Full-Reference Image Quality Assessment Using Neural Networks

Sebastian Bosse*, Dominique Maniry*, Klaus-Robert Müller†, *Member, IEEE*,
Thomas Wiegand*†, *Fellow, IEEE*, and Wojciech Samek*, *Member, IEEE*
*Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany
†Department of Electrical Engineering and Computer Science, Technical University of Berlin, Berlin, Germany

***Abstract*—This paper presents a full-reference (FR) image quality assessment (IQA) method based on a deep convolutional neural network (CNN). The CNN extracts features from distorted and reference image patches and estimates the quality of the distorted ones by combining and regressing the feature vectors using two fully connected layers. Experiments are performed on the LIVE and TID2013 databases and correlations comparable or superior to state-of-the-art IQA methods are achieved.**

I. Introduction

Images and video are ubiquitous today. The share of bits representing digital visual signals is huge and even growing. For transmission or storage, usually lossy compression schemes are applied that introduce distortions into these signals. As humans are typically the ultimate receiver of these signals, it is crucial for such a quality (or: distortion) metric to relate to human visual perception and to be able to predict the visual distortion perceived by humans. As at sender or encoder side of a transmission system the undistorted reference signal is available, common transmission and coding schemes allow for the use of FR IQA. Perceptually relevant FR IQA typically either aims at modeling various processing mechanisms of the human visuals system (HVS) [1], [2] or at applying general assumptions on the general properties of the HVS to infer perceptually relevant features from images [3], [4], [5] in order to estimate perceived quality. In [6], a neural network is used to take several of those IQA outcomes and improve on prediction performance by combining them. Recently, methods applying a third strategy have been proposed; these methods operate purely data driven and do not rely on explicit assumptions about the HVS or perceptual image features. In [7], image patches are k-means clustered in order to learn a general image representation that is used in combination with a support vector machine (SVM) to predict image quality. A linear SVM regression is used to learn a set of linear image features for no-reference (NR) IQA in [8]. An extension of this approach is presented in [9], where only object-like patches of an image are input to [8]. In [10], a five layer CNN is trained to jointly learn features to be extracted and a regression function to estimate the quality of luminance normalized image patches in a NR context. In this paper, a CNN-based method is proposed for FR IQA, thus it uses a different architecture than [11] by running two CNNs in parallel for feature extraction and regression. Our method uses a deep CNN with 16 layers of convolution, nonlinearity and pooling operations. In contrast to [10], no luminance normalization is applied. In order to consider visual saliency and local distortion sensitivity, we apply a weighted average patch aggregation.

In the next section we describe the neural network based FR IQA method. In Section III our method is evaluated on the LIVE [4] and TID2013 [12] image quality databases. We conclude in Section IV with a brief discussion.

II. Proposed Method

The proposed system is applied to individual unpreprocessed 32×32 RGB patches cropped from reference and distorted images. An imagewise estimate of perceived quality can then be obtained by suitable pooling, e.g. by weighted or unweighted averaging, of the patchwise estimated quality. The high-level layout of the proposed method of patchwise estimation of perceived quality can be subdivided in four modules, as shown in Fig. 1.

In a first step, features are extracted from reference and distorted image patches, respectively, by two CNNs that share identical weights. For feature extraction, we choose an architecture with many layers inspired by [13]. The two feature vectors f_r and f_d that are extracted from the reference image and the distorted image by the CNNs are fused in a second step by concatenating f_r, f_d and $f_r - f_d$. After feature fusion, in the third step the fused feature vector is input to a fully connected neural network regressing it to a patch quality estimate. The fourth step aggregates the patch quality estimates to an image quality estimate. Fig. 1 outlines the architectural details of the network.

In order to account for the influence of local image and noise properties on the quality of full images, we propose a weighted average aggregation of patchwise estimated local quality to global quality. To achieve this, two additional fully connected layers are added to the network that run in parallel to the last two layers (the regression part) of the proposed network and that are of the same shape.

The output α_i of these layers can be used to weight the estimated local quality of the corresponding patch i. Activating α_i by a Rectified Linear Unit (ReLU) and using a small stability factor ϵ ensures it to be positive and non-zero with $\alpha_i^* = max(0, \alpha_i) + \epsilon$. The global quality of a full image can then be calculated as

$$q = \frac{\sum_i^{N_p} \alpha_i^* y_i}{\sum_i^{N_p} \alpha_i^*},$$

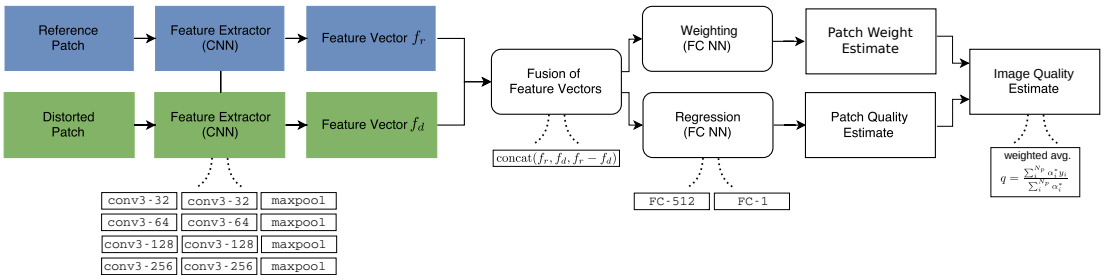

Fig. 1: Layout of the proposed neural network.

where given image is subdivided in N_p patches.

Each patch is annotated with a quality label q_t taken from the full images that the respective patch was cropped from. For end-to-end training, the error of the globally estimated quality $E = |q - q_t|$ is minimzied.

III. Experiments

For evaluating the performance of the proposed method, the CNN is trained on 10 random train-test splits strictly based on separating source reference images: For LIVE, 6 source reference images and corresponding distorted versions were randomly chosen for testing, 6 of the remaining source reference images and corresponding distorted versions were randomly chosen for validation and the remaining 17 source reference images and corresponding distorted versions were used for training. For TID2013, the test set was based on 5 source reference images and corresponding distorted versions, validation set on 5 and training set on 15. In each epoch, 32 random patches are sampled from each image from the training set. Models are trained for 3000 epochs. For evaluation on LIVE, 2048 patches, for TID2013, 1024 patches, are extracted per image.

Table I compares the proposed methods with PSNR, SSIM[3] and FSIM [5], CNNM [6] and, given its high performance, a state-of-the-art NR IQA approach (SOM) [9] in terms of Pearson Linear Correlation Cofficients (LCC), Spearman Rank Order Coerrelation Coefficient (SROCC) and Kendall Rank Order Coerrelation Coefficient (KROCC) as far as reported.

	LIVE		TID2013	
Method	LCC	SROCC	SROCC	KROCC
PSNR	0.856	0.866	0.64	0.47
SSIM [3]	0.906	0.913	0.64	0.65
FSIM [5]	0.962	0.964	0.8	0.64
SOM [9]	0.962	0.964	-	-
CNNM [6]	-	-	0.93	0.77
Proposed	0.968	0.959	0.94	0.78

TABLE I: Performance of the proposed methods on LIVE and TID2013 compared to state-of-the-art IQA methods.

IV. Conclusion

We applied a deep CNN that combines features extraction and regression in one framework to tackle the problem of FR IQA in a purely data-driven approach. In order to deal with local differences of perceived distortions, a weighted aggregation is proposed. Although no assumption about the HVS or image statistics is made and only 32 randomly choosen patches are considered per image, the proposed method achieves performances comparable or superior to state-of-the-art IQA methods. In future work we will study the influence of the CNN architecture (e.g. depth) on FR IQA and investigate what features are actually learned by the network. Furthermore we will explore the performance of our method in cross-dataset and cross-distortion scenarios.

References

[1] A.B. Watson, R. Borthwick, and M. Taylor, "Image quality and entropy masking," in *SPIE Proc.*, 1997, vol. 3016, pp. 1–11.

[2] J. Lubin, "A human vision system model for objective picture quality measurements," *Int. Broadcast. Conv.*, pp. 498–503, 1997.

[3] 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.

[4] 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. 3440–51, nov 2006.

[5] L. Zhang, L. Zhang, X. 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.

[6] V. Lukin, N. Ponomarenko, O. Ieremeiev, K. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," vol. 9394, pp. 93940K–93940K–12, 2015.

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

[8] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-Time No-Reference Image Quality Assessment Based on Filter Learning," *2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 987–994, 2013.

[9] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2394–2402, 2015.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional Neural Networks for No-Reference Image Quality Assessment," in *Comput. Vis. Pattern Recognit. (CVPR), 2014 IEEE Conf.*, 2014, pp. 1733–1740.

[11] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016.

[12] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Color Image Database TID2013 : Peculiarities and Preliminary Results," *Vis. Inf. Process. (EUVIP), 2013 4th Eur. Work.*, pp. 106–111, 2013.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Iclr*, pp. 1–14, 2015.