved. In each different method comparison, our proposed method also outperformed other methods. This confirmed the generalization ability of our method. Overall, our proposed deep CNNs model has state-of-the-art performance. It is worth noting that the proposed model is easy to use and train, because it does not need hand-crafted features. For an input image, after pre-processing, we can input it to the proposed CNNs model and get the quality score of this image.

4. CONCLUSION

In this paper, we proposed a deep learning based no-reference image quality assessment method for single image super resolution. We designed a convolutional neural networks architecture for the no-reference quality assessment of single image super resolution. By applying the proposed label distribution method, our proposed model achieved promising results. Experimental results verified the superiority of our method and confirmed that our method achieved a performance leap compared to state-of-the-arts.

5. REFERENCES

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

[2] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang, "Single-image super-resolution: a benchmark," in *European Conference on Computer Vision*. Springer, 2014, pp. 372–386.

[3] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[4] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[5] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[6] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[7] Michele A Saad, Alan C Bovik, and Christophe Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[8] Anush Krishna Moorthy and Alan Conrad Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal processing letters*, vol. 17, no. 5, pp. 513–516, 2010.

[9] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[10] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang, "Learning a no-reference quality metric for single-image super-resolution," *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[11] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2001, vol. 2, pp. 416–423.

[12] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang, "Fast image/video upsampling," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 5, pp. 153, 2008.

[13] Daniel Glasner, Shai Bagon, and Michal Irani, "Super-resolution from a single image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 349–356.

[14] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[15] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838–1857, 2011.

[16] Jianchao Yang, Zhe Lin, and Scott Cohen, "Fast image super-resolution based on in-place example regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1059–1066.

[17] Radu Timofte, Vincent De Smet, and Luc Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.

[18] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] Michal Irani and Shmuel Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[21] Andrea Vedaldi and Karel Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 689–692.

[22] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.