A DEEP NEURAL NETWORK FOR IMAGE QUALITY ASSESSMENT

Sebastian Bosse[1], *Dominique Maniry*[1,2],
Thomas Wiegand[1,2], Fellow, IEEE, and *Wojciech Samek*[1], Member, IEEE

[1] Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany
[2] Department of Electrical Engineering, Technical University of Berlin, Germany.

ABSTRACT

This paper presents a no reference image (NR) quality assessment (IQA) method based on a deep convolutional neural network (CNN). The CNN takes unpreprocessed image patches as an input and estimates the quality without employing any domain knowledge. By that, features and natural scene statistics are learnt purely data driven and combined with pooling and regression in one framework. We evaluate the network on the LIVE database and achieve a linear Pearson correlation superior to state-of-the-art NR IQA methods. We also apply the network to the image forensics task of decoder-sided quantization parameter estimation and also here achieve correlations of $r = 0.989$.

Index Terms— Convolutional neural network, no reference image quality assessment, quality perception, blind qp estimation, image forensics,

1. INTRODUCTION

Enormous amounts of visual media are ubiquitous today and considerable time and resources are brought up to ensure that the captured, transmitted or presented media is of satisfactory quality. For most applications the human visual system is the ultimate receiver of visual data. Thus, image quality assessment (IQA) aims at providing methods to measure the quality of visual data in a way that is consistent with human perception. Considering the ease of this task for humans, judging the visual perceptual quality is a surprisingly hard problem for computers especially when it comes to blind or no-reference (NR) IQA and researchers are working on it since several decades. While in full reference (FR) IQA the algorithm not only has full information about the distorted, but also about the undistorted reference image, NR IQA estimates the quality of an image without access to any information about the reference image.

Recently, several general purpose approaches not assuming specific distortions have been proposed and gained attention [1]. Distortion specific NR IQA is considered as an easier problem. Thus, in [2] a 2-stage framework is presented, where in a first step the specific distortion of an image is identified based on scene statistics. In a second stage, distortion specific image statistics are used to estimate the image quality. For that the coefficients of an oriented multiscale decomposition are modelled by a generalized Gaussian distribution (GGD) and perceived quality is deduced from the deviation of this statistical model. In [3] a GGD model is used for modelling the distribution of the coefficients after applying a discrete cosine transform to the image. Features extracted from the distribution parameters combination are used to estimate the perceived quality. NR IQA in the spatial domain is addressed in [4], where oriented sample intensity differences are modelled by an asymmetric generalized Gaussian distribution and, similar to [3], features are extracted from the distribution parameters in order to predict the perceived quality. A different line of work, not assuming explicit knowledge about natural scene statistics, is presented in [5]. Here, a visual codebook is constructed by k-means clustering. This codebook is used to encode patches of a distorted image and patchwise descriptors are pooled and used as a predictor for perceived quality. An extension of this codebook-based approach is proposed in [6]. In a first step, object-like regions that are assumed to be semantically meaningful are identified and samples of the extracted regions are then input to a similar algorithm as in [5]. Another data driven approach to NR IQA is presented in [7]. In a general convolutional neural network (CNN) framework feature extraction and regression is combined in order to estimate perceived quality. For that, images are subdivided into patches, locally luminance and contrast normalized and fed into a CNN consisting of five layers. The first is a convolution layer with 50 filter kernels. The resulting 50 feature maps are pooled to one max- and one min-feature map. Those two feature maps are fed to two fully connected layers and finally combined by a linear regression to a one dimensional estimate of perceived quality.

The presented study follows the data driven concept of [7]. We propose a deep convolutional neural network (CNN) with 12 weight layers. Deep CNNs have dominated image classification because they are able to automatically learn high-level feature representations. It seems counter-intuitive to extract high-level features for IQA and so far NR IQA systems have used low-level features or relatively shallow neural

This work was supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC, funding mark 01IS14013A.

networks. Nevertheless, we show that extracting high-level features through deep neural networks can lead to superior performance in NR IQA. For that and in contrast to [7] we do not preprocess the image patches. The network architecture is also evaluated in the context of decoder-sided estimation of the quantization parameter in High Efficiency Video Coding (HEVC) relevant in image forensics [8] and also here achieve a considerable prediction performance.

2. PROPOSED METHOD

2.1. Patchwise Training

Because neural networks usually deal with fixed size input, we apply the network to unpreprocessed 32x32 RGB image patches and average the prediction of N randomly sampled patches for each image. During training each image patch is associated with the quality label of the source image and treated as an independent sample (see [7]). We chose a neural network architecture inspired by [9], as it is a fairly simple architecture with many layers and very good results in the ILSVRC image classification challenge[10]. That means we only use 3×3 convolution kernels, the rectifier activation function (ReLU) and reduce the size of feature maps only through max-pooling.

The 12 weight layers are organized as follows (notation from [9]): conv3-32, conv3-32, maxpool, conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool, conv3-256, conv3-256, maxpool, conv3-512, conv3-512, maxpool, FC-512, FC-1.

Except for the last fully-connected layer, all layers are activated through the ReLU activation function. The convolutions are applied with zero-padding, so their output has the same spatial dimensions as their input. All maxpool layers have 2×2 kernels. We apply dropout regularization [11] with a ratio 0.5 to the fully connected layers. This architecture does not contain any design decisions specific to the NR IQA task and thus shows the general applicability of CNNs for image analysis tasks.

Given an image represented by N_p randomly sampled patches and a ground truth quality label of q_t. The quality prediction q is calculated by averaging the CNN output y_i for each patch: $q = \frac{1}{N_p} \sum_i^{N_p} y_i$. During training the mean absolute error (MAE) $E_{patchwise} = \frac{1}{N_p} \sum_i^{N_p} |y_i - q_t|$ is minimized. As in [7], we use MAE as this loss function puts less emphasis on outliers than mean squared error (MSE). The optimization is done through the adaptive learning rate optimizer ADAM [12] with $\alpha = 0.0001$.

2.2. Weighted Average Patch Aggregation

Even though we only consider global image distortions, not every image patch contains the same amount of task relevant information. For instance, flat regions like a blue sky or a white wall might be unaffected by certain distortions, while highly textured regions like bushes might be unaffected by other distortions. Furthermore, distortions in more salient image regions are more detrimental to perceived quality (e.g. blurry foreground vs. blurry background). In the method described above, this problem is only addressed by averaging over a number of image patches. We propose a second architecture that employs a weighted average patch aggregation layer. That way we can train our network end-to-end from raw image to the image quality label and implicitly learn an importance rating for every patch that is processed. To achieve this we add two additional fully connected layers that run parallel to the two last layers of the original architecture and have the same dimensions. This way we get a second output α_i, that we can use for weighting the estimated local quality of a patch. To ensure that the weight is positive and non-zero it is activated through a ReLU and a small stability factor ϵ is added:

$$\alpha_i^* = max(0, \alpha_i) + \epsilon$$

The small constant (e.g. $\epsilon = 0.000001$) is applied to avoid the division by zero when all weights are zero. The weighted average can then be calculated as follows:

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

End-to-end training means the error of the quality estimation $E_{weighted}$ of each image is directly minimized in training:

$$E_{weighted} = |q - q_t|$$

This loss function has the disadvantage that the gradient for improving the quality score is suppressed whenever the weight α_i^* is low. The quality score of patches that consistently get assigned low weights stops improving and thus the low weight gets reinforced. Ideally we want the network to assign a quality score that is as accurate as possible for each patch, regardless of the weight that is assigned. To achieve this we also explore an additional loss formulation where we optimize both the patchwise loss and the weighted imagewise loss jointly with

$$E_{weighted+} = |q - q_t| + \frac{1}{N_p} \sum_i^{N_p} |y_i - q_t|$$

3. EXPERIMENTS

We evaluate and compare the proposed approach in terms of prediction accuracy measured as Pearson product-moment correlation coefficient (LCC) and mean square error (MSE) and in terms of prediction monotonicity measured by Spearman's rank order correlation coefficient (SRCC).

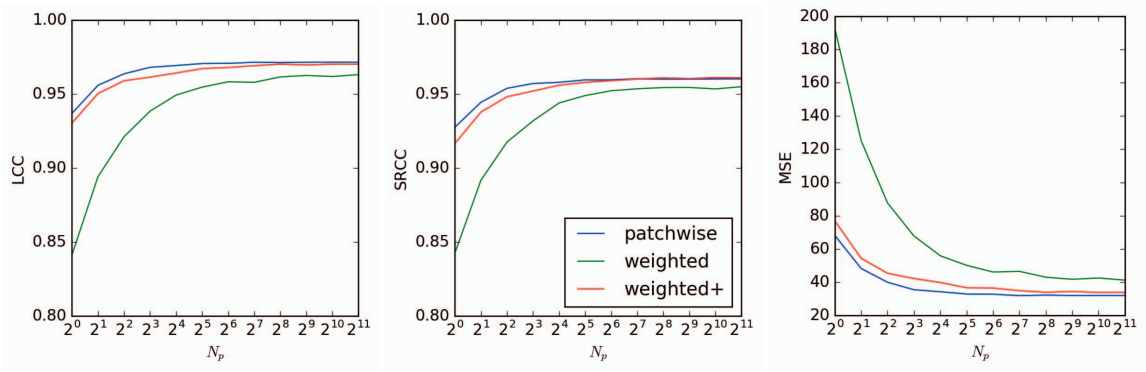

Fig. 1: Performance of the proposed CNN for NR IQA in terms of LCC, SRCC and MSE in dependence of the number of randomly sampled patches.

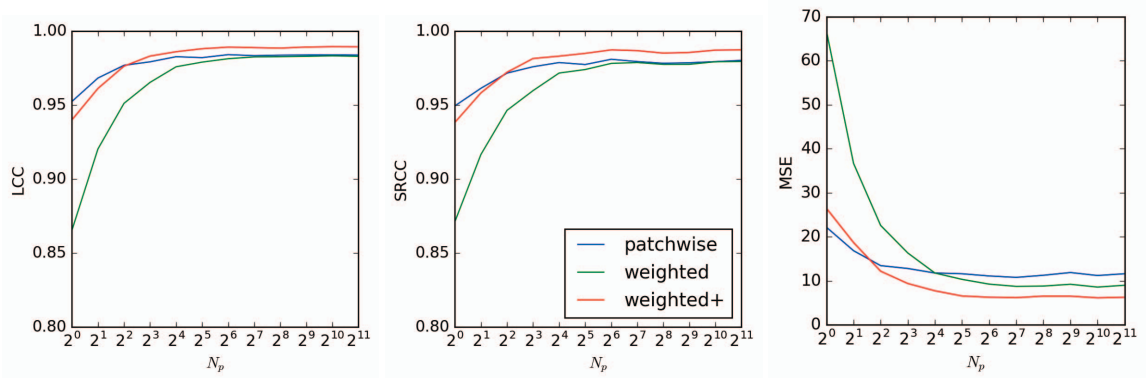

Fig. 2: Performance of the proposed CNN for decoder-sided QP estimation in terms of LCC, SRCC and MSE in dependence of the number of randomly sampled patches.

3.1. No-Reference Image Quality Assessment

To evaluate our approach in a non distortion specific setting, we use the LIVE database [15]. It consists of 981 images that are distorted versions of 29 reference images. The distortions used are varying levels of JPEG compression, JPEG2000 compression, Gaussian blur, white noise and fast fading.

In order to evaluate the performance of the proposed CNN, we train it on 10 random train-test splits. For that 6 reference images and their distorted versions where chosen randomly for testing, 6 other reference image images and their distorted version for validation and the remaining 17 reference images for training. For each epoch we sample 32 random patches from each image from the training set. The models are trained for 3000 epochs, resulting in about 48 million patches used for training; training takes around $6s$ per epoch on a Titan X GPU.

Fig. 1 shows LCC, SRCC and MSE in dependence of the number of patches randomly sampled from the test image. The patchwise averaging outperforms the method of weighted averaging. For the patchwise averaging around 16 patches randomly sampled from the test image are enough to achieve the maximal performance. Using the combined loss function for weighted averaging (*weighted+*) with more than 32 patches shows equal performance with the patchwise approach in terms of SRCC and only slightly lower performance in terms of LCC and MSE. Table 1 lists and compares the proposed method with 6 other state-of-the-art NR IQA methods and 3 popular FR IQA methods. Our method achieves the highest prediction accuracy in terms of LCC among all compared methods and even compared to the FR IQA approaches. In terms of SRCC, only SOM [6] achieves a slightly stronger correlation.

3.2. Decoder-sided HEVC Quantization Parameter Estimation

We also tested the proposed method on the problem of decoder-sided estimation of the quantization parameter in image and video forensics. For this, we compressed the UCID [16] database, containing 1338 uncompressed color images. For compression we choose the BPG image format [17], that is based on a subset of coding tools of the Main 4:4:4 16 Still Picture Profile [18] of HEVC [19]. To generate our image material for the evaluation, for each reference

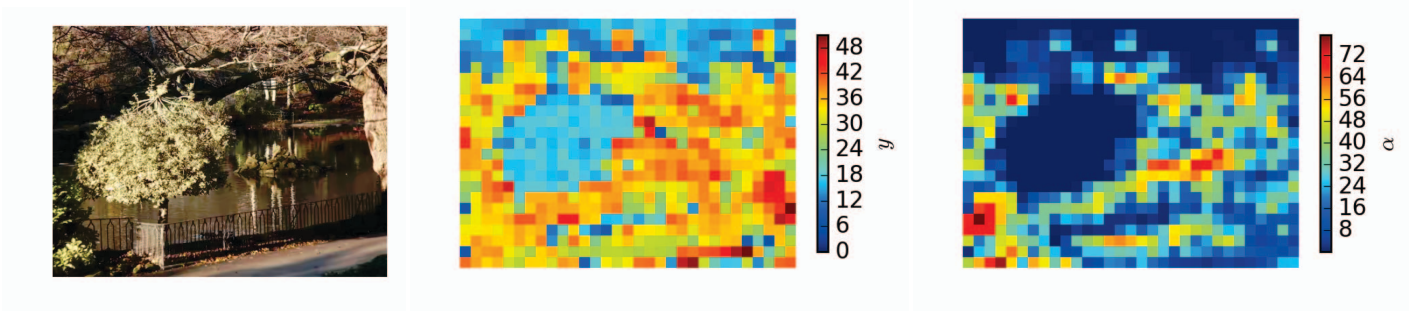

Fig. 3: Example outputs for densely sampled patches with the weighted average patch aggregation. Left: Input image, middle: patchwise QP estimate, right: weight in weighted average calculation. The QP during encoding is 35. The estimated QP is 35.38.

Method	LCC	SRCC
PSNR	0.856	0.866
SSIM[13]	0.906	0.913
FSIM[14]	0.960	0.964
DIIVINE [2]	0.917	0.916
BLIINDS-II [3]	0.930	0.931
BRISQUE [4]	0.942	0.940
CORNIA [5]	0.935	0.942
CNN [7]	0.953	0.956
SOM [6]	0.962	**0.964**
Patchwise (proposed)	**0.972**	0.960
Weighted (proposed)	0.963	0.955
Weighted+ (proposed)	0.970	0.961

Table 1: Comparison of different NR IQA methods based on the LIVE database. For further reference, in the first three rows performances of three prominent FR IQA methods are reported as well in italic. The highest LCC and SRCC for the NR IQA methods are set in bold. The reported correlations are the average correlation achieved on the test sets of 10 random train-test splits.

image we created 52 compressed versions by compression with a quantization parameter QP $\in \{0 \dots 51\}$. The first 100 reference images are put aside as test set and the second 100 reference images are used as validation set during training. Because it is easier to achieve correlations of predictions between distorted versions of the same reference image, we sample only one distorted version per reference image in the validation and test set. The training set is resampled for each epoch. For each epoch, 128 random reference images are randomly picked. For each of these, 10 random distorted versions are sampled and 32 random patches extracted. This creates a total of 40960 patches used in each epoch. Each mini-batch consists of 4 images with 32 patches each. The network is then trained for 1000 epochs, each of which takes around $20s$ to train on a Titan X GPU. After each epoch, the model is evaluated on a fixed subsample of the validation set and only the model with the lowest validation loss is kept.

Fig. 2 shows the evaluation of our three models on the test set. The simple patchwise training performs well, even with low number of patches. With 128 and more patches, the weighted model performs as well in terms of correlations and slightly superior in terms of MSE. It needs more patches to perform well, because it uses the weighting mechanism to select a subset of relevant patches. The weighted model that jointly minimizes both loss functions (*weighted+*) performs best in this experiment if it is pooling over more than 4 patches and achieves the maximal performance using about 16 patches. Fig. 3 gives an impression about the patchwise QP estimation and the corresponding weights. The true QP in this example is QP $= 35$ and was estimated as $\text{QP}_{est} = 35.38$. It can be seen that for patches for which the estimated QP (middle image) is far from the ground truth value, the according weight is reduced (right image).

4. CONCLUSION

We applied a deep CNN to the problem of NR IQA and the somewhat related task of blind QP estimation. For the NR IQA task, the proposed CNN achieved the best LCC among all evaluated NR and FR IQA methods. This result is counter-intuitive as IQA is commonly believed to rely, in contrast to image classification tasks, on rather low-level features. This motivates to continue working on a better understanding of the features driving the perception of image quality [20]. In order to deal with locally variable sensitivity, we proposed a weighted average patch aggregation method. Although we were not able to show an improvement on the NR IQA results, this method clearly increased LCC as well as SRCC for QP estimation. Thus, future studies should further address local distortion sensitivity in order to take it into account for the IQA problem. As CNNs achieve high performance for FR IQA [21], the CNN approach offers an excellent framework to explore the domain of reduced reference IQA [22], living between NR and FR IQA.

5. REFERENCES

[1] R. A. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: A survey," *Information Sciences*, vol. 301, pp. 141–160, 2015.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] P. Bestagini, K. M. Fontani, S. Milani, M. Barni, A. Piva, M. Tagliasacchi, and K. S. Tubaro, "An overview on video forensics," *Eur. Signal Process. Conf.*, , no. 2012, pp. 1229–1233, 2012.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[12] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] 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.

[14] 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.

[15] 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.

[16] G. Schaefer and M. Stich, "Ucid: an uncompressed color image database," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2003, pp. 472–480.

[17] F. Bellard, "http://bellard.org/bpg/," .

[18] F. Bossen, "Common test conditions and software reference configurations Output," document JCTVC-H1100 of JCT-VC, 2013.

[19] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. e0130140, 2015.

[21] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Full-reference image quality assessment using neural networks," in *Int. Work. Qual. Multimed. Exp.*, 2016.

[22] S. Bosse, Q. Chen, M. Siekmann, W. Samek, and T. Wiegand, "Shearlet-based reduced reference image quality assessment," in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016.