QUALITY ASSESSMENT OF IMAGE PATCHES DISTORTED BY IMAGE COMPRESSION USING CROWDSOURCING

Sebastian Bosse[1], *Mischa Siekmann*[1], *Jennifer Rasch*[1],
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

Three experiments addressing the assessment of perceived image quality in a patch-based manner are compared for HEVC compression artifacts. It is shown that image patches of a size small as 128x128 pixel are large enough to evaluate the perceived image quality in a Degradation Category Rating (DCR) setting. Ratings obtained with 128x128 pixel sized images patches and 512x512 pixel sized images of the same spatial statistics show a correlation of r=0.99. Based on this finding, image quality assessment of 128x128 pixel sized image patches degraded by HEVC compression is compared for controlled lab environment and uncontrolled crowdsourcing settings. Although we find high overall correlation between the quality ratings obtained in the two environments, observers tend to give worse ratings in the crowdsourcing setting and for conditions of higher quality a reduction of correlation is observed. These findings have implications for choosing controlled vs. uncontrolled viewing conditions for image quality assessment for real-life applications.

Index Terms— image quality assessment, image coding, psychophysics, quality rating, crowdsourcing, mean opinion score, perception threshold, HEVC

1. INTRODUCTION

A huge amount of data generated, stored and transmitted today is visual information such as images or videos. In order to allow for transmission or storage at bit rates suitable for today's channels or storage devices, these signals are digitized and usually compressed. Compression not only decreases the number of bits used for the representation of the source signal, but also introduces distortions into the signal visible to humans. Thus, an encoder balances the trade-off between bit rate and distortion. Typically, in modern image and video coding algorithms such as the state-of-the-art High Efficiency Video Codec (HEVC) [1], this trade-off is decided on in a block-wise manner, as the image or video signal is subdivided into image or frame blocks to be encoded in succession.

In most multimedia systems, humans are the ultimate receiver of processed and/or transmitted signals. Evidently, the measurement of visual quality degradation is essential in most video transmission systems, e.g. for controlling the encoder or for the assessment of quality properties of a whole transmission system. For multimedia applications within the scope of entertainment, the operational point of an image or video transmission system is usually desired to be close to the perception threshold at the lowest possible bit rate in order to provide the consumer a high quality of experience in terms of visual quality, latency and storage size. However, the determination of this operational point is still a challenge as no satisfying measure for visual quality is at hand.

Although the field of image and video quality assessment has been studied for many years, the question of how to quantify visual quality (or reversely phrased: visual distortion) computationally remains unsatisfyingly answered. Hence, for testing the performance of computational quality estimators and to quantify novel kinds of distortion, psychophysical tests are commonly performed in order to provide ground truth in terms of quantified perceived quality.

In such psychophysical experiments, a human observer is presented with a stimulus and gives an overt response regarding the perceived quality of the stimuli. The typical procedure in any of these experiments is that the human observer has to rank or rate the quality of a set of test images and/or videos. This may be done with or without showing an explicit or hidden reference. These subjective tests are widely used in practice, e.g. for the evaluation of image and/or video transmission systems, and provide quantitative quality assessments for visual signals when averaged over many human observers. Quantified visual quality is then usually referred to and reported as Mean Opinion Score (MOS). Such subjective testing methodologies have been formalized by the International Telecommunication Union (ITU) in [2, 3] and more elaborated methods are still an ongoing research area. However, for the sake of reproducibility, those tests are usually performed in a laboratory environment under highly controlled viewing conditions [2, 3]. Since multimedia content in real life is usually not consumed under laboratory conditions but in e.g. living-room-like environments, it can be argued that psychophysical tests performed in a lab are somewhat unnat-

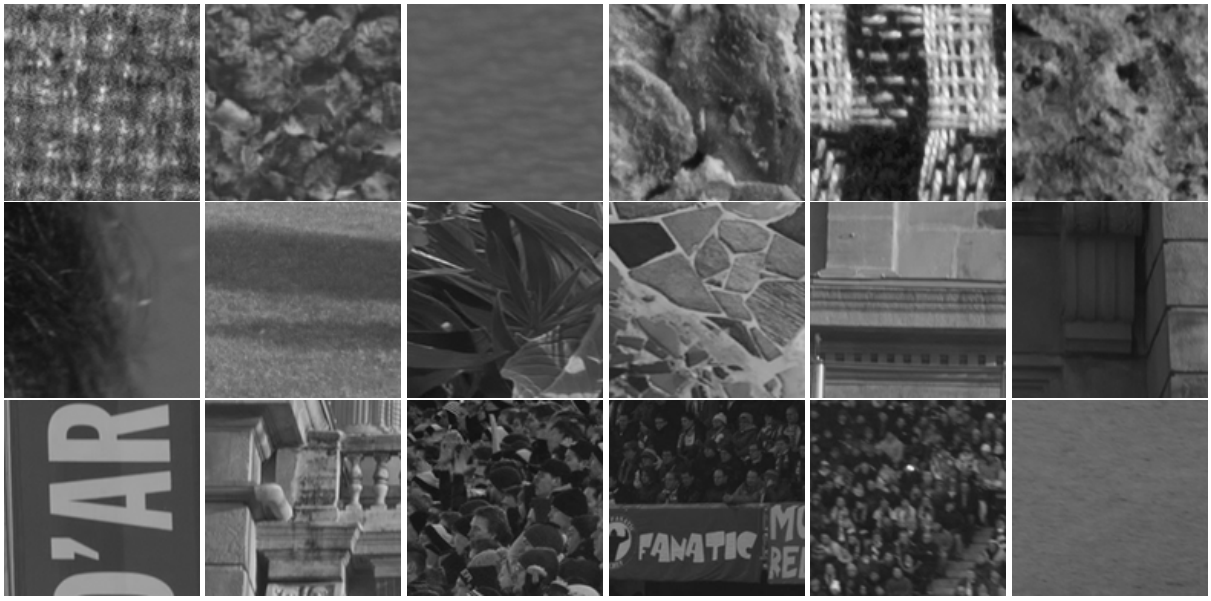

Fig. 1: Image patches used in the experiments. First row: Patches cropped from texture images. Middle and bottom rows: Patches cropped from images with real life content.

ural. As experiments should not take more than 30 minutes in order to prevent observers from becoming unreliable because of fatigue, the number of possible conditions (image/quality pairs) that can be evaluated in one test is restricted. The effort of psychophysical tests is further increased by the need of having ratings of sufficient (15 is recommended in [2]) participants. Being lab-bounded and the need for an experimenter limits the possibilities to parallelize those tests, confining them to the domain of *small data*.

In order to delegate the assessment of perceived visual quality into the *big data* domain, it has been suggested to crowdsource the evaluation (e.g. in [4] for listening test and in [5] for visual quality). Here, the basic idea is to bring the stimuli over the internet to many participants and let them evaluate the perceived quality browser-based at their home PCs in a non-lab environment. This allows for a massive parallelization as the bottleneck of the laboratory access is brought out of consideration. This means that the viewing conditions of the test participants are not only uncontrolled, but that the experimenter is also uninformed about them. On the other hand it can be argued that the viewing conditions in a crowdsourcing approach are closer to those of a real-life multimedia consumer and thus more natural. Several systems for crowdsourcing image and video quality assessment have been presented [4, 5, 6, 7, 8, 9, 10, 11] and overviewed in [12]. In general, high correlations between quality ratings in terms of mean opinion scores obtained in the lab and obtained by crowdsourcing have been reported, indicating the feasibility of the crowdsourcing approach to image and video quality assessment.

So far, psychophysical quality assessment has been performed on whole images only and, to the knowledge of the authors, no studies on patch-based approaches to image quality assessment in order to learn about locality of quality have been carried out.

This paper aims at patch-based quality assessment, as this perspective is closer to the optimization in real-life block-based video and image coding. For that, we present three quality assessment studies. Sec. 2 describes the experimental setup and the stimulus material of these studies. In Sec. 3, at first we analyze whether quality assessment is possible for relatively small patches of a size of 128×128 pixels. Then image quality assessment under controlled lab conditions is compared to crowdsourced image quality assessment for image patches of 128×128 pixel. Sec. 4 concludes the comparisons and discusses implications for practitioners.

2. EXPERIMENTAL DESIGN

2.1. Stimulus Material

In the first experiment, MOS values for 6 texture images used in [13, 14, 15] were obtained under lab-conditions.

For the other two experiments, the 18 image patches used as stimuli had two sources. In order to test if the assessment of perceived image quality based on rather small 128×128 pixel sized image patches is valid, 6 patches were cropped from 512×512 pixel sized texture images from the first experiment. Texture images were chosen for this test. Due to the relative homogeneous structure of textures, cropping does not alter the image statistics. The other 12 patches were cropped from 1920×1080 pixel sized natural grayscale images, show-

Fig. 2: Example of distorted images used in both experiments.

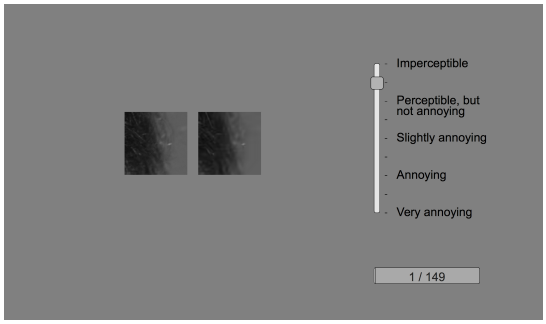

Fig. 3: Screenshot of the rating screen used in the online and the lab quality assessment experiment.

ing real life content. Fig. 1 shows the reference image patches used in the test. Those images were degraded to different distortion levels. The distortions were introduced by encoding the images using the HM13.0 [16] test model of the High Efficiency Video Coding (HEVC/H.265) standard [1] using *Intra only*-settings [17] and different quantization parameters (QP). We refer to an image of a specific size and of a specific distortion level used in the experiments as (test) condition. For all conditions, image patches were cropped from the compressed and thus distorted images. No image patch was scaled after cropping and thus preserving the local statistics of the original images.

Fig. 2 gives an impression over the distortions presented in the test. As the study's goal is to learn more about image distortions close to the perception threshold, the density of test images was higher in the upper end of the quality scale.

2.2. Experimental Setup

Two different experimental setups were used: A laboratory-based, controlled setup and an uncontrolled, online (or: crowdsourced) setup. In the first experiment, the perceived quality of the distorted 512×512 pixel sized texture images was assessed lab-based. In the other two experiments, the 128×128 pixel sized image patches were used as stimuli in a lab-based, and in an online experiment, respectively.

In all experiments, the quality assessment followed a Degradation Category Rating (DCR) procedure using Simultaneous Presentation [2]. Image pairs were presented side-by-side with the distorted image on the right hand side and the undistorted reference image on the left hand side within a 50% gray background. The presentation of the stimuli was self-paced in both cases, leaving the duration of evaluation to the observers. On the right of the screen, a slider operated by a computer mouse was shown to let the observer report his or her quality evaluation. A screenshot of the stimulus presentation screen is shown in Fig.3.
A nine-grade degradation scale was used where the ratings 1, 3, 5, 7 and 9 corresponded to the semantic annotations *Very annoying*, *Annoying*, *Slightly annoying*, *Perceptible, but not annoying* and *Imperceptible*, respectively. In this scale, grade 8 is considered as the psychophysical perception threshold of the impairment [2].
Learning effects were reduced by including a training session in which 5 stimuli were presented at the beginning of each session. Stimuli presented during the training session were not included in the statistical analysis of the test results. In the test session, every condition's presentation was replicated once, resulting in two presentations per condition per session. Before evaluating the collected quality ratings, subjects were screened for reliability according to BT.500 [2].

In order to make the online and lab-based experiments comparable, subjects were introduced in all experiments with the same instructions (although an experimenter was present during the lab sessions) with two successive texts on screen. In these texts, participants were welcomed, the test procedure was explained and subjects were asked to reset the zooming factor of their browsers and maintain a constant viewing distance.

In all tests, subjects were recruited mainly among students at the department and friends or acquaintances of the lab members and not compensated for their participation.

2.2.1. Lab-based Image Quality Assessment

For the lab-based image quality assessment experiment, viewing conditions were controlled in a lab environment and set according to ITU recommendation BT.500 [2], and P.910 [3], respectively. Stimuli were presented on a calibrated $27''$ Dell U2711b display at its native resolution of 2560×1440 pixel. Viewing distance was set to 1 meter, which corresponds to 4 times of the active screen size of the images (relative to the original image size of 1920×1080 pixel) where the patches where cropped from. Relative to the patch size of 128×128 pixels, this corresponds to a visual angle of the $\alpha = 1.72°$.

2.2.2. Crowdsourcing Image Quality Assessment

For the crowdsourced image quality assessment, the same test was delivered to participants over the internet and quality

ratings were reported online through a browser application. Since being out of the lab is inherent to the crowdsourcing approach to image quality assessment, precise viewing conditions, such as screen resolution, contrast and luminance settings, viewing distance and environment illumination, are uncertain.

3. EVALUATION

Neither in the online nor in the lab-based tests subjects had to be rejected after screening [2]. For all evaluations, perceived quality is calculated in terms of mean opinion scores (MOS) and is calculated per condition as the average of all subjects ratings [2].

3.1. Patch-Based Quality Assessment

Pearson correlation between MOS values obtained under lab conditions for 128×128 and 512×512 pixel sized texture images is found as $r_P = 0.99, p < 0.01$. Spearman rank order correlation is measured as $r_S = 0.96, p < 0.01$. Fig. 4 scatters the MOS values of both tests. The high correlation between the MOS values obtained for the two different image patch sizes shows the validity of quality assessment based on 128×128 pixel sized image patches.

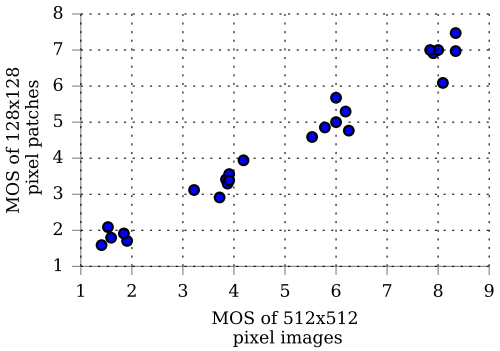

Fig. 4: MOS values obtained for 128×128 pixel sized texture image patches vs. MOS values obtained for 512×512 pixel sized texture image patches

3.2. Lab-Based vs. Crowdsourced Quality Assessment

For the further evaluation, the ratings gathered in the lab-based assessment are considered as ground truth in the sense that the crowdsourced online assessment is desired to predict the results obtained in the lab.

Fig. 5 plots the normalized averaged 0.95%-confidence interval for the MOS values obtained under lab conditions (labMOS) and those obtained in a crowdsourcing, online approach (webMOS) in dependence of the number of observances (since participants in the online test were free to quit

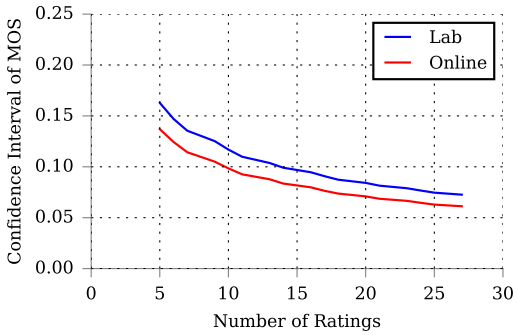

Fig. 5: Normalized averaged 0.95%-confidence interval in dependance of the number of ratings.

the quality evaluation any time, we choose 'observances' or 'number of individual ratings' to be more meaningful than the commonly used parameter 'number of observers' in order to make the two modalities comparable). Assuming Gaussianity of the ratings, the 0.95%-confidence interval is therefore calculated for each condition as $\mathrm{CI}(c, N) = t_{N-1,0.95} \cdot \frac{\sigma_c}{\sqrt{N}}$ with $t_{N-1,0.95}$ being the value of Student's t distribution according to a two-sided 0.95%-critical regions with $N - 1$ degrees of freedom, σ_c being the standard deviation of condition c and N being the number of observances. The conditions $\mathrm{CI}(c, N)$ are then averaged over all conditions using bootstrapping by averaging confidence intervals of randomly resampled permutations of subsets of N observances. Normalization is performed with respect to the range of possible ratings to make the confidence interval comparable to other rating scale ranges. The normalized, averaged confidence intervals fall in the range of previous studies [18], indicating a reliable statistical power in terms of number of subjects/observances in these studies. As Fig. 5 shows, according to the confidence interval, among the two methods, the online assessment of perceived image quality provides a more accurate estimate of MOS than the lab-based assessment given the same number of observances. This is an interesting and puzzling result, since, intuitively, one would assume the crowdsourced environment to be less controlled, leading to a wider range on quality opinion, leading to a higher variance and thus, for the same number of ratings, resulting in a higher confidence interval.

Fig. 6 shows the distribution of individual ratings collected in the lab-based assessment (blue histogram) and in the crowdsourced online assessment (red histogram). The distribution of the rating obtained in the lab are as expected: due to the choice of the study the distribution histogram is skewed to the right (higher quality). Mid-points between semantically annotated rating values (2,4,6,8) are less frequently selected by the observers than semantically annotated ratings (1,3,5,7,9). Similar rating behavior is reported in comparisons of other image quality assessment procedures [18]. The distribution of individual opinion scores collected in the online

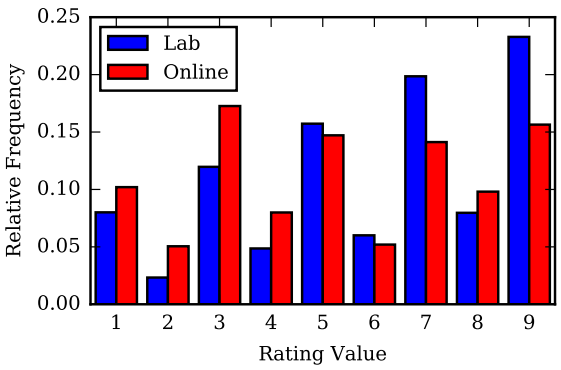

Fig. 6: Distribution of individual quality ratings in lab-based experiment (blue) and crowdsourced experiment (red).

experiment is less skewed, but almost follows a uniform distribution for the semantically annotated rating values (or the mid-points, respectively), while again, the mid-points are less often selected.

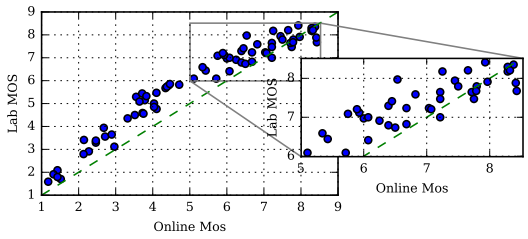

Fig. 7: Online assessed MOS vs. lab-based assessed MOS for the full quality regime (left) and the high quality regime zoomed in (right). The green dashed line indicates the hypothetical identity of labMOS and webMOS.

The tendency of online assessment towards worse quality ratings compared with scores obtained in the lab environment can be seen in the left plot of Fig. 7 as well: The MOS values obtained for different conditions (image/distortion pairs) in the lab are scattered against MOS values gathered in the online test. The green bisecting line indicates hypothetical identity of the MOS values gathered in the different modalities. The plot indicates that the online assessed MOS as predictor of the lab-based assessed MOS has a negative bias that is not constant. For conditions of a labMOS > 7 this bias does not exist.

However, although the MOS values from the lab and those from the online experiment are not identical, over the whole range of conditions, a Pearson correlation of $r_P = 0.96$, $p < 0.001$ and a Spearman rank correlation of $r_S = 0.96$, $p < 0.001$ can be measured. This is in line with previous studies [7, 5] reporting similar high correlations between the MOS values obtained in a lab experiment and in an online experiment.

The right plot in Fig. 7 shows the labMOS plotted against the webMOS for the high quality regime. We define all conditions to be within this regime if labMOS > 6 or webMOS > 6. Here, we find the MOS values obtained in the two modalities to be more scattered. Also, the correlations ($r_P = 0.8$ with $p < 0.001$, $r_S = 0.74$ with $p < 0.001$) are strongly reduced compared to the full range of conditions.

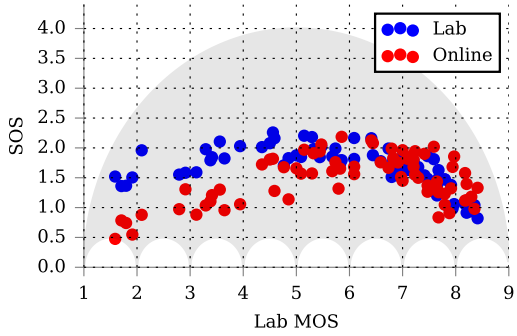

Fig. 8: Standard deviation of opinion scores (SOS) in dependance of MOS for online and lab-based assessed quality vs. lab-bases assessed MOS. The gray shaded area indicates the possible SOS values.

As argued in [19], it is not sufficient to compare MOS values in order to compare test results, since condition-wise differences in dispersion of individual ratings are not taken into account. In Fig. 8, the standard deviation of the opinion scores (SOS) of the two modalities (blue: lab, red: online) is scattered against the labMOS values. For the low quality regime (defined as labMOS < 6), we find the quality ratings obtained in the lab environment to have a bias towards higher dispersion compared to the those obtained in a crowdsourcing setting. This also explains the higher average confidence interval observed in Fig. 5.

4. DISCUSSION AND CONCLUSIONS

This paper discussed the feasibility of patch-based image quality assessment and found that humans can evaluate perceived quality on patch size of 128×128. For image patches of this size image quality assessment in a controlled laboratory environment are compared to image quality assessment in an uncontrolled crowdsourcing setting. We find quality ratings obtained as webMOS and labMOS to be of comparable statistical power. Although the quality ratings obtained in the two different environments are highly correlated over the whole range of conditions, quality ratings obtained by crowdsourcing show a bias towards lower ratings given the same conditions compared to quality ratings gathered under controlled lab conditions. We also find a clear drop in correlation when we compare only conditions in a regime of high quality.

A possible explanation for the bias might be that people

in the online setting adapt to the stimulus, e.g. by adjusting the viewing distance to the specific task (in the study's case: detection and evaluation of image degradation) and stimulus distortion. This can also explain the confidence intervals of the webMOS to be lower than the one of the labMOS. As the condition specific inter-subject dispersion of the webMOS is not higher than the one of the labMOS, it can be argued that the bias indicates the crowdsourcing approach estimates some 'lower bound' of perceived quality of a specific condition. When designing image quality assessment studies, this is to be kept in mind. Thus, future work should investigate reproducibility of results obtained by crowdsourcing compared to those obtained in controlled lab settings.

Our next studies will further evaluate the influence of different image patch sizes to image quality assessment. A question directly connected to the image patch size is how image quality gets pooled along different dimensions such as spatial position, scale and time. Effects found and studied psychophysically should also be evaluated in psychophysiologically oriented experimental settings using EEG [13, 14, 15]. By learning more about the different aspects of stimulus locality, block-based coding schemes might be improved significantly.

5. REFERENCES

[1] G. J. Sullivan, J. 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.

[2] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," Tech. Rep., International Telecommunication Union, Geneva, Switzerland, 2012.

[3] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Tech. Rep., International Telecommunication Union, Geneva, Switzerland, 2008.

[4] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *ICASSP*, 2011, pp. 2416–2419.

[5] F. Ribeiro, "Crowdsourcing subjective image quality evaluation," in *ICIP*, 2011, pp. 3158–3161.

[6] D. Ghadiyaram and A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *arXiv:1511.02919*, 2015.

[7] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowda framework for crowd-based quality evaluation," in *PCS*, 2012, pp. 245–248.

[8] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," in *ACM Int. Conf. Multimed.*, 2012, pp. 359–368.

[9] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Video quality evaluation in the cloud," in *Pack. Video Work.*, 2012, pp. 155–160.

[10] O. Figuerola Salas, V. Adzic, and H. Kalva, "Subjective quality evaluations using crowdsourcing," *PCS*, pp. 418–421, 2013.

[11] O. Figuerola Salas, V. Adzic, A. Shah, and H. Kalva, "Assessing internet video quality using crowdsourcing," *CrowdMM*, pp. 23–28, 2013.

[12] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment," in *MMSP*, 2014, pp. 22–24.

[13] L. Acqualagna, S. Bosse, A. K. Porbadnigk, G. Curio, K.-R. Müller, T. Wiegand, and B. Blankertz, "EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs)," *Journal of Neural Engineering*, vol. 12, no. 2, 2015.

[14] S. Bosse, L. Acqualagna, A. K. Porbadnigk, B. Blankertz, G. Curio, K.-R. Muller, and T. Wiegand, "Neurally informed assessment of perceived natural texture image quality," in *ICIP*, 2014, pp. 1987–1991.

[15] S. Bosse, L. Acqualagna, A. K. Porbadnigk, G. Curio, K.-R. Müller, B. Blankertz, and T. Wiegand, "Neurophysiological assessment of perceived image quality using steady-state visual evoked potentials," in *SPIE Optical Engineering+ Applications*, 2015, pp. 959914–959914.

[16] JCT-VC, "Subversion Repository for the HEVC Test Model reference software," 2014.

[17] F. Bossen, "Common test conditions and software reference configurations," document JCTVC-H1100 of JCT-VC, 2012.

[18] D. M. Rouse, R. Pépion, P. Le Callet, and S. S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment," in *IS&T/SPIE Electr. Imag.*, 2010, pp. 75270F–75270F.

[19] T. Hoßfeld, R. Schatz, and S. Egger, "SOS : The MOS is not enough," in *QoMEX*. IEEE, 2011, pp. 131–136.