Estimation of Distortion Sensitivity for Visual Quality Prediction Using a Convolutional Neural Network

Sebastian Bosse[a,*], Sören Becker[a], Klaus-Robert Müller[b,c,d,*], Wojciech Samek[a,*], Thomas Wiegand[a,b,*]

[a] *Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, Berlin 10587, Germany*
[b] *Department of Electrical Engineering & Computer Science, Technische Universität Berlin, Marchstr. 23, Berlin 10587, Germany*
[c] *Department of Brain & Cognitive Engineering, Korea University, Anam-dong 5ga, Seongbuk-gu, Seoul 136-713, South Korea*
[d] *Max Planck Institute for Informatics, Stuhlsatzenhausweg, Saarbrücken 66123, Germany*

Abstract

The PSNR and MSE are the computationally simplest and thus most widely used measures for image quality, although they correlate only poorly with perceived visual quality. More accurate quality models that rely on processing on both the reference and distorted image are potentially difficult to integrate in time-critical communication systems where computational complexity is disadvantageous. This paper derives the concept of distortion sensitivity as a property of the reference image that compensates for a given computational quality model a potential lack of perceptual relevance. This compensation method is applied to the PSNR and leads to a local weighting scheme for the MSE. Local weights are estimated by a deep convolutional neural network and used to improve the PSNR in a computationally graceful distribution of computationally complex processing to the reference image only. The performance of the proposed estimation approach is evaluated on LIVE, TID2013 and CSIQ databases and shows comparable or superior performance compared to benchmark image quality measures.

Keywords: Deep learning, distortion sensitivity, image quality assessment, perceptual coding, visual perception

*Corresponding author
Email addresses: sebastian.bosse@hhi.fraunhofer.de (Sebastian Bosse), klaus-robert.mueller@tu-berlin.de (Klaus-Robert Müller), wojciech.samek@hhi.fraunhofer.de (Wojciech Samek), thomas.wiegand@hhi.fraunhofer.de (Thomas Wiegand)

1. Introduction

Digital images and videos are ubiquitous in modern society and their availability relies on efficient transmission systems. For transmission over today's channels, signals are digitized and compressed, leading to distortions in the signal at the receiver. Hence, a crucial aspect for designing, benchmarking and optimizing communication systems is the quality of the received signal. The ultimate receiver in most multimedia communication systems is a human, thus, the decisive criterion for quality is the human judgement. Unfortunately, no reliable model for quality judgement is at hand. Therefore, perceived quality is typically assessed in psychophysical judgment tests, during which observers are presented with a stimulus and asked for a response on the respective quality. Individual observer's ratings are pooled to the famous mean opinion score (MOS), or, when referenced to a rating of the reference stimulus to the differential mean opinion score (DMOS) [1]. Recommendations of the International Telecommunication Union specify the different procedures for such assessment [1, 2].

However, quality assessment by humans is cumbersome, expensive and in many application scenarios not accessible, e.g. due to real-time constraints. Computational approaches for image quality estimation aim at bypassing these problems by estimating the quality of signals without the direct involvement of humans. Computational quality models are typically categorized based on the amount of information about the reference signal available to the model as full reference (FR), reduced reference (RR) and no reference (NR) approaches. Unarguably, NR quality estimation poses the most ambitious challenge as it has to the least information available. Yet conceptually, NR quality estimation may not be a feasible approach for certain applications with an important example being encoder control in video compression [3]. An unreferenced rate-distortion optimization would steer the encoder towards coding decisions that remove any type of noise or artifact. In some videos, however, noise and artifacts as for instance film grain, motion blur, or camera shakes are artistic components that are intentionally introduced in order to evoke a certain emotional response in the viewer. Prominent examples for this are the movies *The Blair Witch Project* or *Cloverfield*. However, even with a reference available, perceptual aspects of quality are still not efficiently used for optimizing compression schemes.

The simplest FR image quality measure (IQM) is presumably the mean square error (MSE) between reference image and distorted image. Since it has convenient features, it is perhaps also the most widely used IQM, as it *a)* is of low computational complexity, *b)* is memoryless, *c)* qualifies mathematically as a distance metric in R^N, *d)* has a clear physical interpretation as the energy of the error signal, *e)* features convexity, symmetry and differentiability, allowing for simple optimization procedures, and *f)* is additive [4]. Despite all these advantageous properties the MSE has one crucial disadvantage: As a quality estimator it does not correlate well with visual quality as perceived by humans [5]. This lack of correlation with human perception led scientists and quality researchers to build IQMs around models specifically incorporating engineering as well as biological domain knowledge. Two main strategies are classically distinguished for FR IQMs [6]: whereas *bottom-up* approaches explicitly emulate the human visual system (HVS) [5, 7, 8], *top-down* approaches model hypothesized abstract processing properties of the HVS from a signal processing perspective [9, 10, 11, 12]. Motivated by success of machine learning in image processing areas, purely data-driven approaches [13] represent a recently emerging third strategy with the potential advantage of circumventing deficient domain knowledge of human visual processing.

In general, FR IQMs can benefit from adaptations to the specific content of the images whose perceptual quality is to be estimated [14] and this adaptation is mostly implemented by a weighting scheme. Proposed weighting schemes consider for instance models of the HVS such as saliency [15], scale-wise divisive normalization [16], information content [17], conditional probability [18], contrast sensitivity [19, 20], contrast and luminance perception [21, 22] or shearlet-based measurements of local activity [23]. These weighting schemes model different aspects of the HVS but relate to the same concept of *distortion sensitivity*, suggesting that distortions measured by a given quality model are more (or less) visible in one image area than in another and hence that this image area is more (or less) sensitive to distortions than another. However, here the estimation of distortion sensitivity relies on explicit domain knowledge. Interestingly and in contrast to most previous approaches, we will see that our psychophysical derivation will lead to a non-normalized weighting scheme.

Although accurate quality models have been proposed, most of them are infeasible in time-critical applications such as video compression [3]. In modern video codecs such as High Efficiency Video Coding (HEVC) [24] frames of a video are subdivided into blocks. For the encoding of each of these blocks, a multitude of coding modes are available, each of which has to be tested for its resulting rate-distortion costs, i.e. for each block and coding mode the induced distortion has to be calculated. Evidently, from an efficiency perspective a complex perceptual distortion measure is unsuitable here.

Thus, in this paper, distortion sensitivity is modelled as a property of the reference image. This is particularly appealing as in combination with a low-complex quality model, i.e. the mean squared error (MSE) or peak signal-to-noise ratio (PSNR), computationally demanding processing could be restricted to the reference image only. For time-critical applications such as block-based hybrid video coding, this is a crucial property, as complex processing would be gracefully taken out of the search loop [3].

The first contribution of this paper is the derivation of a functional definition of distortion sensitivity. This is based on a conceptual and statistical discussion of the parameters of the regression function that is used to map the output of a computational quality model into the perceptual domain. In a second contribution, the limits of the proposed framework are explored for a full image-wise compensation of the PSNR for distortion sensitivity. The adaptation of the proposed framework to other quality models is straight-forward. In a third contribution, the concept of distortion sensitivity is adapted from a global to a local scale and it is shown for the PSNR how this leads to a weighting scheme that can be applied to the MSE. A neural network-based approach for the estimation of local distortion sensitivity from the reference image in an end-to-end trained image quality prediction framework is evaluated on LIVE, TID and CSIQ and compared to existing approaches in the literature as fourth contribution.

The paper is structured as follows: In Section 2 the concept of distortion sensitivity as a property of the reference image is derived and discussed. The neural network-based estimation of local distortion sensitivity is presented in Section 3. Performance of the presented approach for neural network-based compensation for distortion sensitivity is evaluated and compared to other relevant approaches on the LIVE [25], the TID2013 [26] and the CSIQ [27] databases in Section 4. Section 5 concludes the paper with a discussion and sketches application opportunities and future work.

2. Distortion Sensitivity

2.1. Psychometric Relation between Computational and Perceptual Quality

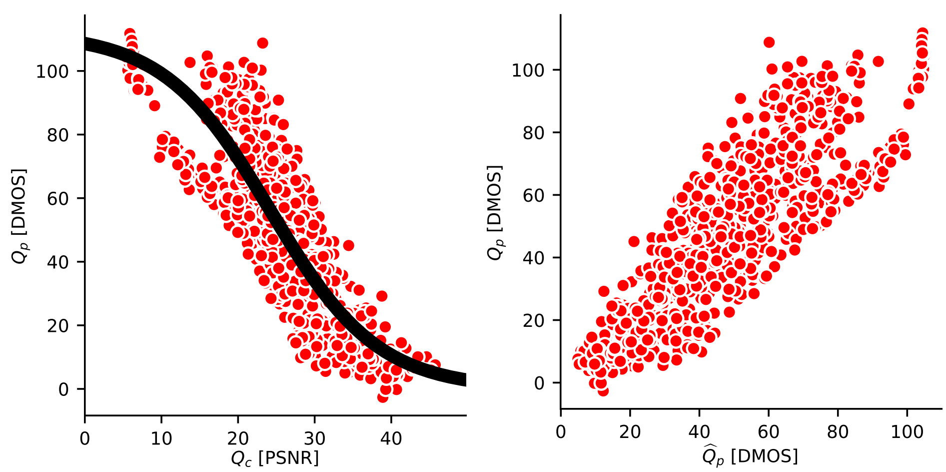


Figure 1: Relation between Q_c, Q_p and $\widehat{Q}$. **Left**: Mapping of Q_c to Q_p for the LIVE database. Red circles indicate Q_c vs. Q_p for individual images, the black curve shows the resulting regression function for the full set. **Right**: Resulting quality predictions $\widehat{Q}_p$ vs. true quality values Q_p.

Due to saturation effects in the extreme cases of imperceptible quality loss or strong impairments, subjective image quality ratings typically do not relate linearly to many computational quality measures. The relation is commonly linearized by a nonlinear mapping from the computational to the perceptual domain. A widely used function is the 4-parameter generalized logistic function [28]

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

Parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ are estimated as $\widehat{\boldsymbol{\beta}}$ based on image-wise pairs of computational quality values Q_c, output of a computational quality model, and perceptual quality values Q_p, output of a quality assessment, e.g. a psychophysical test. Resulting estimates of the regression parameters are then used to predict perceptual quality values from computational quality values as

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

Regression parameters[1] $\boldsymbol{\beta}$ are not valid globally, but dependent on the quality assessment procedure used to obtain Q_p and the quality model computing Q_c, where the consistency of the relation between Q_p and Q_c relies on the performance of the computational quality model. In practice, regression parameters can only be estimated on a limited number of images that need to be sufficiently representative in order to ensure generalization of the prediction to unseen images.

Fig. 1 exemplifies a typical regression based on Eq. 1 from computational to perceptual quality (left) and the resulting prediction of perceptual quality from computational quality (right) with Q_c calculated as peak signal-to-noise ratio (PSNR) and $\boldsymbol{\beta}$ estimated on the full LIVE database [25]. Red circles denote pairs of (Q_p, Q_c) or $(Q_p, \widehat{Q}_p)$ respectively, for individual images. The black line represents the estimated regression function from Q_c to $\widehat{Q}_p$.

Although estimation of regression parameters is typically data-driven, β_0 and β_1 relate directly to the lower and upper bounds of the perceptual quality values. As such, β_0 and β_1 are mainly determined by the range of the perceptual quality scale and, thus, defined by the experimental design of the subjective test and therefore in principle known a-priori. Regression parameter β_3 denotes a horizontal shift of the regression function with respect to Q_c. The slope of the regression, which, with a value of $\frac{\partial \widehat{Q}_p}{\partial Q_c}(Q_c = \beta_3) = \frac{\beta_1 - \beta_0}{4} \cdot \beta_2$, is steepest at $Q_c = \beta_3$, is controlled by β_2 and scaled by the range of β_0 to β_1. Disregarding this scaling, β_2 and β_3 are not depending on the quality scale, but on the relation between the values of a specific quality measure and the ground-truth quality scores for the image set used to estimate the regression parameters. Hence, β_0, β_1 in Eq. 1 can be fixed to the lower and upper bound of the rating scale a and b and $\boldsymbol{\beta}$ can be reduced to $\boldsymbol{\beta} = (\beta_2, \beta_3)$.

Note that another often used regression function, e.g. in [20, 29, 30], the

[1] In order to simplify notation, the $\widehat{\ }$-sign is dropped from now on and estimated regression parameters $\widehat{\boldsymbol{\beta}}$ are referred to as as $\boldsymbol{\beta}$.

5-parameter logistic regression

$$f_5(x;\boldsymbol{\alpha}) = \alpha_0 \left(\frac{1}{2} - \frac{1}{1 + e^{\alpha_1 \cdot (x - \alpha_2)}} \right) + \alpha_3 \cdot x + \alpha_4$$

extends the 4-parameter logistic regression by a linear term $\alpha_3 \cdot Q_c$ as readily seen by reparameterizing $f_5(x;\boldsymbol{\alpha})$ with $\alpha_0 = \beta_0 - \beta_1$, $\alpha_1 = -\beta_2$, $\alpha_2 = \beta_3$ and $\alpha_4 = \frac{1}{2}(\beta_0 + \beta_1)$. In contrast to the psychophysical ratings scale, the 5-parameter logistic function is therefore not bounded and furthermore might yield non-monotonic regression functions contradicting psychophysical quality ratings. For these reasons, the 4-parameter logistic regression function is favored in this analysis. In principle, however, the proposed framework can also be used with the 5-parameter logistic function. In this case, parameters α_0 and α_1 could not simply be set to a and b but instead α_0, α_1 and α_3 would need to be estimated on the training set – similar to what will be presented for β_2.

2.2. Distortion Sensitivity as an Image Property

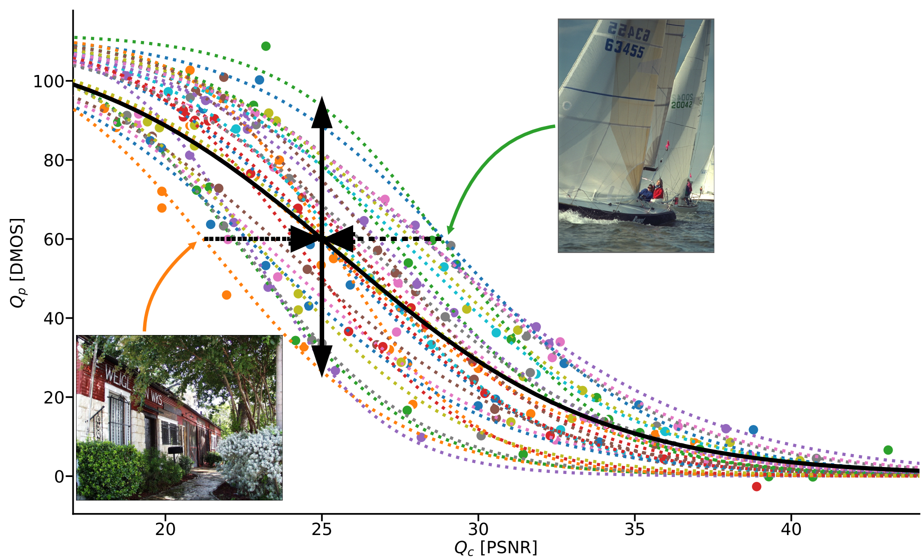


Figure 2: PSNR vs. DMOS for the JPEG-subset of the LIVE database [25]. High values of DMOS denote low subjective quality. Colored dashed curves and circles indicate regressed and measured DMOS values for individual reference images. The thick black curve shows regressed DMOS values for the whole ensemble. Examples images are given for the two extreme cases of distortion sensitivity.

Regression parameters are commonly estimated over a set of images based on an ensemble of reference images that are subject to different distortion types at different distortion magnitudes. However, given enough samples, i.e., impairment levels, regression parameters $\boldsymbol{\beta}^{i,d}$ can also be found per reference image i and distortion type d. Note that in practice this would result in the loss of any generalization ability.

Such a reference image specific estimation of $\boldsymbol{\beta}$ is shown in Fig. 2 for JPEG-distorted images from the LIVE database [25] with Q_c measured as PSNR. The database provides Q_c as DMOS, high values of DMOS denote low subjective quality. Circles denote (Q_c, Q_p) pairs of distorted images and are colored according to the base reference image. Colored dashed curves represent the regression functions estimated for the different reference images, the black curve represents the regression function estimated for the full ensemble. Reference-specific regression curves are widely dispersed around the ensemble-wide regression. This gives raise to the notion of *distortion sensitivity*, as for a given PSNR distorted versions of some reference images exhibit a rather high perceptual quality, while others are reported to appear highly distorted. This is indicated for the extreme cases by vertical black arrows; with regard to the PSNR, the relatively flat image of the sailing boat, represented by green, is perceptually more sensitive towards JPEG distortions than the highly textured image, represented by orange. Based on the previous interpretation of the regression parameters $\boldsymbol{\beta}$ (and as such $\boldsymbol{\beta}^{i,d}$) and the insight that β_0, β_1 are only dependent on the experimental setup for quality assessment, distortion sensitivity can be functionally captured by β_2 and β_3. Hypothetical compensation for the shifting parameter β_3 is sketched for two reference images by dashed black horizontal arrows in Fig. 2.

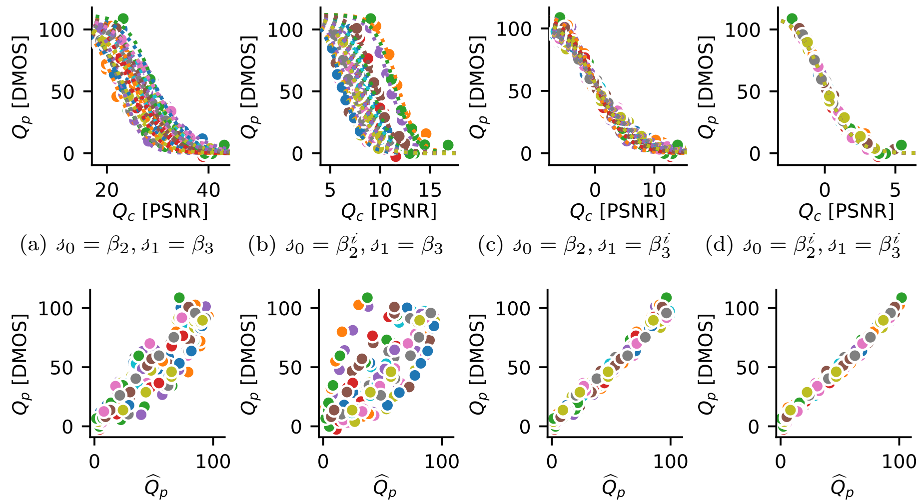
With a functional quantification (for simplicity neglecting different distortion types for the moment) of distortion sensitivity $\mathscr{s}_0^i$ and $\mathscr{s}_1^i$ of a reference image i such a compensation can be used to adapt a computational quality value Q_c as

$$Q_{ac} = \mathscr{s}_0 \cdot (Q_c - \mathscr{s}_1). \tag{3}$$

Assuming a regression according to Eq. 1, β_2^i and β_3^i are optimal predictors of $\mathscr{s}_0$ and $\mathscr{s}_1$. With β_0 and β_1 being the upper and lower bounds a and b of the rating scale, Eq. 2 can be rewritten as

$$\begin{aligned} \widehat{Q}_p =& a + \frac{b-a}{1+e^{-\mathscr{s}_0 \cdot (Q_c - \mathscr{s}_1)}} \\ =& a + \frac{b-a}{1+e^{-Q_{ac}}}. \end{aligned} \tag{4}$$

Although β_2^i and β_3^i are generally not available in practice, assuming their availability helps to analyse the influence and limits of full image-wise distortion sensitivity-based compensation in quality estimation. For this we distinguish four different cases in which we assume available *a*) no reference image-specific information: $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3$; *b*) optimal estimation of $\mathscr{s}_0$ only: $\mathscr{s}_0 = \beta_2^i, \mathscr{s}_1 = \beta_3$; *c*) optimal estimation of $\mathscr{s}_1$ only: $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3^i$; and *d*) optimal estimation of $\mathscr{s}_0$ and $\mathscr{s}_1$: $\mathscr{s}_0 = \beta_2^i, \mathscr{s}_1 = \beta_3^i$, where $\beta_{(\cdot)}$, in contrast to $\beta_{(\cdot)}^i$, denotes a parameter estimation over the full ensemble of reference images. Note that, with regard to correlations between Q_p and Q_{ac}, $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3$ and $\mathscr{s}_0 = 1, \mathscr{s}_1 = 0$ are equivalent, but not with regard to the Pearson correlations between Q_p and



(a) $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3$ (b) $\mathscr{s}_0 = \beta_2^i, \mathscr{s}_1 = \beta_3$ (c) $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3^i$ (d) $\mathscr{s}_0 = \beta_2^i, \mathscr{s}_1 = \beta_3^i$

(e) $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3$ (f) $\mathscr{s}_0 = \beta_2^i, \mathscr{s}_1 = \beta_3$ (g) $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3^i$ (h) $\mathscr{s}_0 = \beta_2^i, \mathscr{s}_1 = \beta_3^i$

Figure 3: Influence of compensating the PSNR for distortion sensitivity on JPEG subset of LIVE database. **Top**: Adapted PSNR vs. ground truth DMOS. **Bottom**: Estimated DMOS compensated for distortion sensitivity vs. ground truth DMOS.

Table 1: Correlations between PSNR compensated for distortion sensitivity and ground truth DMOS (Q_{ac} vs. Q_p), and predicted DMOS compensated for distortion sensitivity and ground truth DMOS ($\widehat{Q}_p$ vs. Q_p). All correlations are calculated on the JPEG subset of the LIVE database.

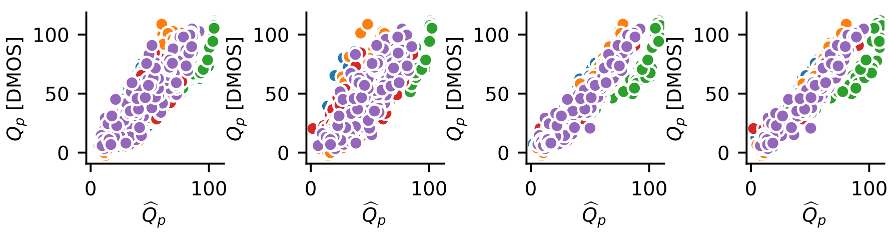
		Q_{ac} vs. Q_p		$\widehat{Q}_p$ vs. Q_p	
		ρ_P	ρ_S	ρ_P	ρ_S
$\mathscr{s}_0 = \beta_2$,	$\mathscr{s}_1 = \beta_3$	-0.88	-0.9	0.9	0.9
$\mathscr{s}_0 = \beta_2^i$,	$\mathscr{s}_1 = \beta_3$	-0.71	-0.72	0.73	0.72
$\mathscr{s}_0 = \beta_2$,	$\mathscr{s}_1 = \beta_3^i$	-0.96	-0.98	0.98	0.98
$\mathscr{s}_0 = \beta_2^i$,	$\mathscr{s}_1 = \beta_3^i$	-0.96	-0.99	0.99	0.99

$\widehat{Q}_p$. Hence, for simplified, yet consistent notation the no adaptation case is represented as $\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3$.

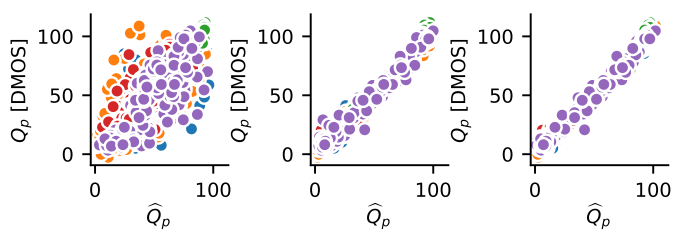
The effect of compensating the PSNR for distortion sensitivity is shown in Fig. 3 for JPEG compressed images from the LIVE database: The top row shows the adapted PSNR (Eq. 3) vs. the ground truth DMOS, the bottom row the predicted DMOS (Eq. 4) vs. the true DMOS for previously defined assumptions, i.e., the left hand side column (Fig. 3a and Fig. 3e) is equivalent to no adaptation. Fig. 3b and Fig. 3f suggest that image-wise compensation for the slope disperses the quality estimates even further, while compensating image-wise for the offset (Fig. 3b and Fig. 3f) and even more a joint compensation for slope and offset (Fig. 3d and Fig. 3h) achieves a clean alignment of quality estimates. Corresponding correlations are summarized in Table 1 and corroborate this observation. It is noteworthy that a joint compensation for slope and offset achieves only small additional improvement over offset-only compensation.

2.3. Distortion Sensitivity and Different Distortion Types

The previous subsection discussed reference image-specific distortion sensitivity subject to a specific distortion type and exemplified this by JPEG distortion. However, different distortion types affect different statistical properties of natural images, hence, also the distortion type may have an influence on distortion sensitivity. This can be accounted for by extending previous considerations and modelling distortion sensitivity not only as a property of a reference image i with respect to a given computational quality measure, but also in dependency of a specific distortion type d.



(a) $s_0 = \beta_2, s_1 = \beta_3$ (b) $s_0 = \beta_2^i, s_1 = \beta_3$ (c) $s_0 = \beta_2, s_1 = \beta_3^i$ (d) $s_0 = \beta_2^i, s_1 = \beta_3^i$



(e) $s_0 = \beta_2^{d,i}, s_1 = \beta_3$ (f) $s_0 = \beta_2, s_1 = \beta_3^{d,i}$ (g) $s_0 = \beta_2^{d,i}, s_1 = \beta_3^{d,i}$

Figure 4: Influence of considering distortion sensitivity on the adapted PSNR for different distortion types. **Top**, from left to right: Distortion type-agnostic consideration of s_0 only, s_1 only, and s_0, s_1 jointly. **Bottom**, from left to right: Distortion type-specific consideration of s_0 only, s_1 only, and s_0, s_1 jointly.

Fig. 4 plots the relation between the estimated quality $\widehat{Q}_p$ and the ground truth quality Q_p for different distortion sensitivity compensation schemes, where again $\beta_{(\cdot)}$ (without superscript) denotes a parameter estimated over the full dataset, $\beta_{(\cdot)}^i$ denotes a reference image-wise estimation over all distortion types in the database, and $\beta_{(\cdot)}^{d,i}$ denotes parameter estimation per reference image i and distortion type d. Clearly, a joint compensation of distortion type d and reference image i can improve the prediction accuracy. Corresponding correlations are summarized in Table 2. Interestingly, as observed previously in

Table 2: Correlation between adapted PSNR and true DMOS (Q_{ac} vs. Q_p) and predicted DMOS and true DMOS ($\widehat{Q}_p$ vs. Q_p) for different adaptions of Q_c by considering neither $\mathscr{s}_0$ nor $\mathscr{s}_0$, only $\mathscr{s}_0$, only $\mathscr{s}_1$, both $\mathscr{s}_0, \mathscr{s}_1$ when accounting for specific distortion types $\mathscr{d}$ or over the set of all distortion types $\mathscr{D}$.

			Q_{ac} vs. Q_p		$\widehat{Q}_p$ vs. Q_p	
			ρ_P	ρ_S	ρ_P	ρ_S
$\mathscr{d}$ agnostic	$\mathscr{s}_0 = \beta_2$,	$\mathscr{s}_1 = \beta_3$	-0.84	-0.87	0.86	0.87
	$\mathscr{s}_0 = \beta_2^{i}$,	$\mathscr{s}_1 = \beta_3$	-0.80	-0.83	0.81	0.83
	$\mathscr{s}_0 = \beta_2$,	$\mathscr{s}_1 = \beta_3^{i}$	-0.88	-0.93	0.90	0.93
	$\mathscr{s}_0 = \beta_2^{i}$,	$\mathscr{s}_1 = \beta_3^{i}$	-0.88	-0.94	0.91	0.94
$\mathscr{d}$ specific	$\mathscr{s}_0 = \beta_2^{i,\mathscr{d}}$,	$\mathscr{s}_1 = \beta_3$	-0.52	-0.50	0.77	0.77
	$\mathscr{s}_0 = \beta_2$,	$\mathscr{s}_1 = \beta_3^{i,\mathscr{d}}$	-0.93	-0.96	0.98	0.99
	$\mathscr{s}_0 = \beta_2^{i,\mathscr{d}}$,	$\mathscr{s}_1 = \beta_3^{i,\mathscr{d}}$	-0.96	-0.99	0.99	0.99

Table 1 for the single distortion case, a compensation solely based on the slope parameter decreases prediction accuracy also in the multi-distortion case, be it estimated per reference image over all distortions type ($\mathscr{s}_0 = \beta_2^{i}, \mathscr{s}_1 = \beta_3$) or per reference image and distortion type ($\mathscr{s}_0 = \beta_2^{\mathscr{d},i}, \mathscr{s}_1 = \beta_3$). Compensation for the offset over all distortions per reference image ($\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3^{i}$) improves prediction accuracy, also considering the distortion type ($\mathscr{s}_0 = \beta_2, \mathscr{s}_1 = \beta_3^{\mathscr{d},i}$) further improves the quality estimation. However, a joint consideration of slope and offset ($\mathscr{s}_0 = \beta_2^{i}, \mathscr{s}_1 = \beta_3^{i}$ and $\mathscr{s}_0 = \beta_2^{\mathscr{d},i}, \mathscr{s}_1 = \beta_3^{\mathscr{d},i}$) achieves only little additional improvement.

The discussion and findings presented in this section suggest distortion sensitivity can be efficiently modelled as a feature of a reference image and functionally captured based on the shifting parameter of the 4-parameter generalized logistic function. Additional image-wise compensation for the slope parameter achieves only little further improvements in prediction accuracy. Hence, in the following only the shifting parameter will be considered as a functional representation of distortion sensitivity. For simplified notation β_2 is replaced by c. This modifies Eq. 3 and Eq. 4 to

$$Q_{ac} = Q_c - \mathscr{s}. \tag{5}$$

and

$$\begin{aligned}\widehat{Q}_p =& a + \frac{b-a}{1+e^{-c\cdot(Q_c-\mathscr{s})}} \\ =& a + \frac{b-a}{1+e^{-c\cdot Q_{ac}}},\end{aligned} \tag{6}$$

where c is estimated over full datasets and distortion sensitivity is denoted as $\mathscr{s}$.

2.4. Localized Distortion Sensitivity

Previous considerations studied distortion sensitivity as a full image feature. However, statistics of natural images are locally structured and spatially highly

non-stationary [31, 32] so that distortion sensitivity not only varies globally across different images, but also spatially within a given image.
Although in principle applicable to any computation distortion measure, the PSNR allows for a very simple consideration of local distortion sensitivity. According to Eq. 5 the PSNR (instantiating the computational quality value Q_c)is compensated for distortion sensitivity and the perceptually imagewise adapted PSNR (paPSNR_I) written as

$$\begin{aligned} \mathrm{paPSNR}_I &= \mathrm{PSNR} - \mathscr{s}_I \\ &= 10 \cdot \log_{10} \frac{C^2}{10^{\frac{\mathscr{s}_I}{10}} \cdot \mathrm{MSE}}, \end{aligned} \tag{7}$$

with $\mathscr{s}_I$ denoting the image-wise distortion sensitivity and C the maximum (peak) sample value of the given signal class, e.g. for 8-bit SDR images $C = 255$. While PSNR and paPSNR_I do not allow for a direct local weighting, the mean squared error (MSE) can be adopted image-wise to the perceptually adapted MSE (paMSE_I)

$$\mathrm{paMSE}_I = 10^{\frac{\mathscr{s}_I}{10}} \cdot \mathrm{MSE}. \tag{8}$$

By localizing distortion sensitivity to a pixel position (x, y) as $\mathscr{s}(x, y)$, we define the perceptually adapted MSE (paMSE) with $s(x, y)$ being the reference and $\tilde{s}(x, y)$ the distorted image samples is defined as

$$\mathrm{paMSE} = \frac{1}{M \cdot N} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} 10^{\frac{\mathscr{s}(x,y)}{10}} \left(s(x, y) - \tilde{s}(x, y) \right)^2 \tag{9}$$

leading directly to a perceptually adapted PSNR (paPSNR)

$$\mathrm{paPSNR} = 10 \cdot \log_{10} \frac{C^2}{\mathrm{paMSE}}. \tag{10}$$

Note that when distortion sensitivity is available only globally for a full image, with $\mathscr{s}(x, y) = \mathscr{s}_I$ then Eq. 10 simplifies to Eq. 7.
The resulting compensation for local distortion sensitivity is very similar to the normalized weighting scheme often used in the literature [17, 15], but does not employ a image-wise normalization of the weights.
Due to the scarceity of samples, i.e. distortion levels per reference image, no performance limits can be derived for local compensation of distortion sensitivity.

3. Estimation of Distortion Sensitivity using Neural Networks

The neural network used for end-to-end trained image quality estimation proposed in [33] is re-used here for the estimation of patch-wise distortion sensitivity. Input to the network are 32×32 pixel-sized patches of the gray-scale

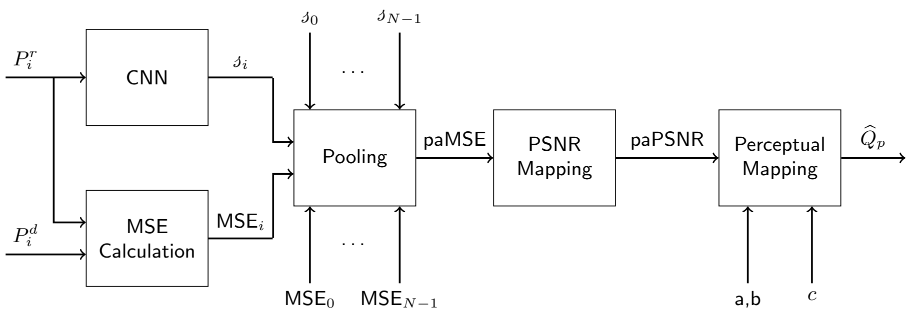


Figure 5: CNN-based compensation of the PSNR for distortion sensitivity. Distortion sensitivity $\mathfrak{s}_i$ is estimated by the CNN from the reference patch P_i^r. The image-wise paPSNR is calculated from all sensitivity-weighted MSEs between collocated reference patches P_i^r and distorted patches P_i^d and mapped into the perceptual domain on the quality estimate $\widehat{Q}_p$.

converted reference image. The proposed CNN comprises 12 weight layers that are used to estimate the distortion sensitivity $\mathfrak{s}_i$ of a given reference image patch P_i^r. 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 as shown in Fig. 6. Convolutional layers are activated through a Leaky Rectified Linear Unit (LReLU) activation function [34] with a leakyness of 0.2. To allow for the estimation of distortion sensitivity for patch sizes other than 32×32 pixels, the network architecture is adapted for the processing of patches of *a)* 8×8 pixels by removing the first two pooling layers; *b)* 16×16 pixels by removing the first pooling layer; *c)* 64×64 pixels by introducing an additional pooling layer succeeding the 7th convolution layer; and *d)* 128×128 pixels by introducing two additional pooling layers succeeding the 7th and the 9th convolution layer.

Analogous to Section 2.4, the distortion sensitivity estimate $\mathfrak{s}_i$ output of the network is used to weight the patch-wise MSE$_i$, measured between a reference image patch P_i^r and the collocated image patch P_i^d from the distorted image. The resulting image-wise paMSE from Eq. 9 leads with Eq. 10 directly to the image-wise paPSNR. The image-wise paPSNR is mapped into the perceptual domain by Eq. 6. Based on previous considerations, parameters a and b are fixed as the lower and upper value of the quality scale used in the psychophysical quality assessment; an additional parallel branch consisting of only 1 weight with a constant input of 1 is used for estimating a global value of c. The overall architecture is sketched in Fig. 5.

Commonly the MSE is used as minimization criterion in regression tasks. However, optimization with respect to mean absolute error (MAE) puts less emphasis on outliers and reduces their influence. Hence, MAE is chosen as a less outlier sensitive alternative to MSE. The loss function to be minimized is then

$$E = |\widehat{Q}_p - Q_p|. \tag{11}$$

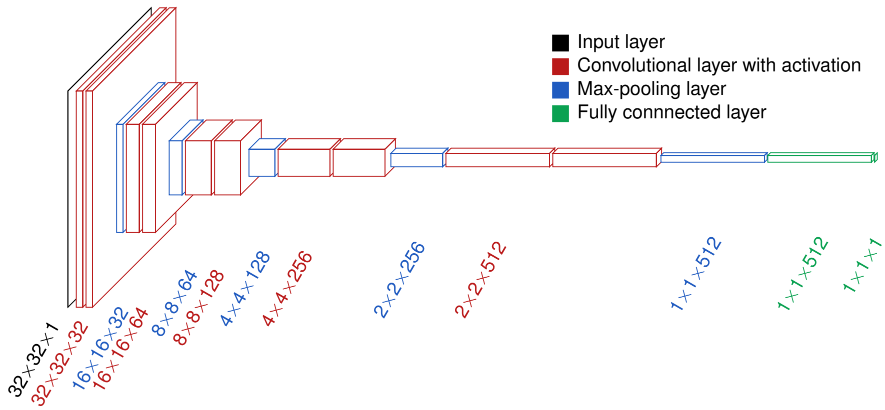


Figure 6: Architecture of the network. Layers are represented by cuboids. Height and depth of a cuboid represent spatial resolution at different levels, width represents the number of channels. The input layer is denoted by a black cuboid, the output of convolutional layers (including LReLu-activations) by red, of max-pooling layers by blue and of fully connected layers by green cuboids. Note that the last layer (output of the last fully connected layer with the dimensionality $1 \times 1 \times 1$) is the output of the network and, as such, holds the distortion sensitivity estimate.

4. Experiments and Results

4.1. Datasets

Experiments are performed on the LIVE [25], TID2013 [26] and CSIQ [27] image quality databases.

The LIVE [25] database comprises 779 quality annotated images based on 29 source reference images that are subject to 5 different types of distortions at different distortion levels. Distortion types are JP2K compression, JPEG compression, additive white Gaussian noise, Gaussian blur and a simulated fast fading Rayleigh channel. Quality ratings were collected using a single-stimulus methodology, scores from different test sessions were aligned. Resulting DMOS quality ratings lie in the range of $[0,\ 100]$, where a lower score indicates better visual image quality.

The TID2013 image quality database [26] is an extension of the earlier published TID2008 image quality database [35] containing 3000 quality annotated images based on 25 source reference images distorted by 24 different distortion types at 5 distortion levels each. The distortion types cover a wide range from simple Gaussian noise or blur over compression distortions such as JPEG to more exotic distortion types such as non-eccentricity pattern noise. This makes the TID2013 a more challenging database for the evaluation of quality models. The rating procedure differs from the one used for the construction of LIVE, as it employed a competition-like double stimulus procedure. The obtained MOS values lie in the range $[0,\ 9]$, where larger MOS indicate better visual quality.

The CISQ image quality database contains 866 quality annotated images. 30 reference images are distorted by JPEG compression, JP2K compression, Gaussian blur, Gaussian white noise, Gaussian pink noise or contrast change. For quality assessment, subjects were asked to position distorted images horizontally on a monitor according to its visual quality. After alignment and normalization resulting DMOS values span the range [0, 1], where a lower value indicates better visual quality.

4.2. Experimental Setup

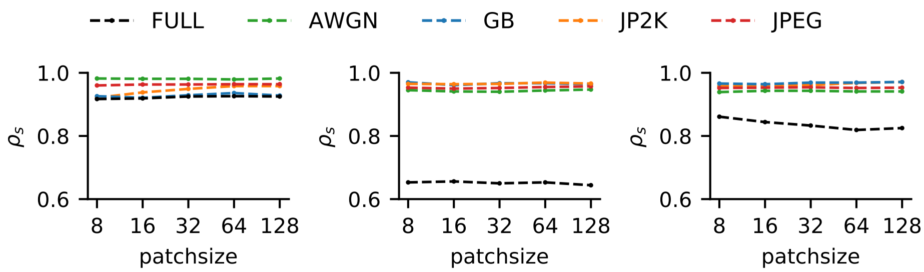
Networks are trained and tested either on LIVE, TID2013, or CSIQ for single database-evaluation. The databases are randomly split in training, validation and test set. To guarantee that no distorted or undistorted version of an image used in testing or validation has been seen by the network during training, the datasets are split by reference image. For each database validation and test set each contain 6 reference images, whereas the training set consists of 17, 13 and 18 reference images for LIVE, TID2013 and CSIQ. Results are reported as the average over 30 random splits. Models are trained for 150 epochs after which the model with the lowest validation loss is selected and tested; this amounts to early stopping [36]. Training and validation of models with an input patch size of 32×32 pixels is based on 32 patches, randomly sampled from one image per iteration. This allows to train the network based on image-wise quality annotations from the datasets. To keep the amount of data seen by the neural network in each training iteration constant for different patch sizes, the number of sampled patches per image is scaled inversely proportionally with the square of the patch size, i.e. 512 patches of 8×8 pixel, 128 patches of 16×16 pixels, 8 patches of 64×64 pixels and 2 patches of 128×128 pixels. Patches are densely sampled, i.e. the full image is considered, for testing.

To assess the generalization ability of the proposed methods the CSIQ image database is used for cross-dataset evaluating the models trained either on LIVE or on TID2013 and models trained for single database evaluation were reused. LIVE and TID2013 share a lot of reference images, thus, tests between these two are unsuitable for evaluating generalization for unseen images. For cross-distortion evaluation, models trained on LIVE are tested on TID2013 in order to determine how well a model deals with distortions that have not been seen during training and in order to evaluate whether a method is truly non-distortion or just many-distortion specific.

Note that, in contrast to many results reported in the literature, if not explicitly stated differently, we use the full TID2013 database and do not ignore any specific distortion type.

4.3. Influence of Patch Size

In a first evaluation, the influence of the patch size is investigated for distortion types that are shared among LIVE, TID2013 and CSIQ and for the full databases. Spearman rank order coefficient (SROCC) obtained with the proposed method is plotted with regard to the patch-size on which distortion sensitivity is estimated in Fig. 7. The prediction monotonicity is surprisingly little affected by the size of the patch on which distortion sensitivity is estimated. As will be discussed in detail in Section 5, this can be explained by



(b) LIVE (c) TID2013 (d) CSIQ

Figure 7: Influence of the patch-size on the prediction performance measured as SROCC on LIVE, TID2013 and CSIQ evaluated for selected distortion types (Gaussian blur, white Gaussian noise, JP2K and JPEG compression) and over the full databases.

the non-normalizing weighting scheme inherent to the paMSE defined in Eq. 9. In accordance with the patch size used in [33], further results in this section are achieved based on distortion sensitivity estimation on 32×32 pixel sized patches.

4.4. Performance Evaluation

Table 3: Average SROCC over 20 runs of the proposed method for the distortion types of LIVE and CSIQ databases and the *actual* subset of TID2013 in comparison to PSNR, SSIM [10], MS-SSIM [9], FSIM [12] and HaarPSI [11].

		PSNR	SSIM	MS-SSIM	FSIM	HaarPSI	**paPSNR**
LIVE	JP2K	0.895	0.961	0.962	0.972	0.968	0.949
	JPEG	0.881	0.976	0.981	0.984	0.983	0.963
	AWGN	0.985	0.969	0.973	0.972	0.985	0.981
	GB	0.782	0.952	0.954	0.971	0.967	0.929
	FF	0.891	0.956	0.947	0.952	0.951	0.941
TID2013	AWGN	0.929	0.865	0.865	0.91	0.937	0.94
	SCN	0.92	0.852	0.854	0.89	0.931	0.944
	MN	0.832	0.777	0.807	0.809	0.786	0.856
	HFN	0.914	0.863	0.86	0.904	0.907	0.948
	IN	0.897	0.75	0.763	0.825	0.867	0.916
	GB	0.915	0.967	0.967	0.955	0.912	0.967
	DEN	0.948	0.925	0.927	0.933	0.947	0.943
	JPEG	0.919	0.92	0.927	0.934	0.951	0.952
	JP2K	0.884	0.947	0.95	0.959	0.97	0.965
	MGN	0.891	0.78	0.779	0.857	0.89	0.934
	LCNI	0.915	0.906	0.907	0.949	0.962	0.963
CSIQ	AWGN	0.936	0.897	0.947	0.936	0.967	0.943
	JPEG	0.888	0.955	0.963	0.966	0.97	0.954
	JP2K	0.936	0.961	0.968	0.97	0.982	0.961
	GPN	0.934	0.892	0.933	0.937	0.954	0.939
	GB	0.929	0.961	0.971	0.973	0.978	0.969
	CTRST	0.862	0.792	0.952	0.944	0.945	0.901

[2]Note that [10] and [11] use the 4-parameter logistic function that is also employed in this work whereas [12], [37], [20], [38], [17] and [39] use a 5-parameter logistic function to regress quality predictions onto MOS values before correlations are computed.

Table 4: Comparison of the proposed method to the state-of-the-art FR image quality estimation models based on the LIVE and TID2013 databases. The highest LCC and SROCC are set in bold. The reported correlation for the proposed models are achieved on the test sets of 30 random train-test splits. Correlations for all other models are taken from the literature[2].

Feature extraction from distorted image		LIVE		TID2013	
		LCC	SROCC	LCC	SROCC
Yes	SSIM [10]	0.945	0.948	0.790	0.742
Yes	FSIM_C [12]	0.960	0.963	0.877	0.851
Yes	GMSD [37]	0.956	0.958	-	-
Yes	DOG-SSIM [20]	0.963	0.961	0.919	0.907
Yes	DeepSim [38]	0.968	**0.974**	0.872	0.846
Yes	HaarPSI [11]	0.967	0.900	0.87	0.863
Yes	IW-PSNR [17]	0.933	0.933	-	0.689
Yes	PSIM [39]	0.958	0.962	0.908	0.893
Yes	DIQaM-FR [33]	0.977	0.966	0.88	0.859
Yes	WaDIQaM-FR[33]	**0.980**	0.97	**0.946**	**0.940**
No	PSNR	0.872	0.876	0.675	0.687
No	$\mathrm{paPSNR}_{\gamma=1}$ (proposed)	0.904	0.925	0.588	0.65
No	$\mathrm{paPSNR}_{\gamma=\gamma^*}$ (proposed)	0.938	0.943	0.863	0.876

The performance of the presented paPSNR-based quality estimation is summarized and compared to related methods for selected distortion types on LIVE, TID2013 and CSIQ in terms of SROCC in Table 3. The proposed method clearly outperforms the PSNR for almost all distortion types and databases. An exception that is observable in all databases is additive white Gaussian noise (AWGN), for which the original PSNR is already a very good predictor and thus difficult to improve. Although applying complex processing on the reference image only, the SROCC of the proposed method is in general close to methods that perform complex processing on the distorted image as well.

Table 4 presents a comparison of the proposed method to state-of-the-art methods, evaluated on the full LIVE and TID2013 databases. Although the proposed method ($\mathrm{paPSNR}_{\gamma=1}$) outperforms the PSNR on LIVE, its prediction accuracy is clearly inferior to all other approaches. Here, the distinction of the proposed approach from methods employing complex processing on the distorted image is important to note; the computational advantage of the proposed approach will be discussed in detail in later. In contrast to the single distortion results shown in Table 4, on TID2013 the proposed approach not only performs inferior to other sophisticated state-of-the-art approaches, but even worse as compared to the PSNR. This can be explained by the distortion type dependency of distortion sensitivity analyzed in Section 2.3.

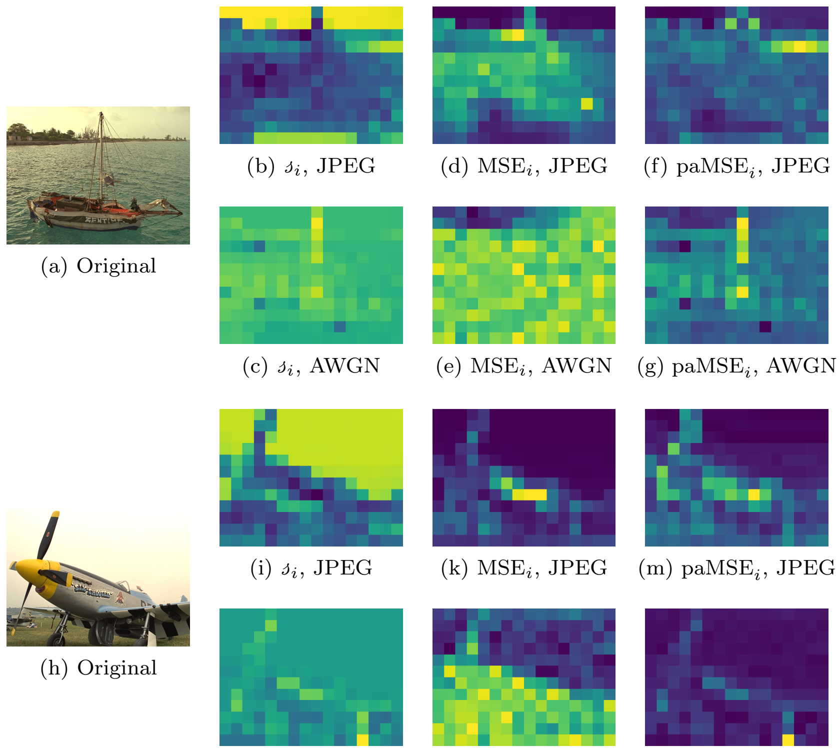
This distortion type dependency can be effectively approximated by simple linear scaling of $\jmath$ with a distortion type-specific factor γ [23]. The scaling is incorporated as an additional trainable parameter into the sensitivity estimation described in Section 3 and γ is distortion type-specific jointly optimized with all other distortion type-agnostic parameters of the network. The resulting performance over the full dataset is referred to as $\text{paPSNR}_{\gamma=\gamma^*}$ in Table 4. Note that this evaluation relies on the (for most applications reasonable) assumption that the distortion type by which the test image is affected is known. As Table 4 shows, considering distortion type dependency increases the prediction performance substantially, especially when tested on TID2013 containing a multitude of different distortion types.

Remarkably, the proposed paPSNR shares eminent conceptual similarity with the information content weighted PSNR (IW-PSNR) [17] which also achieves accuracy improvements for the low-complex PSNR through an image-dependent local weighting function. However, a methodological key difference between the two frameworks lies in the amount of information accessible to the weighting function: whereas local weights are based exclusively on the reference image for paPSNR, both the reference and corresponding distorted image are taken into consideration in case of IW-PSNR. Correlations of the two approaches as listed in Table 4 are therefore not directly comparable. Although a meaningful notion of distortion sensitivity as an image property as described in Section 2 appears to be restricted to the reference image, switching the role of reference and distorted images in the proposed framework allows for an easy adaptation to estimate patch weights based on distorted images. For clarity, models employing this adaptation are denoted as $\text{paPSNR}^{\text{dst}}$. Note that the adapted weighting function employed by $\text{paPSNR}^{\text{dst}}$ still disposes of less information than in the IW-PSNR framework as patch weight estimates are exclusively based on distorted images. Linear Pearson correlation and Spearman rank order correlation on LIVE and TID2013 for this adaptation are listed in Table 6. Performance on LIVE is comparable or even superior to other state-of-the-art methods. In case of TID2013, performance also clearly increases, yet state-of-the-art performance is only accomplished when distortion type dependency is compensated for. In comparison with IW-PSNR, the $\text{paPSNR}^{\text{dst}}$ achieves a superior performance on LIVE as well as clearly higher Spearman rank order correlations on TID2013.

4.5. Local Weights

The spatial distribution of patch-wise estimated distortion sensitivity $\mathscr{s}_i$ and the resulting distortion sensitive MSE is exemplified in Fig. 8 for two reference images and two distortion types, namely JPEG compression and additive white Gaussian noise. Original images are presented in Fig. 8a and Fig. 8h, corresponding sensitivity maps for JPEG compression distorions in Fig. 8b and Fig. 8i and those for AWGN in Fig. 8c and Fig. 8j. Examples for patch-wise MSE maps are visualized in the second from right column, resulting paMSE maps in the right column of Fig. 8. Distortion sensitivity maps are presented in the same color scale representing values of $\mathscr{s}_i$ from 21 to 34, thus are directly comparable. Local distortion sensitivities values lie in a range expected from Fig. 2. Color scales differ between the visualizations of different MSE and paMSE maps in order to use full ranges for each map.

Comparing the distortion sensitivity maps shows that for the case of JPEG distortions, local distortion sensitivity varies largely within the images. While low values of sensitivity are assigned to textured regions of the images, high



(a) Original (b) $\mathscr{s}_i$, JPEG (c) $\mathscr{s}_i$, AWGN (d) MSE_i, JPEG (e) MSE_i, AWGN (f) paMSE_i, JPEG (g) paMSE_i, AWGN

(h) Original (i) $\mathscr{s}_i$, JPEG (j) $\mathscr{s}_i$, AWGN (k) MSE_i, JPEG (l) MSE_i, AWGN (m) paMSE_i, JPEG (n) paMSE_i, AWGN

Figure 8: Examples of local distortion sensitivity for two reference images and two distortion types. The left-most column show the reference images from which patch-wise distortion sensitivity is estimated. The second from left column shows the resulting maps of distortion sensitivity for JPEG compression and AWGN distortions. In the second from right column the patch-wise MSE is shown, the perceptually adapted MSE resulting from patch-wise MSE and patch-wise distortion sensitivity is shown in the right-most column. Low values are represented by blue, high values by yellow. For comparability, colors are aligned for the distortion sensitivity maps.

values of sensitivity are estimated for rather flat areas, e.g. the sky in Fig. 8a and Fig. 8h. This is expected as distortions in textured regions are subject to masking effects, whereas JPEG-specific distortions such as blocking are highly visible in flat areas.

For the case of additive white Gaussian noise, local values of $\mathscr{s}_i$ do not show this wide range of variation, but are relatively uniformly distributed over image. This suggests that, disregarding a global shift, the (unadapted) PSNR already is a good quality predictor for images affected by additive white Gaussian noise. This is in line with the numerical results presented in Section 4.4.

Table 5: Average SROCC over 100 runs of paPSNR trained and tested on different databases for selected distortion types and over full databases.

Trained on	LIVE		TID2013	
Tested on	TID2013	CSIQ	LIVE	CSIQ
JP2K	0.96	0.962	0.945	0.956
JPEG	0.923	0.958	0.949	0.935
AWGN	0.932	0.95	0.983	0.932
GB	0.906	0.97	0.893	0.959
FULL	0.637	0.815	0.897	0.815

4.6. Cross-Database Evaluation

The generalization ability of the neural network-based adaptation of the PSNR is studied in a cross-database evaluation for selected distortions and over full databases. For cross-database evaluation on the full database, no knowledge about the distortion type is assumed, i.e. $\gamma = 1$. The results are presented in terms of SROCC in Table 5.

High generalization ability is achieved for the single distortion case. Given the large amount of reference images shared between LIVE and TID2013, this is not surprising. For single distortions the approach also generalizes well for images unseen during training in CSIQ. Cross-database evaluation over full image databases results in low prediction accuracies. As shown in Section 4.4, the proposed method does not perform well without consideration of the distortion type; hence, high accuracies can neither be expected for distortion-type agnostic cross-database evaluation.

4.7. Weight Estimation on Distorted Images

Table 6: Performance comparison on LIVE and TID2013 databases with models trained on the distorted image instead of the reference image.

	LIVE		TID2013	
	LCC	SROCC	LCC	SROCC
$\text{paPSNR}^{\text{dst}}_{\gamma=1}$	0.971	0.971	0.739	0.741
$\text{paPSNR}^{\text{dst}}_{\gamma^*}$	0.972	0.971	0.898	0.902

Although it does not follow the previously derived concept of distortion sensitivity and gives away the advantage of graceful distribution of complex processing to the reference image only, local weights can in principal also be estimated from the distorted image. The resulting prediction performance is presented in Table 6. The results show that adaptation of the PSNR based on the distorted image achieves higher prediction accuracy compared to adaptation based on the reference image both in terms of Pearson linear correlation coefficient (LCC) and SROCC. From the perspective of distortion sensitivity this is very surprising. However, it was shown e.g. in [40, 33] that a neural network can learn to extract quality related information from the distorted image only; such an information is not available from a reference image. Further, the distorted image contains information about the distortion type [41] that can be exploited by the network to improve prediction accuracy. It can be hypothesised that a network trained on the distorted image in fact learns a different representation compared to a network trained on the references. The inferior performance obtained by predicting 'distortion sensitivity' from a undistorted image by a network trained on distorted images (LCC: 0.877, SROCC: 0.921) and predicting 'distortion sensitivity' from an distorted image by a network trained on undistorted images (LCC: 0.79, SROCC: 0.807) corroborates this conjecture.

5. Discussion & Conclusion

In this paper, a conceptual framework for distortion sensitivity for visual quality estimation was derived. Parameters of the non-linear regression function used to map computational quality values into the psychophysical domain were discussed and functionally interpreted. It was shown and exemplified for the PSNR that the shift parameter of the psychometric mapping function can serve efficiently as a functional definition of distortion sensitivity. Distortion sensitivity was modelled as a distortion type-dependent property of a reference image; being a reference image property allows for an offline estimation of distortion sensitivity. It was shown that compensating for distortion sensitivity can efficiently improve the prediction performance of a given computational quality model. Limits of such approaches were explored quantitatively. A neural network-based method for patch-wise estimation of distortion sensitivity within an image quality estimation framework was presented that significantly improves the quality estimation accuracy of the base quality model, i.e. the PSNR.

The presented definition of distortion sensitivity and the proposed framework for estimation thereof can be easily adapted to other quality models than the PSNR and extended to other signal modalities such as videos, assuming the availability of quality annotated data.

The neural network-based patch-wise compensation for distortion sensitivity significantly improves the performance of the PSNR. However, comparing the achieved performance with the limits determined by (hypothetical) optimal image-wise compensation shows that the method still has further potential for improvement. The sub-optimality indicates that there is some room for improving the generalization ability of the model with regard to unseen images.

Weights used for spatial pooling are commonly normalized. The weighting scheme derived from distortion sensitivity does not comprise any normalization. This also explains the independence of the SROCC from patch-size as non-normalized weights are capable of capturing a global image property (cf. Section 2.2. Imagine one image of high and spatially uniform distortion sensitivity and another image of low and spatially uniform distortion sensitivity. While a non-normalized weighting scheme could differentiate between high and low sensitivity, this information would be lost by normalization of the weights. However, in future work, differences between normalized and non-normalized weighting can be studied within the presented framework. This potentially also brings better understanding on how humans spatially pool perceptual visual quality.

The proposed method works better if local weights are estimated from the distorted images rather than from the reference images. This does not follow the concept of distortion sensitivity that was presented as a property of the reference image and, thus, appears surprising. It is however not unexpected, since networks, as shown in e.g. in [40, 33], are able to predict quality relatively accurately from the distorted image alone as well. More insight into the nature of distortion sensitivity and relevant features driving distortion perception might be gained by investigating differences in the internal representations in networks trained based on the original and distorted images using explaining methods [42, 43, 44]. Also note that the reference image-based models were trained on a smaller sample size regarding the input signal compared to the distortion image-based models, while the number of quality labels is identical. At this point it is not clear how this imbalance impacts the training. However, although achieving higher prediction accuracy, estimating quality based on weights extracted from the distorted image forfeits the crucial advantage of performing complex computations on the reference image only.

The derivation of a distortion sensitive PSNR led to a local weighting scheme for a perceptual adaptation of the MSE. This has a very interesting application perspective, as such a weighting scheme could be, analogously to [45], incorporated into the bit allocation in hybrid block-based video compression. This would directly bridge from psychometric properties to bit allocation for perceptual image compression. The presented approach could play out its real strength, as for mode decision [3] the computationally complex estimation of distortion sensitivity has to be performed only once per reference block, whereas per mode decision iteration only the computationally low-complex MSE has to be calculated. The conceptual decoupling of of distortion sensitivity estimation from quality estimation (enabled by modelling distortion sensitivity as a property of the reference image only) further allows for parallelization and/or offline estimation of distortion sensitivity in time-critical systems. However, as computation time is crucial for real-time systems, the estimation approach and the influence of the network architecture should be thoroughly analysed in terms of computational complexity.

Although the perceptual adaptation of the low-complex MSE is particularly appealing, the proposed framework can be directly applied to other FR quality models.

The discussion of the limits of the proposed framework shows that the availability of quality annotated images and videos is crucial for the success of data-driven approaches to quality assessment. This is especially important for an application such as the previously sketched distortion sensitive bit allocation as most databases do not consider modern compression algorithms such as High Efficiency Video Coding (HEVC) as a distortion type and typically only contain images and videos of resolutions that are practically not of highest relevance any more. Hence, until larger and more suitable database are available, the method could be trained on images annotated by quality models that are computationally less graceful, but more accurate.

Combining the concept of distortion sensitivity with psychophysiological methods [46, 47] for determination for perceptual thresholds such as the sweep-steady-state visual evoked potential (SSVEP) [48, 49] are a promising for a directed assessment of distortion sensitivity, as SSVEP were shown to be highly correlated with perceived quality [50, 51]. Also event-related potentials (ERPs) were shown to be feasible to assess quality at different distortion levels [52].

However, the main advantage and technical motivation of the proposed distortion sensitive quality assessment is not primarily a remarkable high accuracy, but the allocation of computational complex processing to the *reference image* only.

References

References

[1] ITU-R Rec. BT.500-13, Methodology for the subjective assessment of the quality of television pictures (2012).

[2] ITU-T Rec. P.910, Subjective video quality assessment methods for multimedia applications (2008).

[3] T. Wiegand, H. Schwarz, Video Coding: Part II of Fundamentals of Source and Video Coding, Foundations and Trends in Signal Processing 10 (1–3) (2016) 1–346.

[4] Z. Wang, A. C. Bovik, Mean squared error: Love it or leave it? A new look at signal fidelity measures, IEEE Signal Processing Magazine 26 (1) (2009) 98–117.

[5] B. Girod, What's Wrong with Mean-squared Error?, in: Digital Images and Human Vision, MIT Press, 1993, pp. 207–220.

[6] W. Lin, C.-C. J. Kuo, Perceptual visual quality metrics: A survey, Journal of Visual Communication and Image Representation 22 (4) (2011) 297–312.

[7] S. J. Daly, Application of a noise-adaptive contrast sensitivity function to image data compression, Optical Engineering 29 (8) (1990) 977–987.

[8] J. Lubin, A human vision system model for objective picture quality measurements, International Broadcasting Convention (1997) 498–503.

[9] Z. Wang, E. P. Simoncelli, A. C. Bovik, Multiscale structural similarity for image quality assessment, in: IEEE Asilomar Conference on Signals, Systems and Computers, 2003, pp. 1398–1402.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[11] R. Reisenhofer, S. Bosse, G. Kutyniok, T. Wiegand, A Haar wavelet-based perceptual similarity index for image quality assessment, Signal Processing: Image Communication 61 (2018) 33–43.

[12] L. Zhang, L. Zhang, X. Mou, D. Zhang, X. Mou, FSIM: A feature similarity index for image quality assessment, IEEE Transactions on Image Processing 20 (8) (2011) 2378–2386.

[13] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, A. C. Bovik, Deep Convolutional Neural Models for Picture Quality Prediction, IEEE Signal Processing Magazine 34 (November) (2017) 130–141.

[14] B. Ortiz-Jaramillo, J. Niño-Castañeda, L. Platiša, W. Philips, Content-aware video quality assessment: predicting human perception of quality using peak signal to noise ratio and spatial/temporal activity, in: Image Processing: Algorithms and Systems XIII, Vol. 9399, 2015, p. 939917.

[15] W. Zhang, A. Borji, Z. Wang, P. Le Callet, H. Liu, The application of visual saliency models in objective image quality assessment: A statistical evaluation, IEEE Transactions on Neural Networks and Learning Systems 27 (6) (2016) 1266–1278.

[16] V. Laparra, J. Ballé, A. Berardino, E. P. Simoncelli, Perceptual image quality assessment using a normalized Laplacian pyramid, Electronic Imaging 2016 (16) (2016) 1–6.

[17] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Transactions on Image Processing 20 (5) (2011) 1185–1198.

[18] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, C.-C. J. Kuo, Compressed image quality metric based on perceptually weighted distortion, IEEE Transactions on Image Processing 24 (12) (2015) 5594–5608.

[19] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, M. Carli, Modified image visual quality metrics for contrast change and mean shift accounting, in: 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), IEEE, 2011, pp. 305–311.

[20] S.-C. Pei, L.-H. Chen, Image quality assessment using human visual DOG model fused with random forest, IEEE Transactions on Image Processing, 24 (11) (2015) 3282–3292.

[21] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, V. Lukin, On between-coefficient contrast masking of dct basis functions, in: Proceedings of the third international workshop on video processing and quality metrics, Vol. 4, 2007.

[22] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, M. Carli, New Full-Reference Quality Metrics based on HVS, Proceedings of the Second International Workshop on Video Processing and Quality Metrics, (2006) 2–5.

[23] S. Bosse, M. Siekmann, W. Samek, T. Wiegand, A perceptually relevant shearlet-based adaptation of the PSNR, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2017, pp. 315–319.

[24] G. J. Sullivan, J. R. Ohm, W.-J. J. Han, T. Wiegand, Overview of the High Efficiency Video Coding (HEVC) Standard, IEEE Transactions on Circuits and Systems for Video Technology 22 (12) (2012) 1649–1668.

[25] H. R. Sheikh, M. F. Sabir, A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on image processing 15 (11) (2006) 3440–51.

[26] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. J. Kuo, Color Image Database TID2013: Peculiarities and preliminary results, 4th European Workshop on Visual Information Processing (EUVIP) (2013) 106–111.

[27] E. C. Larson, D. M. Chandler, Consumer subjective image quality database (2009).

[28] VQEG, Objective perceptual assessment of video quality: full reference television (2004).

[29] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, W. Lin, Unified Blind Quality Assessment of Compressed Natural, Graphic, and Screen Content Images, IEEE Transactions on Image Processing 26 (11) (2017) 5462–5474.

[30] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, C. W. Chen, Blind quality assessment based on pseudo-reference image, IEEE Transactions on Multimedia 20 (8) (2018) 2049–2062.

[31] D. Ruderman, The statistics of natural images, Network: Computation in Neural Systems 5 (1994) 517–548.

[32] A. J. Bell, T. J. Sejnowski, The 'independent components' of natural scenes are edge filters, Vision Research 37 (23) (1997) 3327–3338.

[33] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Transactions on Image Processing 27 (1) (2018) 206–219.

[34] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.

[35] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, F. Battisti, TID2008-A database for evaluation of full-reference visual quality assessment metrics, Advances of Modern Radioelectronics 10 (January 2016) (2009) 30–45.

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

[37] W. Xue, L. Zhang, X. Mou, A. C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Transactions on Image Processing 23 (2) (2014) 668–695.

[38] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, Y. Zhu, DeepSim: Deep similarity for image quality assessment, Neurocomputing 257 (2017) 104–114.

[39] K. Gu, L. Li, H. Lu, X. Min, W. Lin, A fast reliable image quality predictor by fusing micro- and macro-structures, IEEE Transactions on Industrial Electronics 64 (5) (2017) 3903–3912.

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

[41] A. K. Moorthy, A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Transactions on Image Processing 20 (12) (2011) 3350–3364.

[42] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (7) (2015) e0130140.

[43] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K. R. Müller, Explaining nonlinear classification decisions with deep Taylor decomposition, Pattern Recognition 65 (May 2016) (2017) 211–222.

[44] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digital Signal Processing 73 (2018) 1–15.

[45] S. Bosse, C. Helmrich, H. Schwarz, D. Marpe, T. Wiegand, Perceptually optimized QP adaptation and associated distortion measure, in: Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-H0047, Macao, China, 2017.

[46] S. Bosse, K.-R. Müller, T. Wiegand, W. Samek, Brain-computer interfacing for multimedia quality assessment, in: Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 2834–2839.

[47] U. Engelke, D. Darcy, G. Mulliken, S. Bosse, M. Martini, S. Arndt, J.-N. Antons, K. Chan, N. Ramzan, K. Brunnström, Psychophysiology-Based QoE Assessment: A Survey, IEEE Journal of Selected Topics in Signal Processing 11 (1) (2017) 6–21.

[48] C. W. Tyler, P. Apkarian, D. M. Levi, K. Nakayama, Rapid assessment of visual function: an electronic sweep technique for the pattern visual evoked potential., Investigative Ophthalmology & Visual Science 18 (7) (1979) 703.

[49] J. M. Ales, F. Farzin, B. Rossion, A. M. Norcia, An objective method for measuring face detection thresholds using the sweep steady-state visual evoked response., Journal of Vision 12 (10) (2012) 1–18.

[50] S. Bosse, L. Acqualagna, W. Samek, A. K. Porbadnigk, G. Curio, B. Blankertz, K.-R. Müller, T. Wiegand, Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition, IEEE Transactions on Circuits and Systems for Video Technology 28 (8) (2018) 1694–1706.

[51] L. Acqualagna, S. Bosse, A. K. Porbadnigk, G. Curio, K.-R. Müller, T. Wiegand, B. Blankertz, EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs)., Journal of Neural Engineering 12 (2) (2015) 026012.

[52] S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K.-R. Müller, T. Wiegand, Toward a direct measure of video quality perception using EEG., IEEE Transactions on Image Processing 21 (5) (2012) 2619–29.