Exploiting Neural Models for No-reference Image Quality Assessment

Cenhui Pan [1], Yi Xu, Yichao Yan, Ke Gu, Xiaokang Yang

Shanghai Jiao Tong University, Shanghai, China
Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai, China
Cooperative Medianet Innovation Center, Shanghai, China
[1]pancenhui920123@outlook.com

***Abstract*—We propose an improved algorithm for no-reference image quality assessment (NR-IQA) using the convolutional neural network (CNN) and neural theory based saliency detection. Firstly, we extract non-overlapping patches from the input image. For each patch, we obtain the quality score by CNN network, which consists of seven layers and integrates feature learning and regression into image patch quality estimation. Considering that the patches attracting much attention take significant role in visual perception, an efficient technique based on free energy based neural model is used to detect the saliency map. This saliency map is then applied as a weighting mask to output the quality score of the whole image. Results of experiments show that our algorithm achieves state-of-the-art performance, as compared with the prevailing IQA methods.**

***Index Terms*—No-Reference, Image Quality Assessment, convolutional neural network, free energy**

I. Introduction

Based on the availability of reference images, image quality assessment is classified into three categories: full reference (FR) IQA, reduced-reference (RR) IQA and no-reference (NR) IQA. FR-IQA algorithms can measure the quality of image directly by comparing the distorted image with the undistorted reference image. Typical examples of FR-IQA algorithms include VIF [1], SSIM [2], FSIM [3] and ADD-SSIM [4]. RR-IQA methods are provided with partial information about the original reference image. However, ideal reference images are actually unavailable in many practical computer vision applications. The study of NR-IQA algorithms is required. No-reference image quality assessment (NR-IQA), which aims at predicting the quality of digital images without the non-distorted reference image, is one of the most challenging tasks of objective image quality measures.

Many successful research teams use the natural scene statistics (NSS) model, because the regularity of nature images has been proved in many visual science literatures. The NSS features can be extracted in the wavelet transform, the DCT transform domain or the spatial domain. The DIIVINE [5] approach uses a two-stage framework and obtains the statistics properties from a wavelet coefficient model. BLIINDS-II [6] is a fast single-stage algorithm, which relies on the image DCT coefficients since the energy of the input image concentrates on a small block of DCT coefficients. The CORNIA [7] algorithm shows good performance by directly using raw-image-patches as local descriptors and learning a dictionary to obtain effective image representations. The BRISQUE [8] algorithm establishes the statistical model of the local normalized luminance values and outputs the quality scores through parameter analysis.

Recently, a Convolutional Neural Network (CNN) has been applied to NR-IQA. In the literature of [9], the researchers modify a $32 \times 32 - 26 \times 26 \times 50 - 2 \times 50 - 800 - 800 - 1$ network structure, such that it can predict the quality on small patches. The problem of this model is that it obtains the image quality scores by averaging the predicted patch scores, which ignores the importance of the human vision system (HVS). In [10], Li Jie *et. al* combined the CNN network and the Prewitt magnitude of segmented images, considering that the HVS is sensitive to the image edges and contours. The visual quality scores are obtained by the quality score and the weight of each image patch, which is inferred from the gradient map.

Nevertheless, the HVS is not just about the edges and contours. Saliency detection, which highlights the salient object regions in a scene, includes more factors of human visual system, such as the surrounding environment, the luminance and the location. It is widely accepted that the corresponding visual saliency can improve the performance of the IQA.

In this paper, we propose an improved algorithm for NR-IQA. Instead of gray images, we design a new CNN structure focusing on color images. We then perform the saliency detection with free energy based neural theory. After that, we calculate the weight of small patches by the corresponding saliency map. The final quality score is yielded with the weighted average of each image patch.

The rest of the paper is organized as follows: In Section 2, we describe the improved NR-IQA algorithm in detail. In Section 3, we provide the comparative experiments and evaluate the performance of our approach. The conclusions are drawn in Section 4.

978-1-5090-5316-2/16/$31.00 © 2016 IEEE

II. OUR ALGORITHM FOR NR-IQA

We propose an improved NR-IQA framework using CNN and saliency detection. As compared with Li Jie's work [10], which only exploits edges to be an important factor in HVS properties, we introduce saliency map computation into image quality score estimation. It is noted that saliency computation considers many HVS characteristics, including center surround mechanism and pop-up object properties. Therefore, it is expected that we can achieve more consistent results with visual perception when saliency computation is integrated into image quality estimation. In our NR-IQA framework, firstly, we estimate the quality score of each image patch by CNN. Secondly, we compute the saliency map based on free energy theory. Finally, this saliency map is used as the weight mask of the image patches, and we get the predicted score of the whole image by computing the weighted averaging scores of all the image patches.

A. Saliency detection

More than hundreds of saliency detection models have been proposed in the recent years. Lately, a free energy principle explains that there exists a relationship between the real scene and the brain's prediction, which easily surprises the human viewers and attracts more human attention [11]. The Free Energy inspired Saliency detection model (FES) [12] searches for the gap between an image and its predicted version that is reconstructed from the input signal with a semi-parametric model, which provides a natural ground and connection to saliency detection. The final saliency map is formed to be the weighted sum across three local entropy maps in different color channels. Some examples for the saliency detection are shown in Figs 1, where (a)-(f) are respectively the original image and its corresponding images of five distortion types (fast-fading channel simulation, Gaussian blur, JP2K compression, white noise and JPEG compression) from LIVE database [13], and the associated saliency maps are shown in Fig 1(g)-(l). We can easily find that the FES algorithm could predict human salient regions correctly and the saliency maps for nearly all types of distortions share similar characteristics, because the FES calculates the saliency map based on the resized 63×47 pixel representation of the input color images. It is shown that the FES is almost no damage to the image quality information.

B. CNN for NR-IQA of color images with saliency computation

For an RGB image, we use a local normalization for each channel separately, and sample non-overlapping patches of size 32×32 from it. Suppose the intensity value of a pixel at location (x, y) is $I(x, y)$. Based on the work of the spatial NSS model works, we compute normalized values $\widetilde{I}(x, y)$ as follows:

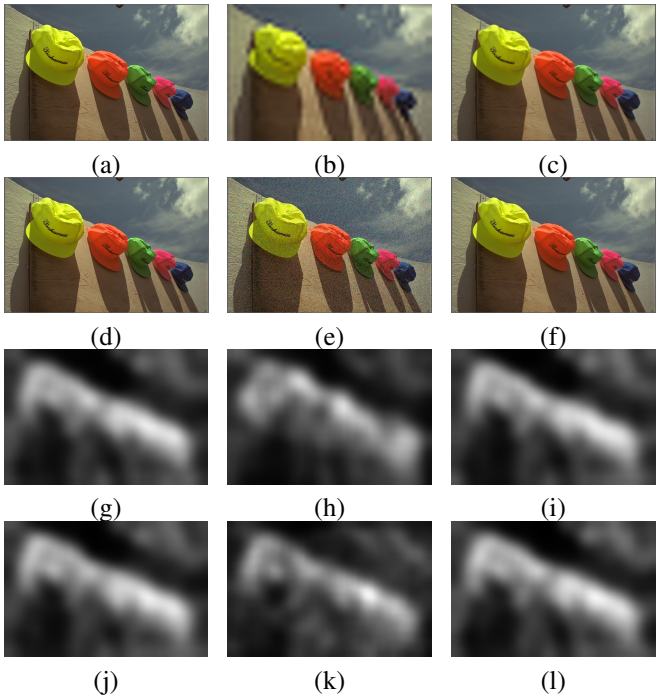

Fig. 1. Examples of representative images from LIVE [13] and the associated saliency map: (a)original image; (b)fast fading; (c)Gaussian blur; (d)JP2K compression; (e)white noise; (f)JPEG compression; (g)-(l) saliency maps of (a)-(f).

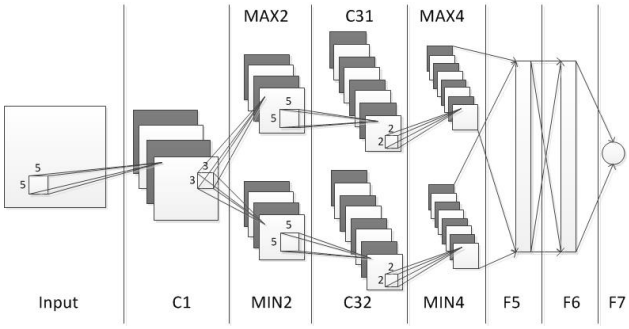

Fig. 2. The architecture of our CNN

$$\widetilde{I}(x,y) = \frac{I(x,y) - m(x,y)}{\sigma(x,y) + c}$$
$$m(x,y) = \frac{1}{(2P+1)(2Q+1)} \sum_{p=-P}^{P} \sum_{q=-Q}^{Q} I(x+p, y+q) \quad (1)$$
$$\sigma(x,y) = \sqrt{\sum_{p=-P}^{P} \sum_{q=-Q}^{Q} (I(x+p, y+q) - m(x,y))^2}$$

where c is a small positive constant avoiding division-by-zero. Parameters $P = 3$ and $Q = 3$ are the normalized window sizes.

We establish a CNN architecture, which contains: two convolutional, two pooling and three full-connected layers, as shown in Fig 2. The first convolutional layer(C1) filters the

32×32 input patch with 16 kernels of size 5×5 with a stride of 1 pixel and produces 16 feature maps of size 28×28. A max and min pooling layers(MAX2, MIN2) come after the convolutional layer. For each feature map, we take the max and min over 3×3 with a stride of 2. After the pooling, we get two $14 \times 14 \times 16$ feature maps. The second convolutional layer(C31, C32) filters the input with 32 kernels of size 5×5 with a stride of 1 pixel. The outputs of layer C31 and layer C32 are both 32 feature maps of size 10×10. They are followed by a max or min pooling layer(MAX4,MIN4) respectively, which takes the max or min over 2×2 with a stride of 2. Two full-connected layers(F5, F6) of 800 nodes are connected to all outputs of max and min pooling layers. The last full-connected layer(F7) only contains one node and gives the predicted quality score. It works as a simple linear regression. We use the logistic neurons on the convolutional layers (C1,C31,C32), but we apply the ReLUs on the pooling and full-connected layer.

Let $f(x_i : para^t)$ denote the estimated score of the input patch x_i with our CNN's parameters $para$, and y_i denotes the ground truth score. The loss function is:

$$\begin{aligned} Loss &= \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} \|y_i - f(x_i; para)\|_{l_2} \\ para' &= \min_{para} Loss \end{aligned} \tag{2}$$

where N_{total} is the number of total image patches.

We update the parameters $para$ with momentum as follows:

$$\begin{aligned} \Delta para^t &= \gamma \Delta para^{t-1} + \epsilon^t para' \\ para^t &= para^{t-1} + \Delta para^t \end{aligned} \tag{3}$$

where $para^t$ is the parameters at epoch t. The momentum γ stays at 0.9 as the training process. The learning rate ϵ^t for the three convolution layers (C1) starts at a value of 0.0001 and decreases to 0.00001 when $t > 50$. While, the learning rate for other layers is 0.00001. We find out this setting could almost save half of the training time and achieve better performance.

Due to the fact that each of our training images shares the homogeneous distortions, we are able to take non-overlapping 32×32 patches from the large images and regard the source image's quality score as the ground truth score of each patch. In this way, we can get a much larger number of training samples, which is important for CNN training. For test stage, the CNN network is likely to work as a local distortion-based pooling method by operating on the small patches, which has proved to improve the performance of the IQA algorithm in [14][15]. We are going to fine tune the final quality score for better IQA performance, by combining the scores of every patches by the weighted average. It is generally claimed that the salient regions can easily attract the human attentions at first glance and affect the human judgement of the images. In this work, we take the influence of saliency into consideration with examining the saliency maps of the test images. As a result, the patches in salient region will contribute to the whole image quality score greatly with large weight values. We get the saliency map S_{final} by FES, and compute the weight w_i of image patch h_i according to the saliency map S_{final}:

$$w_i = \sum_{j=1}^{32 \times 32} S_{final}(j) \tag{4}$$

where pix j belongs to the patch h_i . It should be noted that the weights $\{w_i\}_{i=1}^{N_p}$ are then normalized between 0 and 1.

Finally, we get the predicted quality score Z of the test image by computing:

$$Z = \frac{\sum_{k=1}^{N_p} w_k \times z_k}{\sum_{k=1}^{N_p} w_k} \tag{5}$$

where N_p is the number of the patches sampled from the test image, and z_k is the predicted quality score by our CNN network.

III. EXPERIMENTAL RESULTS

The comparative experiments are implemented on popular LIVE database [13] and TID 2008 [16]. The LIVE database consists of 29 reference images and 779 distorted images with five distortions: JPEG2000 compression (JP2K), JPEG compression (JPEG), White Gaussian (WN), Gaussian blur (BLUR) and a Rayleigh fast-fading channel simulation (FF). Differential Mean Opinion Scores (DMOS) in the range of $[0, 100]$ represent the subjective quality of the image. Higher DMOS corresponds to lower image quality. The TID2008 database contains 25 reference images and 1700 distorted images with 17 different types of distortions. Mean Opinion Score (MOS) in the range of $[0, 9]$ is associated with each image. Higher MOS indicates higher image quality.

The two measures of Spearman Rank Order Correlation Coefficient (SROCC) and Linear Correlation Coefficient (LCC) are used to evaluate the performance of IQA algorithm. Spearman Rank Order Correlation Coefficient assesses how well the relationship between two variables can be described using a monotonic function and Linear Correlation Coefficient measures the degree of linear dependence between two variables.

A. Performance Evaluation on LIVE Dataset

The experiments are non-distortion-specific, i.e. our network is trained with all five distortions together. The results are shown in Table I and Table II, as compared with the current typical IQA algorithms. We randomly select 80% reference images and their corresponding distorted images as the training set and the remaining 20% as the test set. Our results are obtained as the average score of 100 train-test iterations. Among all the NR-IQA methods, the method with top-ranked performance is highlighted in bold. It is significant that our method achieves the best performance on JPEG, FF distortion types. Meanwhile, our method is competitive in terms of SROCC and LCC if all of five distortions are considered. Especially, the proposed blind IQA model even outperforms state-of-the-art FR-IQA methods, such as FSIM, VSI and ADD-SSIM.

TABLE I
SROCC ON THE LIVE DATABASE

SROCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.870	0.885	0.942	0.763	0.874	0.866
SSIM [2]	0.939	0.946	0.964	0.907	0.941	0.913
FSIM [3]	0.970	0.981	0.967	0.972	0.949	0.964
VSI [17]	0.960	0.976	0.984	0.953	0.943	0.952
ADD-SSIM [4]	0.967	0.984	0.984	0.968	0.951	0.965
DIIVINE [5]	0.913	0.910	**0.984**	0.921	0.863	0.916
BLIINDS-II [6]	0.929	0.942	0.969	0.923	0.889	0.931
BRISQUE [8]	0.914	0.965	0.979	0.951	0.877	0.940
CORNIA [7]	0.943	0.955	0.976	**0.969**	0.906	0.942
Le's CNN [9]	0.952	0.977	0.978	0.962	0.908	0.956
Li's CNN [10]	**0.964**	0.935	0.988	0.941	0.945	0.958
our method	0.955	**0.981**	0.953	0.927	**0.983**	**0.968**

TABLE II
LCC ON THE LIVE DATABASE

LCC	JP2K	JPEG	WN	BLUR	FF	ALL
PSNR	0.873	0.976	0.926	0.779	0.870	0.856
SSIM [2]	0.921	0.955	0.893	0.893	0.939	0.906
FSIM [3]	0.91	0.985	0.978	0.978	0.912	0.960
VSI [17]	0.965	0.981	0.966	0.942	0.938	0.948
ADD-SSIM [4]	0.975	0.985	0.974	0.970	0.954	0.959
DIIVINE [5]	0.922	0.921	0.988	0.923	0.888	0.917
BLIINDS-II [6]	0.935	0.968	0.980	0.938	0.896	0.930
BRISQUE [8]	0.922	0.973	0.985	0.951	0.903	0.942
CORNIA [7]	0.951	0.965	0.987	**0.968**	0.917	0.935
Le's CNN [9]	0.953	0.981	0.984	0.953	0.933	0.953
Li's CNN [10]	**0.978**	0.977	**0.993**	0.945	0.960	0.966
our method	0.961	**0.989**	0.954	0.948	**0.987**	**0.969**

TABLE III
SROCC AND LCC OBTAINED BY TRAINING ON THE LIVE DATABASE AND TESTING ON TID2008 DATABASE

	SROCC	LCC
CORNIA [7]	0.890	0.880
BRISQUE [8]	0.882	0.892
Le's CNN [9]	0.920	0.903
our method	0.922	0.916

B. Cross-database test

In this section, the experiment is planed to test the robustness of our method. We train our CNN on LIVE dataset and test its performance on TID2008 dataset. There are four distortion types are shared by LIVE and TID2008, including JP2K, JPEG, WN and BLUR. The value range of DMOS in LIVE is from 0 to 100. But the MOS values in TID2008 fall in the range from 0 and 9. We perform a nonlinear mapping with a logistic function, which is usually applied to transform the quality score gained by FR-IQA methods into a specified value range. We randomly repeat to select 80% images of TID2008 to estimate the parameters of the logistic function and test the performance on the remaining 20% images 150 times. We report the results in Table III. Our algorithm is comparable to the other state of the art methods.

IV. CONCLUSION

In this paper, we have proposed a no-reference image quality assessment algorithm, which introduces saliency detection to modify the CNN structure. Our method first estimates the quality score of each image patch by CNN network. Next, we compute the saliency map using free energy based neural theory. The salient patches is assigned to be large weights in the pooling stage. Finally, we get the predicted score of the whole image by computing the weighted averaging scores of all the image patches. In contrast to modern NR-IQA methods, it emphasizes the fact that the distortions of those image patches attracting much human's attention affect visual quality dominantly. Experimental results on LIVE and TID2008 databases are provided to confirm the superiority of our algorithm, achieving better performance with humans visual perception.

Acknowledgements. The work was supported by S-tate Key Research and Development Program (2016YF-B1001003), NSFC (61527804, 61521062, 61502301), STCSM (14XD1402100) and the 111 Program (B07022).

REFERENCES

[1] H. R. Sheikh, A. C. Bovik, and G. 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.

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

[3] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[4] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *Trans. Broadcasting*, 2016.

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

[6] M. Saad, A. 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.

[7] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *CVPR*, 2012, pp. 1098–1105.

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

[9] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neutral networks for no-reference image quality assessment," in *CVPR*, 2014.

[10] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, and G. Xie, "No-reference image quality assessment using prewitt magnitude based on convolutional neural networks," *Signal, Image and Video Processing*, pp. 1–8, 2015.

[11] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Rev. Neuroscience*, vol. 11, pp. 127–138, 2010.

[12] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Visual saliency detection with free energy theory," *IEEE Singal Processing Letter*, vol. 22, no. 10, pp. 1552–1555, 2015.

[13] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik, *LIVE image quality assessment database release 2. Online*, http://live.ece.utexas.edu/research/quality.

[14] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 192–201, 2009.

[15] K. Gu, G. Zhai, X. Yang, and W. Zhang, "An efficient color image quality metric with local-tuned-global model," in *ICIP*, 2014, pp. 506–510.

[16] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radio Electronics*, vol. 10, pp. 30–45, 2009.

[17] L. Zhang, Y. Shen, and H. Yi, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.