NEURAL NETWORK-BASED ESTIMATION OF DISTORTION SENSITIVITY FOR IMAGE QUALITY PREDICTION

Sebastian Bosse[1],*Sören Becker*[1], *Zacharias V. Fisches*[1], *Wojciech Samek*[1], *and Thomas Wiegand*[1,2]

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

ABSTRACT

Due to its computational simplicity, the PSNR is a popular and widely used image quality measure, although it correlates poorly with perceived visual quality. Distortion sensitivity, a reference image specific property, can be used to compensate for the lack of perceptual relevance of the PSNR. Based on the functional mapping between perceptual and computational quality a deep convolutional neural network is used to estimate patchwise distortion sensitivity. The local estimates are used for an imagewise perceptual adaptation of the PSNR. The performance of the proposed estimation approach is evaluated on the LIVE and TID2013 databases and shows comparable or superior performance as compared to benchmark image quality measures.

Index Terms— Neural network, distortion sensitivity, image quality assessment, perceptual models

1. INTRODUCTION

Reliable estimation of perceptual quality is crucial for the evaluation, design and optimization of image and video communication systems. Image quality measures (IQMs) are categorized according to the amount of information about the reference signal available to the quality estimator as full reference (FR), reduced reference (RR) and no reference (NR) approaches. NR quality estimation targets very general perceptual quality models, but may not be suitable for all applications. An important example is video coding [1], where some distortions, e.g. film grain might be introduced intentionally by the director. An encoder, optimizing reference-free, should not 'correct' for those only seeming distortions.

FR IQMs classically follow two main strategies [2]: While *bottom-up* approaches explicitly model the human visual system (HVS) [3, 4, 5], *top-down* approaches mimic potentially hypothesized abstract properties of the HVS from a signal processing perspective [6, 7, 8, 9]. With recent advances in machine learning, a third branch of data-driven approaches emerged [10] that may not rely on potentially deficient domain knowledge.

Generally, FR IQMs benefit from adaptation to the specific content of the images to be tested [11]. This may be done for a whole image or locally, e.g. by considering HVS models such as saliency [12] or scalewise divisive normalization [13], information content [14], conditional probability [15] or learned features [16]. Adaptations are typically based on the reference signal. In [17] visual sensitivity is estimated for image quality prediction from normalized and filtered distorted images using a convolutional neural network (CNN). However, as we will discuss, this forfeits the practically crucial advantage of processing complex calculations only on the reference signal.

The simplest and probably most widely used FR IQMs are the mean squared error (MSE), computed as the average energy of the samplewise error between reference and distorted signal, and the peak signal-to-noise ratio (PSNR), a logarithmic approximation of human perception based on the Weber-Fechner law [18] applied to the MSE. The low complexity of the MSE/PSNR is particularly appealing in time-critical applications such as block-based hybrid video coding, where during mode decision the blockwise distortion is evaluated for every coding mode considered [1]. A perceptual adaptation based solely on the reference image would increase computational complexity for quality prediction only out of the search loop of mode decision.

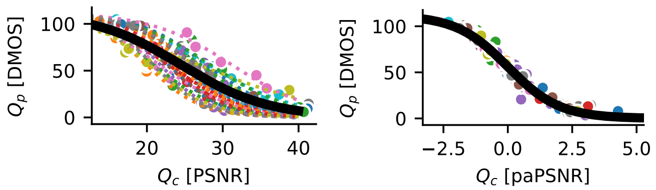
In this paper we build on the previously proposed framework for distortion sensitivity [19] for image quality prediction and extend it by incorporating all functional parameters into an end-to-end learning scheme based on a deep neural network. Parameters of distortion sensitivity are estimated from the reference signal in order to compensate for the perceptual shortcomings of the PSNR and improve its performance for perceptual quality estimation.

The proposed approach is evaluated on the LIVE [20] and the TID2013 [21] databases.

2. METHODS

2.1. Distortion Sensitivity

Perceptual quality assessed in psychophysical tests is characterized by saturation effects in the extreme cases of very high and very poor quality. For the evaluation of computational quality models, this is accounted for by mapping the



(a) No Adaptation

(b) Compensating for c_{r}, d_{r}

Fig. 1: PSNR (*left*: no compensation, *right:* compensated for slope β_2 and shift β_3 per reference image) vs. DMOS for the JPEG subset of the LIVE database [20]. Colored dashed curves and circles indicate regressed and measured DMOS values for individual reference images. Thick black curves show regressed DMOS values over all reference images.

computational quality value Q_c to perceptual quality values Q_p. A commonly used mapping function for mapping from the computational to the perceptual domain is the 4-parameter generalized logistic function [22]

$$\begin{aligned} Q_p =& f(Q_c; \boldsymbol{\beta}) \\ =& \beta_0 + \frac{\beta_1 - \beta_0}{1 + e^{-\beta_2 \cdot (Q_c - \beta_3)}}. \end{aligned} \tag{1}$$

Parameters $\boldsymbol{\beta}$ can be estimated as $\widehat{\boldsymbol{\beta}}$ from a set of images annotated with perceptual quality values Q_p, typically MOS or DMOS, and computational quality scores Q_c, and can be used to predict perceptual quality scores from computational quality values (computed by a specific quality measure) as

$$\widehat{Q}_p = f(Q_c; \widehat{\boldsymbol{\beta}}). \tag{2}$$

A typical regression from Q_c to Q_p exemplified for the JPEG-subset of the LIVE database [20] with Q_c calculated as PSNR is shown in Fig. 1a, where the colored circles indicate DMOS and PSNR values for individual images and the black line represents the regression function from Q_c to $\widehat{Q}_p$ estimated on the full set of images in the database.

Although the parameters are typically estimated, β_0 and β_1 relate directly to the lower and upper bounds of the perceptual quality values of the quality annotations. As such, β_0 and β_1 are mainly driven by the range of the perceived quality scale and, thus, defined by the experimental design of the subjective test and know in principal a-priori. The regression function is shifted with respect to Q_c by β_3. The slope of the regression, that, with a value of $\widehat{Q}_p(Q_c = \beta_3) = \frac{\beta_1 - \beta_0}{4} \cdot \beta_2$, is steepest at $Q_c = \beta_3$, is controlled by β_2, scaled by the range of β_0 to β_1. Disregarding the scaling in the slope of the regression function, β_2 and β_3 are independent of the quality scale, but on the relation between the values of a specific quality measure and the ground-truth quality scores of the image set used to estimate the regression parameters. If the data set used for estimating $\boldsymbol{\beta}$ is large and diverse enough to ensure generalization, the estimated regression parameters can be used to map values of a specific computational quality measure into the perceptual quality domain for other images as well. Fig. 1a suggests that β_3 is predominantly determined by the source image and its estimation specific for the source image r could be used to compensate the systematic deviations of a computational quality measure Q_c. With c compensating globally for β_2 and d_{r} compensating for the source image specific shift we find a perceptually adapted Q_c as

$$Q_{c,adapt} = c \cdot (Q_c - d_{r}). \tag{3}$$

The resulting regression after compensation with c, d estimated as the (in practice unavailable) regression parameters β_2, β_3 is shown in Fig. 1b for the JPEG-subset of the LIVE database. We will refer to d_{r} as *distortion sensitivity* of an image r with regard to a specific computational distortion measure.

2.2. Distortion Sensitivity and the Mean Squared Error

Disregarding c for the moment, with Eq. 3 the PSNR can be shifted to an imagewise perceptually adapted PSNR

$$\begin{aligned} \text{paPSNR} &= \text{PSNR} - d_{r} \\ &= 10 \cdot \log_{10} \frac{C^2}{10^{\frac{d_{r}}{10}} \text{MSE}}, \end{aligned} \tag{4}$$

which leads to an imagewise perceptually adapted MSE as

$$\text{paMSE} = 10^{\frac{d_{r}}{10}} \text{MSE}. \tag{5}$$

Distortion sensitivity is not necessarily a global signal property, but may vary locally over image regions. With distortion sensitivity $d(p)$ of the local image patch p and the local $\text{MSE}(p)$, the MSE between the samples of the distorted and the reference image in patch p, we redefine a locally perceptually adapted MSE as

$$\text{paMSE} = \frac{1}{P} \sum_{p=0}^{P-1} 10^{\frac{d(p)}{10}} \cdot \text{MSE}(p) \tag{6}$$

with P being the number of patches extracted from an image.

2.3. Neural Network-Based Estimation of Distortion Sensitivity

Motivated by its previously shown performance for perceptual quality estimation [16], we use a VGGnet-inspired [23] CNN. Input to the network are 32 × 32 pixel-sized patches of the reference image. Our proposed CNN comprises 12 weight layers that are used to estimate the distortion sensitivity $d(p)$ for the given image patch p. The network is organized as a series of 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 layers, followed by FC-512, FC-1 layers. Convolutional layers are activated through a Leaky Rectified Linear Unit (LReLU) activation function with a leakyness of 0.2 [24]. An additional parallel branch consisting of only 1 weight and having a constant input of 1 is used for estimating a global value of c.

The outputs $\widehat{c}, \widehat{d}(p)$ are used with Eq. 1, Eq. 3 and Eq. 6 for estimating the perceptual quality as

$$\widehat{Q}_p = a + \frac{b-a}{1+e^{-\widehat{c}\cdot Q_c}}, \qquad \text{with} \tag{7}$$

$$Q_c = 10 \cdot \log_{10} \frac{C^2}{\frac{1}{P}\sum_{p=0}^{P-1} 10^{\frac{\widehat{d}(p)}{10}} \cdot \mathrm{MSE}(p)}. \tag{8}$$

As discussed previously, a, b are chosen as the lower and upper limit of the rating scales used during psychophysical quality assessment.

The network is trained by minimizing the mean absolute error (MAE) between reported and predicted perceptual quality

$$E = |Q_p - \widehat{Q}_p| \tag{9}$$

Since the network is estimating d patchwise, but Q_p is estimated imagewise, patches for one image are in the same mini-batch during training. For training, each mini-batch contains 1 image, represented by 32 randomly sampled image patches. Optimization of the weights is controlled using the ADAM method [25]. In order to prevent overfitting, the final model used for evaluation is chosen as the one with the best validation loss during training [26]. For evaluation, patches are densely sampled from the test images.

3. EXPERIMENTS AND RESULTS

The proposed approach is evaluated on the TID2013 [21] and LIVE [20] image quality databases. TID2013 and LIVE differ in the range and orientation of quality scores. In order to make errors and gradients comparable across databases, scores have been linearly mapped to the same range. For cross-validation, databases have been randomly split by reference image into training, evaluation and test set (17/6/6 for LIVE, 13/6/6 for TID2013). Results are reported based on 100 random splits. All models are trained for 150 epochs.

Fig. 2 shows the optimization loss during training, validation and testing over the number of epochs of training for one randomly chosen split from training over TID2013. The loss shows the typical behavior for iterative gradient descent minimization and does not indicate overfitting.

Table 1 summarizes the performance of the proposed approach applied per distortion in terms of Spearman rank order correlation coefficient (SROCC) for the full LIVE database and the *actual* subset of TID2013. The proposed approach consistently outperforms PSNR and shows comparable or superior performance to SSIM, MS-SSIM, FSIM.

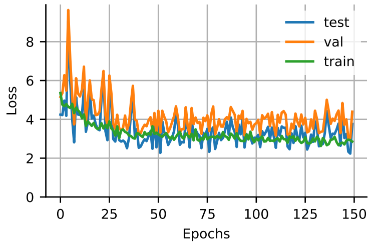


Fig. 2: Course of loss over epochs for test, validation and training.

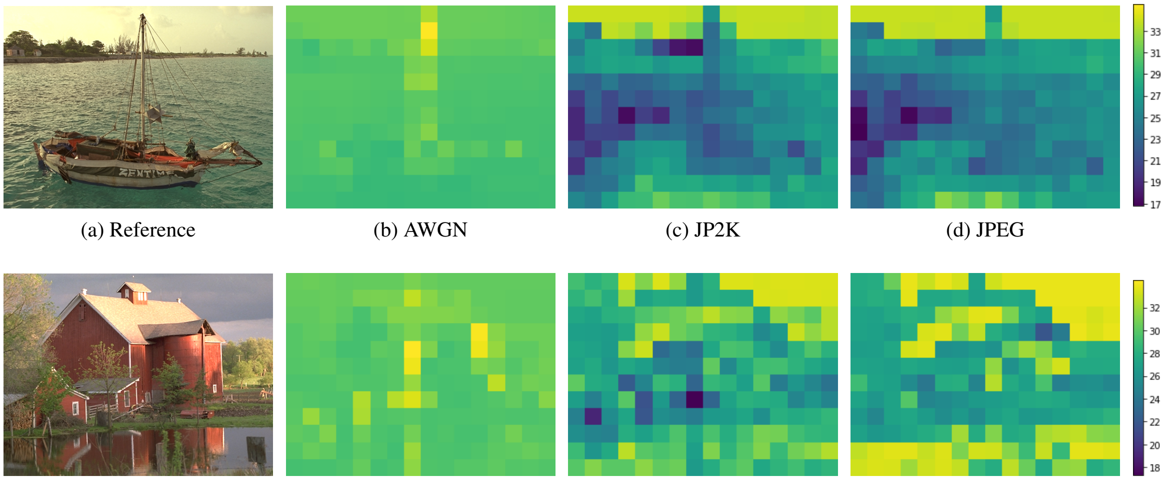
		PSNR	SSIM	MS-SSIM	FSIM	paPSNR
LIVE	jpg2k	0.90	0.96	0.96	0.97	0.95
	jpg	0.88	0.98	0.97	0.98	0.96
	gwn	0.99	0.97	0.97	0.97	0.98
	blur	0.78	0.95	0.95	0.97	0.93
	ff	0.89	0.96	0.95	0.95	0.93
TID2013	gwn	0.93	0.87	0.86	0.91	0.95
	scn	0.92	0.85	0.85	0.94	0.96
	mn	0.83	0.78	0.81	0.81	0.86
	hfn	0.91	0.86	0.86	0.9	0.95
	in	0.90	0.75	0.76	0.82	0.90
	gblur	0.91	0.97	0.97	0.96	0.96
	den	0.95	0.93	0.93	0.93	0.94
	jpg	0.92	0.92	0.93	0.93	0.95
	jpg2k	0.88	0.95	0.95	0.96	0.96
	mgn	0.89	0.78	0.78	0.86	0.92
	lcmi	0.91	0.91	0.91	0.95	0.96

Table 1: Average SROCC over 100 runs of the proposed method for the distortion types of LIVE database and the *actual* subset of TID2013 in comparison to PSNR, SSIM [7], MS-SSIM [6] and FSIM [9].

	PSNR	SSIM	MS-SSIM	FSIM	paPSNR	paPSNR
					$\beta = 1$	$\beta = \beta_{opt}^{dist}$
LIVE	0.88	0.95	0.95	0.96	0.91	0.94
TID2013 (*actual*)	0.82	0.88	0.88	–	0.84	0.92
TID2013 (*full*)	0.64	0.74	0.79	0.8	0.65	0.88

Table 2: Average SROCC over 100 runs for the full LIVE database, the *actual* subset of TID2013 and the full TID2013 database in comparison to PSNR, SSIM, MS-SSIM and FSIM without ($\beta = 1$) and with ($\beta = \beta_{opt}^{dist}$) distortion type specific linear scaling of $\widehat{d}$.

Results obtained for models trained on the full databases and the *actual*-subset from TID2013 are shown in Table 2 in the column '$\beta = 1$'. The quantification of distortion sensitivity as a property of the reference image is dependent on the distortion type and this dependency can be effectively approximated by a distortion-specific linear scaling factor [19]. The column titled '$\beta = \beta_{opt}^{dist}$' lists the SROCC obtained when such a scaling factor is applied to the distortion type-agnostic estimate of d. This simple adaptation significantly increases the prediction monotonicity of the proposed method and renders it comparable (LIVE) or superior (TID2013) to SSIM,



(a) Reference (b) AWGN (c) JP2K (d) JPEG

(e) Reference (f) AWGN (g) JP2K (h) JPEG

Fig. 3: Examples of resulting sensitivity maps $d(p)$ for different distortions resulting from additive white Gaussian noise (AWGN), JPEG2000 and JPEG compression.

Trained on	LIVE(*full*)	TID2013 (*actual*)
Tested on	TID2013	LIVE
jpg2k	0.93	0.91
jpg	0.93	0.91
gwn	0.91	0.98
gblur	0.85	0.83
common subset	0.92	0.89

Table 3: Average SROCC over 100 runs of paPSNR trained on full LIVE database or the *actual* subset of TID2013 and tested for selected distortions types on the other database. Performance for the subset of common distortion types are show in *common subset*.

MS-SSIM and FSIM.

This distortion type dependency can also be seen from the resulting sensitivity maps; examples for AWGN, JP2K and JPEG are shown in Fig. 3. For JP2K and JPEG, lower sensitivities are assigned to image regions of higher activity, probably capturing masking effects. For JP2K and JPEG, sensitivity maps are similar, but differ in detail. For AWGN however, rather constant sensitivity maps are estimated. The reason is that the PSNR as a base quality measure is already very accurate in predicting perceptual quality for AWGN and the consideration of distortion sensitivity cannot achieve any further improvement.

Table 3 shows that the proposed approach is relatively stable in a cross-database evaluation. LIVE contains more reference images than TID2013 (29 vs. 25), and therefore achieves a higher performance than vice-versa. However, per reference image, TID2013 is almost completely contained by LIVE, hence it is difficult to draw a final conclusion regarding the generalization ability.

4. CONCLUSION

The functional parameters of the sigmoidal mapping from the computational domain into the perceptual domain are typically estimated over a full database. Two of the four parameters (β_0, β_1) are approximately predefined by the setup of the psychophysical quality assessment experiment and the other two can be estimated jointly as global (β_2) and reference-specific (β_3) in an end-to-end framework using a CNN for improving the PSNR. Although other IQMs could be used as a basis, the PSNR approach allows to conveniently distribute the computational effort, as complex computations are only performed on the reference signal, e.g. for perceptually relevant mode decision in video coding [1]. The network architecture used is rather deep; for real-time applications it would be beneficial to reduce network depth. We evaluated our method on 32×32 pixel-sized patches, however performance might increase with other patch sizes. Training on more patches might also influence the performance, but would probably demand larger databases. In the proposed approach distortion sensitivity was estimated from the reference image only. While this is argued as well in [16], [10] interprets sensitivity as a feature of the distorted signal. For better understanding of the underlying psychovisual processes this should be studied comparatively. A very interesting property of our approach is that no explicit domain knowledge is used. Therefore it can be directly applied to related domains, such as audio or video quality assessment.

5. REFERENCES

[1] Thomas Wiegand and Heiko Schwarz, “Video coding: Part II of fundamentals of source and video coding,”

Foundations and Trends in Signal Processing, vol. 10, no. 1-3, pp. 1–346, 2016.

[2] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[3] B. Girod, "What's Wrong with Mean-squared Error?," in *Digit. Images Hum. Vis.*, pp. 207–220. 1993.

[4] S. J. Daly, "Application of a noise-adaptive contrast sensitivity function to image data compression," *Opt. Eng.*, vol. 29, no. 8, pp. 977–987, 1990.

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

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *IEEE Asilomar Conf. Signals, Syst. Comput.*, vol. 2, no. 1, pp. 1398–1402, 2003.

[7] 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.

[8] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *Signal Process. Image Commun.*, vol. 61, no. Supplement C, pp. 33–43, 2018.

[9] L. Zhang, L. Zhang, X. Mou, D. Zhang, and X. Mou, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[10] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep Convolutional Neural Models for Picture Quality Prediction," *IEEE Signal Process. Mag.*, vol. 34, no. November, pp. 130–141, 2017.

[11] B. Ortiz-Jaramillo, J. Niño-Castañeda, L. Platiša, and W. Philips, "Content-aware video quality assessment: predicting human perception of quality using peak signal to noise ratio and spatial/temporal activity," in *SPIE/IS&T Electron. Imaging*. International Society for Optics and Photonics, 2015, vol. 9399, p. 939917.

[12] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, 2016.

[13] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electron. Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[14] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.

[15] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C.J. J. Kuo, "Compressed Image Quality Metric Based on Perceptually Weighted Distortion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5594–5608, 2015.

[16] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, 2018.

[17] J. Kim and S. Lee, "Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework," *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1676–1684, 2017.

[18] S. E. Palmer, *Vision science: Photons to phenomenology*, vol. 1, MIT press Cambridge, MA, 1999.

[19] S. Bosse, M. Siekmann, W. Samek, and T. Wiegand, "A Perceptually Relevant Shearlet-Based Adaptation of the PSNR," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 315–319.

[20] 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.

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

[22] VQEG, ITUT Tutorial, and VQEG, "Objective perceptual assessment of video quality: full reference television," *ITU-T Telecommun. Stand. Bur.*, 2004.

[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ImageNet Chall.*, pp. 1–10, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[25] D Kingma and J Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[26] L. Prechelt, "Early stopping–but when?," in *Neural Networks: Tricks of the Trade*, pp. 53–67. Springer, 2012.