Direct Application of Convolutional Neural Network Features to Image Quality Assessment

Xianxu Hou[1], Ke Sun[1], Bozhi Liu[1], Yuanhao Gong[1], Jonathan Garibaldi[2], Guoping Qiu[1,2]

[1] *College of Information Engineering, Shenzhen University, Shenzhen, China*
[2] *School of Computer Science, University of Nottingham, Nottingham, United Kingdom*

***Abstract*—We take advantage of the popularity of deep convolutional neural networks (CNNs) and have developed a very simple image quality assessment method that rivals state of the art. We show that convolutional layer outputs (deep features) of a CNN compute the local structural information of spatial regions of different sizes in the input image. The learned convolutional kernels contain a much richer set of weights thus capturing much more local structural information than hand crafted ones. As the deep features learned from large datasets already contain very rich multi-resolutional structural image information, they can be directly used to calculate visual distortion of an image and it is not necessary to introduce further complicated computational process. We will present experimental results to demonstrate that this is indeed the case, and that simple cosine distance of the deep features is as good as state the art methods for full reference image quality assessment.**

***Index Terms*—CNN, deep features, image quality assessment**

I. Introduction

Deep convolutional neural networks have been demonstrated to be able to achieve state of the art performance in many computer vision tasks and learn remarkably powerful (deep) features for representing visual information. Although still very difficult to interpret, the deep features are able to generalize surprisingly well from one task to another under the framework of transfer learning. In addition to high level visual recognition problems, deep features have been also successfully used to solve different image transformation tasks like style transfer and super-resolution [1] and image generation [2]. In particular, perceptual loss functions defined in the deep feature space have been successfully used to measure high-level image similarities.

In this paper, we demonstrate another potentially very useful image processing application of the deep features and through which we attempt to gain a better understanding of the nature of deep features. We show that the convolutional layers' outputs actually can capture multi-resolution structrual information of a local spatial region of different sizes in the input image. As the learned convolutional kernels contain a much richer set of weights thus capturing more local structural information than hand-crafted features such as those used in SSIM [3], it is expected that using the deep features will outperform hand crafted features. Also, as the deep features have already contained a rich set of multi-resolutional structural image information, they can be directly used to calculate the visual distortion of an image and it is not necessary to introduce more complicated computational process.

II. Related Work

The simplest objective image quality metric is Peak Signal to Noise Ratio (PSNR), which represents the ratio between the maximum possible power of a signal and the power of distortion noise, which is based on per-pixel mean squared error (MSE) between the original image and the distorted image. In general, a higher PSNR value usually indicates better image quality, however in some cases it could not generalize well to perceived visual quality [4], [5], [6], [7]. In particular, for image reconstruction or generation tasks the output images tend to be very blurry when compared to natural images.

Another widely used is SSIM index [3], which is a more perceptual-based approach by comparing the structures of the reference and the distorted signals instead of absolute point-wise error. SSIM is based on the assumption that the human visual system is highly adapted to extract structural information from visual information and considers image degradations as perceived changes in structural information variation.

Additionally complex wavelet structural similarity (CW-SSIM) index [8] is proposed as an extension of the SSIM to the complex wavelet domain as a general purpose image similarity metric. The key idea is that certain image distortions can lead to consistent phase change in the local wavelet coefficients, and the structural content of the image doesn't change with a consistent phase shift of the coefficients.

The image quality assessment indexes mentioned above are purely based on human-crafted features, and the key insight is to design better method by incorporating the structural information to measure the inconsistency (or similarity) of a

978-1-5386-4458-4/18/$31.00 ©2018 IEEE

Fig. 1. 64 convolutional kernels of size $3 \times 3 \times 3$ learned by the first convolutional layer of pretrained VGGNet [13].

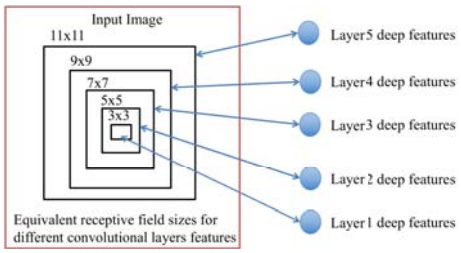

Fig. 2. Deep features at different hidden layers and their corresponding equivalent receptive field size. In essence, each hidden layer unit (deep feature) computes the local pixel structure of its corresponding receptive field region.

given pair of images. In recent years, there have been a number of authors attempted to develop deep learning based image quality assessment methods [9], [10], [11], [12]. Whilst these deep learning based methods use the deep neural network in some sophisticated manner, we argue that the convolutional layer outputs of the deep neural network can be directly used to measure image quality.

III. Direct Deep Feature Based Image Quality Assessment

A. Rationale

The success of SSIM and CW-SSIM is mainly through incorporating structural information of an image either in pixel or complex wavelet domain. In parallel, the fact that natural image signals are highly structured has been exploited by the deep CNNs to build state of the art object recognition models. In particular, CNNs use a set of filters or kernels like $3 \times 3 \times 3$ (i.e. 3 pixel width and height, and 3 channel depth) to spatially convolve across the width and height of the input volume to compute dot products between the entries of each filter and the corresponding pixels of the input image. Thus, each output of a convolutional layer is a weighted combination of the pixels in a local region. Conceptually these kind of convolution operations are able to capture the structural information due to the local connectivity. Exactly what it is capturing depends on the weights of the filters. Previous works on image recognition tasks have shown that deep convolutional neural networks can learn interpretable filters based on a large-scale dataset. Fig. 1 shows the 64 convolutional kernels of size $3 \times 3 \times 3$ learned by the first convolutional layer of VGGNet [13] trained on ImageNet [14]. We can see that there are a variety of meaningful frequency and orientation-selective kernels.

At each hidden layer, a deep feature covers a different size of an input receptive field as illustrated in Fig. 2. Equivalently, a deep feature in convolutional layer 1 computes the local structure of a 3x3 spatial block. Inspecting the convolutional weights in Fig. 1, it is not difficult to see that some filters will compute the local luminance or weighted average (when all weights have the same signs), some will compute local contrasts along different directions (when the weights have both positive and negative values). For higher layers, each hidden unit covers a (larger and larger) local block in the input image and it can be regarded as a multi-resolutional image information.

From above reasoning, it is very clear that the hidden unit outputs of a CNN computes various visually meaningful local features such as the local luminance values and contrasts of various types depending on the weight values. As a pre-trained CNN (such as the VGGNet) will have a variety of filtering kernels, it is very clear that the deep features of a CNN are in fact computing the local structural information of the image. Based on previous successful image quality measurements based on hand-crafted features for computing the local structural information of an image such as the well-known SSIM [3] method, deep features learned from large dataset will contain richer local structure information and can be directly apply to measure image quality and be expected to outperform hand-crafted features. We will demonstrate this is indeed the case.

B. Deep Feature Based Image Quality Assessment

Based on the above rationale, we directly use the deep features of a pre-trained CNN to develop the deep feature based image quality assessment (DFB-IQA). Our DFB-IQA index tries to provide a good approximation to perceived image distortion by incorporating image pixel spatial correlation and object structure information, which are implicitly captured by the learned filters through a large-scale image dataset training instead of human engineering. In other words, the proposed DFB-IQA index measures the structural similarity of two images in the learned feature space. The final DFB-IQA index is defined as the cosine distance of feature representation (flattened as a vector) of a reference image x and the distorted image $\tilde{x}$ as follows:

$$DFB\_IQA(x, \tilde{x}) = \frac{\boldsymbol{\Phi_i(x)} \cdot \boldsymbol{\Phi_i(\tilde{x})}}{\|\boldsymbol{\Phi_i(x)}\|_2 \, \|\boldsymbol{\Phi_i(\tilde{x})}\|_2} \tag{1}$$

where $\boldsymbol{\Phi_i(x)}$ and $\boldsymbol{\Phi_i(\tilde{x})}$ are the i^{th} hidden activations when feeding the input image x and distorted image to pretrained VGGNet work Φ.

IV. Experiments

A. Testing dataset

In this work, we evaluate the proposed DFB-IQA index on Release 2 version of LIVE Image Quality Assessment Database [15], [3], [16], which is a large dataset including 30 reference images and 779 distorted versions with JPEG

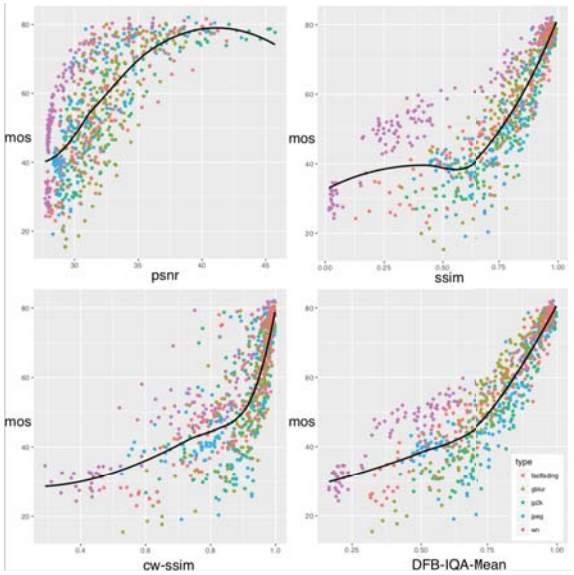

Fig. 3. Scatter plots of mean opinion scores (MOS) versus PSNR, SSIM, CW-SSIM and DFB-IQA. DFB-IQA-Mean is the average result of different image quality indexes based on 5 different convolutional layer.

compression, JPEG2000 compression, gaussian blur, white noise and fast fading rayleigh. The difference mean opinion score (DMOS) (1 to 100) value for each distorted image is provided according to human rating. In our experiments, we convert difference mean opinion score (subtracted by 100) to *mean opinion score* (MOS) for comparison.

B. Results

We conduct experiments to compare the performance of the proposed DFB-IQA index with PSNR, CW-SSIM[8], MSSIM[3], FSIMc[17], BIFS[18], IGM[19] and DeepSim[10]. On the other hand, due to the hierarchical architecture of deep convolutional neural network we also investigate the effect of different level features on the performance of DFB-IQA.

In order to get a more comprehensive comparison of different image quality assessment algorithms, we show the correlation map of MOS versus different model predictions with different types of distortions. The DFB-IQA-Mean index is based on the average of the 5 convolutional layers. From Fig. 3, we can see that the proposed DFB-IQA performs quite well in this test and behaves consistently with MOS. Though SSIM performs well for a single type of distortion, it cannot generalize well for cross-distortion testing. On the contrary, the scatter plot of DFB-IQA index is more compact, demonstrating that it provides remarkably good prediction of the mean opinion scores.

In addition, we investigate the effect of different level features on the performance of DFB-IQA for subjective MOS prediction. In Fig. 4, we show the scatter plots of MOS versus DFB-IQA by using layer conv1_1, conv2_1, conv3_1, conv4_1, conv5_1 and the average results of the 5 layers. We can observe the trend that DFB-IQA indexes constructed from

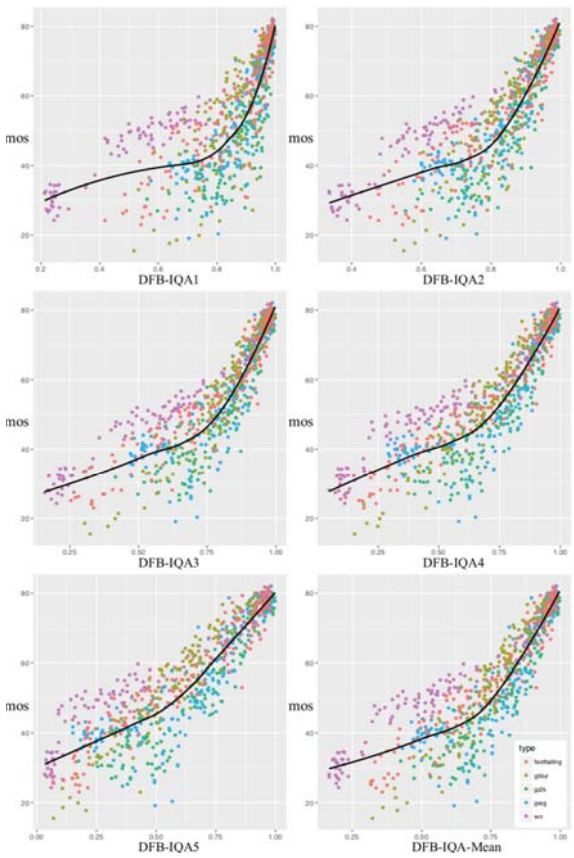

Fig. 4. Scatter plots of mean opinion scores (MOS) versus DFB-IQA indexes calculated by different convolutional layers. DFB-IQA-Mean is the average result of different image quality indexes based on 5 different convolutional layer.

higher layers behave more consistently with mean opinion scores when applied to distorted images created from different types of distortions. Specifically, there is an obvious divergence in the scatter plot for *white noise* distortion in lower layers (conv1_1 and conv2_1), while it shows a more compact distribution for all types of distortions by using higher layers (conv4_1 and conv5_1). This could be explained that higher layers of the convolutional neural network mainly capture the high-level content in terms of objects and overall structures with bigger receptive field size instead of limiting to detailed pixel information. From Fig. 2, it can be easily understood because these higher layers cover a larger receptive field in the input image. Thus, the difference between different types of distortions will not be quite significant, what matters is the degradation degree for each type of distortion method.

We further conduct quantitative experiments to measure the performances of different image quality assessment algorithms, i.e, root mean squared error (RMSE), CW-SSIM [8], MSSIM [3], FSIMc [17], BIFS [18], IGM [19] DeepSim [10]. We adopt three widely used criteria for evaluating the performance of different IQA methods, i.e. the Pearson's linear correla- tion coefficient (PLCC), Spearman's rank-order correlation coefficient (SRCC), and Kendall's rank-order correlation

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT IMAGE QUALITY ASSESSMENT INDEXES.

Method	PLCC	SRCC	KRCC
RMSE	0.742	0.677	0.500
CW-SSIM[8]	0.690	0.788	0.592
MSSIM[3]	0.668	0.646	0.463
FSIMc[17]	0.818	0.864	0.673
BIFS[18]	0.884	0.855	0.667
IGM[19]	0.886	0.856	0.668
DeepSim[10]	0.885	0.877	0.683
DFB-IQA1	0.703	0.833	0.636
DFB-IQA2	0.816	**0.886**	**0.701**
DFB-IQA3	0.849	**0.914**	**0.740**
DFB-IQA4	0.866	**0.913**	**0.737**
DFB-IQA5	**0.890**	**0.906**	**0.727**
DFB-IQA-Mean	0.825	**0.890**	**0.708**

coefficient (KRCC) between the predicted quality scores and MOS.

The evaluation results are shown in Table I. All the results are calculated based on all types of distortions in the Release 2 version of LIVE Image Quality Assessment Database [16]. The DFB-IQA1 to DFB-IQA5 are calculated based on the features extracted from the conv1_1 to conv5_1 layer and DFB-IQA-Mean is the average result of image quality indexes of 5 different convolutional layers. We can see that our proposed DFB-IQA methods work well in general and achieve new state of the art on this dataset. It is demonstrated that DFB-IQA index has a better consistency and more stable correlation with subjective mean opinion scores in cross-distortion evaluation. In addition, when compared the performance with different convolutional layers, it is clear that the image quality indexes computed based on middle and high level features perform better with higher PLCC, SRCC and SRCC values. We can conclude that the learned representation from deep convolutional neural network can effectively capture the perceived changes of image quality degradation. What's more, it could give us a few insights for the learned features from the perspective of image quality assessment that why they can be used to construct perceptual loss functions for different image transformation tasks [1], [20], [21], [2]. Finally, it is interesting to observe that our direct application of layer 5 deep features to compute the image quality performs better than a recent more elaborated deep learning based method [10] in this dataset, further demonstrating correctness of the technical rationale of direct application of deep features for image quality assessment.

V. CONCLUDING REMARKS

In this paper, we propose the use of learned features to design image quality assessment metrics under deep learning framework. The key insight is the capability of pretrained deep convolutional neural network to incorporate structural and perceptual image information in its hidden features, which can be directly used as an alternative to hand crafted features for the design of image quality measures. Our experiments demonstrate the effectiveness of the proposed DFB-IQA index with respect to subjective mean opinion scores prediction in cross-distortion settings. DFB-IQA can be seen as an alternative or complementary to traditional approaches like SSIM, which tries to incorporate structural information by human engineering while DFB-IQA seeks to extract image pixel spatial correlation and object structure information from learned features based on image recognition tasks.

REFERENCES

[1] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *arXiv preprint arXiv:1603.08155*, 2016.

[2] X. Hou, L. Shen, K. Sun, and G. Qiu, "Deep feature consistent variational autoencoder," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 1133–1141.

[3] Z. 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.

[4] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal processing*, vol. 70, no. 3, pp. 177–200, 1998.

[5] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[6] Y. Gong and I. F. Sbalzarini, "Curvature filters efficiently reduce certain variational energies," *IEEE Transactions on Image Processing*, vol. 26, pp. 1786–1798, 2017.

[7] Y. Gong, "Mean curvature is a good regularization for image processing," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.

[8] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: a new image similarity index," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 18, no. 11, p. 2385, 2009.

[9] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks." *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2018.

[10] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "Deepsim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, 2017.

[11] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[12] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal Image & Video Processing*, no. 3, pp. 1–8, 2017.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[15] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[16] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database release 2."

[17] L. Zhang, L. 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.

[18] F. Gao and J. Yu, "Biologically inspired image quality assessment," *Signal Processing*, vol. 124, pp. 210–219, 2016.

[19] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 43–54, 2013.

[20] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *arXiv preprint arXiv:1603.03417*, 2016.

[21] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.