Towards a Direct Measure of Video Quality Perception using EEG

Simon Scholler, Sebastian Bosse, Matthias S. Treder, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Müller, and Thomas Wiegand, *Fellow, IEEE*

***Abstract*—An approach for the direct measurement of video quality change perception using electroencephalography (EEG) is presented. Subjects viewed 8s video clips while their brain activity was registered using EEG. The video signal was either uncompressed at full length or changed from uncompressed to a lower quality level at a random time point. The distortions were introduced by a hybrid video codec. Subjects had to indicate whether or not they had perceived a quality change. In response to a quality change, a positive voltage change in EEG (the so-called P3 component) was observed at a latency of about 400-600 ms for all subjects. The voltage change positively correlated with the magnitude of the video quality change, substantiating the P3 component as a graded neural index of video quality change perception within the presented paradigm. By applying machine learning techniques, we could classify on a single-trial basis whether or not a subject perceived a quality change. Interestingly, some video clips wherein changes were missed (i.e., not reported) by the subject were classified as quality changes, suggesting that the brain detected a change although the subject did not press a button. Concluding, abrupt changes of video quality give rise to specific components in the EEG that can be detected on a single-trial basis. Potentially, a neurotechnological approach to video assessment could lead to a more objective quantification of quality change detection, overcoming the limitations of subjective approaches (such as subjective bias and the requirement of an overt response). Furthermore, it allows for real-time applications wherein the brain response to a video clip is monitored while it is being viewed.**

***Index Terms*—Video quality, perception, electroencephalography, EEG, video coding.**

I. Introduction

VIDEO signals are typically intended to be viewed by humans. For their transmission at bit rates suitable for today's channels or storage devices, these signals are digitized and potentially compressed. With an increasing reduction in bit rate, the compression algorithm starts to introduce distortions that are visible to humans. The measurement of these distortions is essential in most video transmission tasks, e.g. for controlling the trade-off between bit rate and distortion or for assessing the visual quality of a transmitted video signal.

One approach towards measuring subjective distortion has been the modeling of the human visual system [1] [2]. The basic idea for obtaining such a model is to measure transfer functions given various temporal and/or spatial stimuli and then to combine these measurements into a more complete model. Such approaches led to a profound understanding of the limitations of the human visual system, e.g. the spatio-temporal limits of perception.

But the question at hand how to quantify visual distortion remained answered unsatisfyingly by these approaches. One reason is that various top-down mechanisms (which are difficult to model) influence the sensitivity of humans to distortions. The world, as we see it, is based on a number of inferences that are related to the visual input like motion or depth, or content and context-related inferences. These inferences are a major part of (active) perception and constitute the way we perceive. A model that does not dwell on these top-down processes will remain incomplete.

Hence, a precise model for the subjective perception of distortion is not available The most common current approach for quantifying subjective distortion still is a judgment experiment: a human observer is presented a stimulus and gives an overt response. Such subjective testing for visual quality assessment has been formalized in Rec. ITU-R BT.500-11 [3] for television applications and Rec. ITU-T P.910 [4] for multimedia applications. The typical procedure in any of these recommendations is that the subject has to rank the quality of a set of test videos. This may be done with or without showing a reference video. These subjective tests are widely used in practice and deliver quality assessments for video signals when averaged over many subjects. They share the drawback that ratings are highly variable across subjects. Further, these ratings given by the human observer are the result of a conscious process, which may be inferred by various aspects and which is prone to be affected by subjective factors (e.g., bias, expectation, strategies).

A potential solution, proposed here, is to directly monitor brain activity during the observation of video clips using electroencephalography (EEG). For the first time, measurements of brain activity are used to quantify the perception of a human observer when being shown a change in video quality. Our approach capitalizes on the P3 component (or

This study was partly supported by the Bundesministerium für Bildung und Forschung (BMBF), Fkz 01IB001A/B, 01GQ0850.

Correspondence: Simon Scholler is with the Machine Learning Laboratory, Berlin Institute of Technology, Berlin, Germany. E-mail: simon.scholler@tu-berlin.de

Sebastian Bosse is with Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute (Fraunhofer HHI), 10587 Berlin, Germany (email: sebastian.bosse@hhi.fraunhofer.de).

Matthias S. Treder, Benjamin Blankertz, and Klaus-Robert Müller are with the Machine Learning Laboratory, Berlin Institute of Technology, Berlin, Germany (e-mail: matthias.treder@tu-berlin.de, benjamin.blankertz@tu-berlin.de, klaus-robert.mueller@tu-berlin.de)

Gabriel Curio is with the Department of Neurology and Clinical Neurophysiology, Charité, Berlin, Germany (email: gabriel.curio@charite.de)

Thomas Wiegand is jointly affiliated with the Image Processing Department, Fraunhofer HHI, and the Image Communication Chair, Berlin Institute of Technology, 10587 Berlin, Germany (email: twiegand@ieee.org).

Manuscript received July 21, 2011; revised January 3, 2012.

P300 component), a large positivity that is usually observed 300–500 ms after a rare and/or significant event. Its amplitude peaks over central-parietal brain regions [5]. The oddball-paradigm is the classical paradigm in which a P3 response is elicited: Frequent and infrequent stimuli are shown and a P3 component is found in response to infrequent stimuli. The P3 reflects cognitive processing and is observed independent of the sensory modality. In contrast to a visually evoked potential (VEP), it is not directly linked to sensory processing. Despite the high amount of noise in the EEG[1], single-trial detection of the P3 component has been demonstrated in applications such as visual [7], [8] and auditory [9], [10] brain-computer interfaces (BCIs) and in audio quality assessment [11]. For reviews on BCIs and BCI technology, refer to [12]–[14].

This paper is organized as follows. In Section II, the viewing experiment including the stimuli and EEG is described. The measured EEG signals and their analysis is described in Section II-C. Results are presented in Section III followed by a discussion on future directions.

II. Viewing Experiment

Our incentive was to investigate whether the (possibly subconscious) perception of video codec artifacts can be measured with EEG. To this end, an experiment was conducted in which subjects watched short video clips, some of which featured a sudden quality change from high quality to a lower quality level during the video (Section Section II-B). In order to make the measurement independent of the image statistics at the actual gaze position at the time of the quality change, stimulus material was chosen that is spatially and temporally roughly homogeneous.

A. Stimulus material

In order to obtain more control over the properties of the video stimuli than would be possible with real world video sequences, the stimuli were generated synthetically. By using video material without semantically meaningful content and the absence of salient objects, influences due to high level image understanding were minimized. Furthermore, by using homogeneous stimuli, we assumed the exact position of the eye gaze to be of little effect on the experimental data. On the other hand, the video material should not be too abstract, but contain simple real-world textures and motion.

In order to meet these requirements, video sequences were generated based on a synthetic image of a textured checkerboard as shown in Figure 1. The image was deformed over time by simulating a swaying water surface on top of the checkerboard. The deformation was calculated by solving the two-dimensional wave function $\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - p\frac{\partial u}{\partial t}$ (v being speed of wave, p decay of wave, t time, x,y image coordinates) iteratively using a Finite-Differences-approach [15]. The deformation $u(x, y, t)$ was applied to the image by convolution, resulting in a smooth and modest movement. Reflections have not been taken into account.

[1]In simulations by [6], only half of the surface scalp potential comes from sources within a 3 cm radius around the electrode.

The sequence was generated in a resolution of 832 by 480 pixels, with a frame rate of 60 frames per second and a duration of 8 seconds.

From this undistorted sequence 500 new sequences have been generated by introducing one out of 10 magnitudes of quality drops in one out of 50 time instances into the undistorted sequence. The time instances of the quality change were chosen randomly uniformly distributed between 2 seconds and 6 second, in other words between the 121th and 360th frame. This assured that the subject cannot predict the time point of a quality change. The magnitude of the quality change was controlled by the quantization parameter of the video coder (see below). These 500 sequences with one quality change each and the undistorted sequences served as stimuli. The different stimuli have been labeled with QC1 to QC10 and QC0, were QC1 denotes the most subtle quality change, QC10 the strongest quality change and QC0 is the undistorted sequence.

Prior to each presentation of a stimulus, a central fixation point was shown for 1 second. Figure 1 sketches the time course of the stimuli and gives an example of the appearance of a stimulus over time.

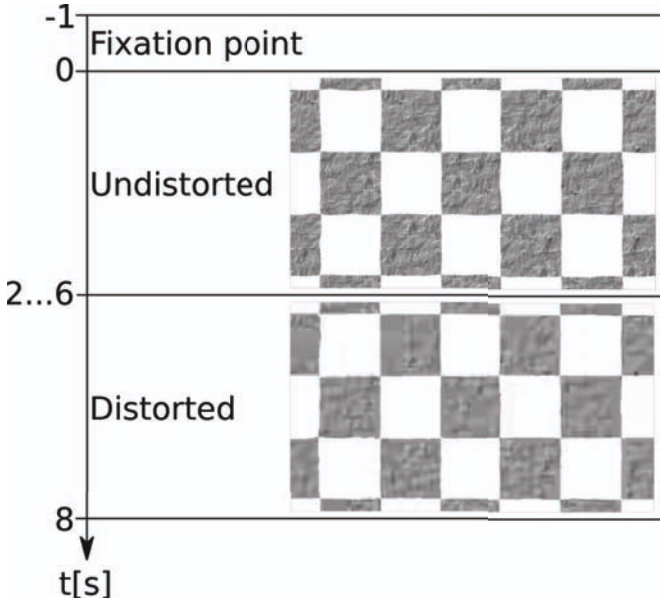

Fig. 1. Time course of the video sequence. For 1 second, a fixation point is shown. For at least 2 seconds, the undistorted video is presented. At a random time point (uniformly distributed between 2 and 6 seconds), the video quality drops instantaneously. On the right, undistorted and distorted (highest quality loss QC10) frames are shown exemplarily.

The quality loss considered in the experiments was induced by lossy compression of the synthesized video sequence. The used video coder is a state-of-the-art, hybrid, motion-compensated, block-based coder [16]. It is architecturally similar to the emerging HEVC standard [17], offering a flexible quad-tree structure for prediction and transform. Statistical redundancies are exploited by block-wise temporal and spatial prediction. The residual signal is transformed block-wise and coefficients are quantized in the transform domain. Coding artifacts, perceived by the human observer as a loss of visual quality, are introduced by the quantization.

The encoder decides about the best representation of the video to be coded by minimizing the Lagrangian R-D cost functional $J = D + \lambda R$ on block basis with D being the distortion and R being the frame rate.

We configured the coder with a maximum prediction block-size of 64×64 and a maximum quad tree depth of 4, equivalent to a minimum prediction. We used a *IPPP* prediction structure and an intra period longer than the coded sequence, so that only the first frame is coded Intra-only.

In order to obtain sequences of different visual qualities, we changed the quantization parameter $QP \in \{32, 34, 36, 38, 40, 42, 44, 46, 48, 51\}$. The quantization parameter QP corresponds to the quantization step size $\Delta_{i,j}$ with $\Delta_{i,j} = 2^{\frac{QP}{6}} \times m_{i,j}$, where $m_{i,j}$ is a multiplier depending on the transform matrix and the position of the coefficient. However, the perceived quality loss in the generation of the stimuli is controlled by and monotone with the quantization parameter QP.

B. Experimental Design

Videos were shown on a 30" screen (Dell UltraSharp 3008WFP) with a native resolution of 2560×1600 pixels. The screen was normalized according to Rec. ITU-R BT.500-11 [3] specifications. Video resolution was 832×480 pixels or 20.8×12 cm, which corresponds to 24×14 degrees of visual angle. Viewing distance was 48 cm (4 times the video height) in compliance with Rec. ITU-R BT.500-11 [3] specifications.

Nine subjects (3 females, 6 males, age group 20–32) participated in the experiment. They were naive with respect to the purpose of the experiment and they had not participated in a video assessment study before. All had normal or corrected-to-normal vision. Subjects sat in front of the screen in a dimly-lit room.

Fig. 2. Experimental setup.

Following a visual acuity test and a general introduction to the experiment, a pretest of 100 trials with 10 different stimulus levels was conducted. In each trial, subjects first had to fixate a central fixation point for 1s. Subsequently, the video was shown for a duration of 8s. In 83% of the video sequences, a quality change occurred in the interval of 2-6s; in the other 17% of sequences, no change occurred (QC0). The stimulus levels of the experiment (QC1-QC10) are the magnitude of the quality change (cf. Section Section II-A). At the end of the video, subjects indicated via a button press whether or not they perceived a change in quality at any point during the video (yes/no-task). Based on the behavioural data from the pretest, 4 subject-specific stimulus levels were chosen that targeted the slope of the psychometric functions, i.e. stimulus levels around the threshold of perception. If QCx is the stimulus level closest to the perception threshold, then QCx-2, QCx-1, QCx, QCx+1 were the selected stimulus levels[2]. In addition, the undistorted condition without a quality change (QC0) and a maximal quality change condition (QCmax) were included in order to have clearly perceived respectively not perceived trials as a reference (Table I).

TABLE I
INDIVIDUAL STIMULUS LEVELS FOR ALL SUBJECTS. FOR THREE SUBJECTS, STIMULUS LEVEL QC-I WAS OMITTED DUE TO TIME CONSTRAINTS.

Subject	undistorted	QC-I	QC-II	QC-III	QC-IV	QCmax
S1	QC0	QC4	QC5	QC6	QC7	QC10
S2	QC0	QC4	QC5	QC6	QC7	QC10
S3	QC0	-	QC5	QC6	QC7	QC10
S4	QC0	-	QC5	QC6	QC7	QC10
S5	QC0	QC4	QC5	QC6	QC7	QC10
S6	QC0	QC3	QC4	QC5	QC6	QC10
S7	QC0	QC3	QC4	QC5	QC6	QC10
S8	QC0	-	QC5	QC6	QC7	QC10
S9	QC0	QC4	QC5	QC6	QC7	QC10

The main experiment consisted of 600 trials (100 trials per stimulus level, randomly shuffled). These trials were subdivided into 8 blocks of 75 trials each. A block lasted about 15 minutes followed by a few minutes break. Including cap preparation, the experiment lasted about 4 hours.

C. EEG data

Brain activity was recorded using a 64-channel actiCAP active electrodes setup and BrainAmp amplifiers (Brain Products, Munich, Germany). The following electrode sites were used: AF3-4; Fp1-2; Fz,1-10; FCz,1-8; Cz,1-8; CPz,1-8; CP,1-9; Pz,1-10; POz,1-6; Oz,1-2. Data was recorded at 1000 Hz with impedances kept below 20 kΩ. For offline analysis, data was downsampled to 200 Hz and lowpass filtered at 40 Hz in order to attenuate line noise. Trials in which subjects responded before the end of the video were omitted. No artifact rejection was performed.

D. Data Analysis

1) Behavioural Data: The psychometric function characterizes the performance of an observer in a detection task as a function of a physical quantity, in this case the change in video quality. The performance is given by the detection rate, i.e. the fraction of trials where subjects reported to have detected the stimulus divided by the total number

[2]Stimulus levels with Latin numbers denote the pooled quality change levels over subjects. Since the stimulus levels are selected individually for each subject, the pooled stimulus levels might have different physical characteristics but are most similar perceptually.

of trials for the stimulus level. To obtain the psychometric curve, a logistic function was fitted to the detection rates of the stimulus levels with the *psignifit* toolbox [18] using bootstrapping to determine the confidence intervals of the fit.

2) Neurophysiology: Event-related potentials (ERPs) are obtained by aligning EEG data from a number of trials according to a predefined time point and averaging over trials. The time instance of alignment was chosen to be the time instance of the quality drop. ERP waveforms reflect only neuronal activity that is phase-locked to the stimulus, because activity that is not phase-locked to the quality change averages out. Averaging also helps to increase the signal to noise ratio of the transient ERP waveform. Common parameters to characterize an ERP are the amplitude and the latency (with respect to stimulus onset) of the peaks in the waveform.

3) Linear Discriminant Analysis (LDA): Classification of EEG data in the temporal domain is normally done by comparing the ERP data of trials from different stimulus levels. For each trial, features are derived from each channel at different time points which are then used for classification. Usually, a single feature corresponds to the voltage averaged in a certain time window for a particular EEG channel. For a linear classifier, the separating hyperplane is defined by $\mathbf{w}^T\mathbf{x}+b=0$, where $\mathbf{w}$ is the projection vector, $\mathbf{x}$ is a data point on the separating hyperplane and b is the bias term (classification threshold). The classification of a data point $\mathbf{x_i}$ is then given by $y_i = sgn(\mathbf{w}^T\mathbf{x_i}+b)$.

Linear discriminant analysis is a linear classifier that aims at finding a projection that maximizes the ratio of the class mean distance (of two or more classes) and the within-class variances. The projection vector of LDA is defined as

$$\mathbf{w} = \hat{\mathbf{\Sigma}}_c^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$$

where $\hat{\boldsymbol{\mu}}_i$ is the estimated mean of the class i and $\hat{\mathbf{\Sigma}}_c = {}^1\!/_2(\hat{\mathbf{\Sigma}}_1 + \hat{\mathbf{\Sigma}}_2)$ is the estimated covariance matrix (the matrix of covariances of all EEG channels), i.e., the average of the class-wise empirical covariance matrices $\hat{\mathbf{\Sigma}}_i$.

A linear classifier trained on temporal EEG features can be regarded as a spatial filter. Thus, the linear classifier may be interpreted as a "backward model" to recover the signal of discriminative sources. The weight vector $\mathbf{w}$ of the classifier can be visualized as a scalp map (cf. Section Section II-D4).

Let $\mathbf{w} \in \mathbb{R}^C$ be the LDA projection vector and $\mathbf{D} \in \mathbb{R}^{C\times T}$ be a matrix of EEG signals with C channels and T samples per channel. Then

$$\mathbf{D}_f := \mathbf{w}^\top \mathbf{D} \quad \in \mathbb{R}^{1\times T}$$

is the result of spatial filtering: each EEG channel in $\mathbf{D}$ gets weighted with the corresponding component of $\mathbf{w}$ and summed up to yield a single virtual channel. In other words, each virtual channel is a linear combination of the original EEG channels.

For Gaussian distributions with equal covariance matrices for both classes, LDA is the optimal classifier, i.e. the risk of misclassification for samples drawn from the same distributions is minimized [19]. Since ERP signals are approximately Gaussian distributed and the covariances are dominated by the background EEG signal rather than by class-specific covariations, LDA is perfectly suitable for classifying ERP signals.

However, the stated optimality of LDA relates to known parameters μ_1, μ_2, $\mathbf{\Sigma}_c$. For real applications, the dimensionality of the feature space is often high while the number of observations is comparatively low. For EEG

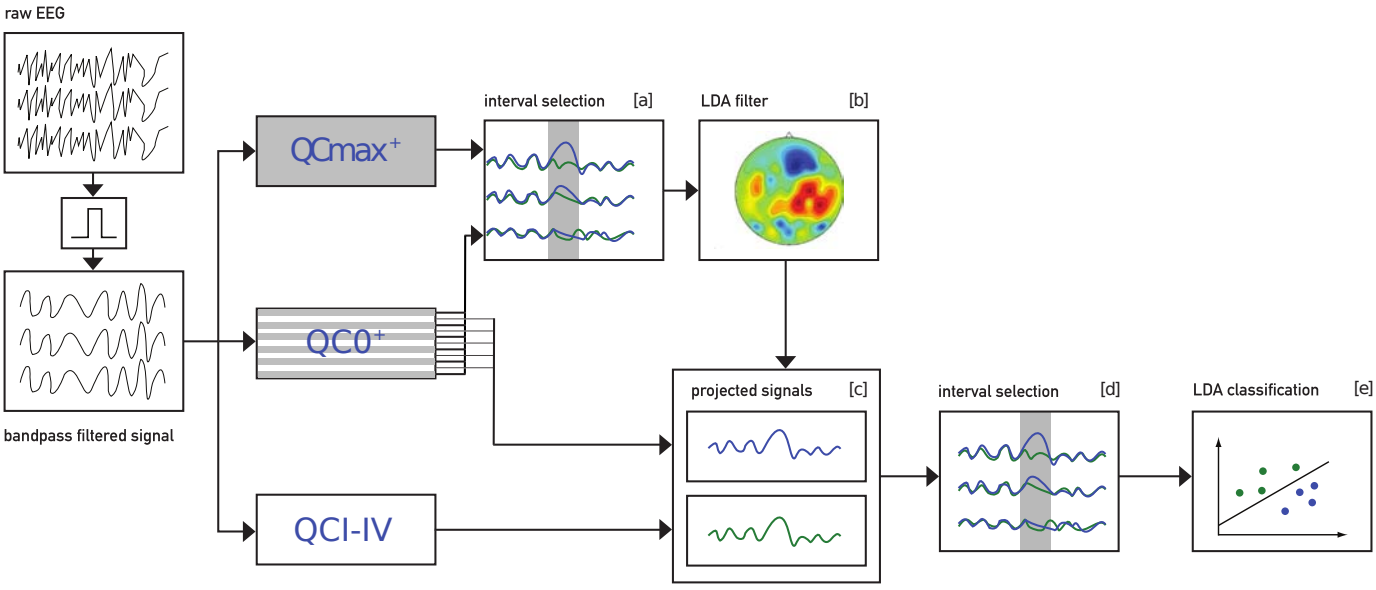

Fig. 3. EEG single-trial classification scheme. First, the raw EEG signal is bandpass filtered. Then, QC0$^+$ trials (where the superscript '+' indicates the subset of trials that were correctly labeled by the subject, i.e. correct rejections) are split into two equisized sets in an even-odd manner, as indicated by the grey-and-white horizontal bars. One of these two sets and the QCmax$^+$ trials (again, '+' indicates correctly labeled trials, in this case hits) are used to train the LDA filter. To this end, the sgn-r^2 between QCmax$^+$ and QC0$^+$ is used to select a discriminative interval **[a]** on basis of which an LDA filter is determined **[b]**. Then the sets of quality change trials QC-I, QC-II, QC-III, and QC-IV (depicted as QCI-IV) and the second set of QC0$^+$ trials are projected to virtual channels using the LDA filter **[c]**. For each QC level and for hits and misses separately, two intervals that discriminate best between quality changes and undistorted trials are selected **[d]**. This yields two-dimensional features which are finally used to classify single trials using LDA **[e]**.

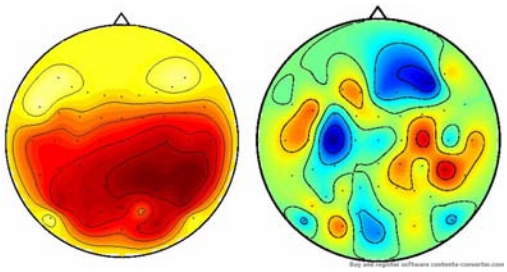

Fig. 4. Scalp plots depicting spatial distribution as a top view on the head, with nose pointing upwards and crosses marking the electrode positions. **Left:** sgn-r^2 values for QCmax$^+$ against QC0$^+$ for subject S1. The spatial distribution is similar to the P3 component, suggesting that class differences are mostly due to the P3. **Right:** LDA filter trained on the two classes. If the channels were uncorrelated, the LDA filter would be proportional to the difference of the class means. If the noise is not substantial (as it is the case for our narrowband filtered data), the spatial distribution roughly consists of dipoles along the gradient of the sgn-r^2 scalp plot.

data, the ratio between the number of observations and the number of channels is low which leads to a systematic misestimation of the covariance matrix [20]. Due to that, the estimated distribution parameters, in particular $\hat{\mathbf{\Sigma}}_c$, are error prone which render classification suboptimal. Accordingly, classification was done by LDA with automatic regularization of the estimated covariance matrix using shrinkage [20] which outperforms the standard LDA when classifying ERPs on a single-trial level. For detailed overviews on single-trial classification of EEG data see [21], [22].

4) Classification: Classification is done subject-wise in a 2-step procedure (cf. Figure 3). First, discriminative time intervals between correctly reported QC0 trials (correct rejections, denoted as QC0$^+$) and correctly detected trials with highest quality change (hits, denoted as QCmax$^+$) are computed (Figure 3a).

To measure class discrimination for each channel and time point (over trials), we use the signed squared biserial correlation coefficient sgn-$r^2 := \mathrm{sgn}(r) \cdot r^2$. The biserial correlation coefficient is defined as

$$r := \frac{\sqrt{N_1 \cdot N_2}}{N_1 + N_2} \frac{\mu_1 - \mu_2}{\sigma},$$

where N_1 and N_2 are the number of samples in the two classes, μ_i is the mean over samples (EEG voltage levels in this case) of class i (stimulus levels) and σ is the standard deviation over all samples. Roughly, the r^2 measures how much of the total variance for all samples can be explained by class membership.

Since classification is done on the P3 which is a single positive component, it is sufficient and neurophysiologically reasonable to choose only one time interval[3]. For the channelwise mean of this interval, the LDA-projection vector is computed (Figure 3b). Figure 4 shows the projection vector obtained for one subject.

The data for the different stimulus levels is projected on the LDA filter (Figure 3c; cf. Section II-D3). The LDA filter is computed on QCmax$^+$ vs. QC0$^+$ since we expect the P3 component to be most prominent for this stimulus level (this is verified in Figure 7, top left). By projecting EEG data of the other stimulus levels onto the LDA filter, we ensure that the classification is done on the P3 component instead of artifacts such as eye blinks that might also be discriminative to some degree. Furthermore, as illustrated in Figure 5, LDA-prefiltering reduces the variability across trials and thus enhances class separation on a single trial level. Also, the dimensionality (i.e. the number of channels) is thereby largely reduced. This is particularly important as LDA relies on an estimation of the covariance matrix, which is systematically skewed if the ratio between the number of features and the number of observations is high.

In the second step, the LDA-prefiltered data is classified stimulus-levelwise (QC I-IV) against EEG data of correctly rejected undistorted trials. Again, the sgn-r^2 heuristic is used to extract the most discriminative time intervals [21] (Figure 3d). Simulations have shown that two intervals suffice. By searching for the most discriminative subject-specific time

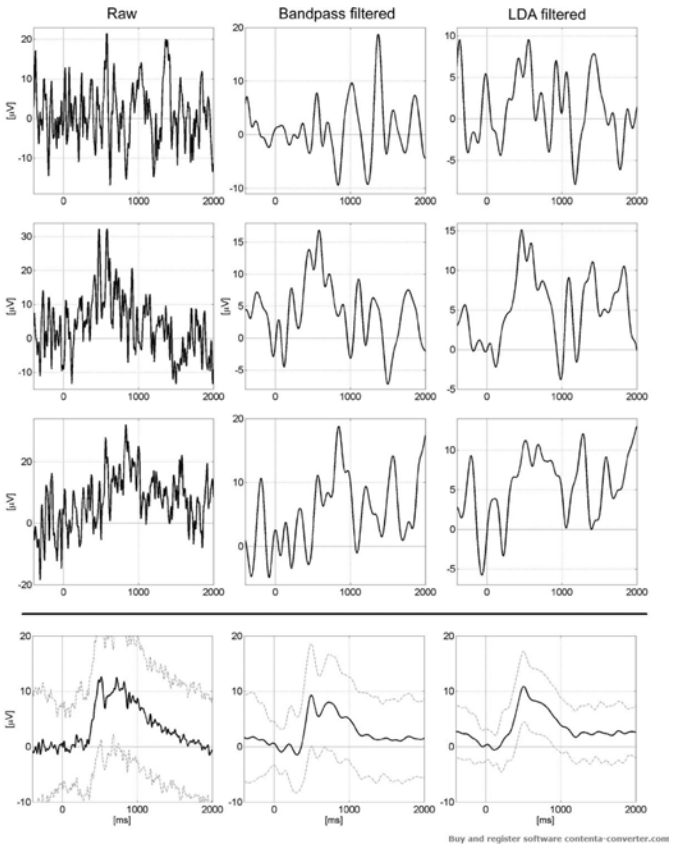

Fig. 5. EEG data from hit trials of stimulus level QC-IV of subject S1. **First column:** raw EEG data at channel CPz; **second column:** bandpass-filtered EEG data at channel CPz; **third column:** LDA-projected EEG data. **Rows 1–3:** Single-trial data from three hit trials of subject S1. **Last row:** ERP and standard deviation over all hit trials of QC-IV for subject S1. The ERP peak is similar for all three data types, but the standard deviation is substantially reduced by each processing step. Further, the processing clearly increases the prominence of the P3 component in single trials. This illustrates that the filtering steps are beneficial for the enhancement the P3 component and, therefore, classification performance.

[3]The VEP is not used for classification since in our study, its amplitude was considerably smaller than the P3 amplitude and was of negligible class discriminability.

intervals in the LDA-projected data, the classification becomes invariant with respect to the exact position of the P3 component relative to stimulus onset. The trialwise means within the selected time intervals are the features (2-dimensional) for the classification with LDA (Figure 3e). Thus, this analysis is based on the assumption that a P3 component for lower quality changes has a similar spatial distribution to the P3 component of QCmax although amplitude and latency might differ. Note that two disjunct sets of undistorted trials from QC0, created by alternating trials chronologically, are used for step 1 and 2.

For each subject, hits (true positives) and misses (false negatives) of each stimulus level are cross-validated separately against correctly reported undistorted trials ($QC0^+$) in a leave-one-out fashion. Classification is only performed on stimulus levels with sufficient numbers of hit or miss trials ($n \geq 20$). Classifying hits against $QC0^+$ serves as a proof of concept. It shows that the features derived from EEG indeed reflect the subject's perception of the quality change and that EEG-based classification on a single-trial basis is possible. By classifying misses against $QC0^+$ trials, we investigate whether the EEG signals of these two classes differ although their behavioural response is the same. A classification performance above chance level would substantiate that EEG offers a sensitivity that goes beyond what can be inferred from the behavioural data.

Classification performance is measured by the AUC, the area under the curve of the receiver operating characteristic (ROC) [23]. The ROC curve displays the true positive rate and the false positive rate of a 2-class problem as the discrimination criterion is varied. It is frequently used in machine learning as a measure of classification separability that is invariant to the number of observations per class. The AUC value equals the probability that a randomly chosen instance of class 1 has a higher value than one of class 2. An AUC of 1 (or 0) thus reflects perfect class separation, an AUC of 0.5 classification by chance.

For classification, data is bandpass filtered with a butterworth filter between 0.2 and 7 Hz to attenuate the influence of slow wave and alpha activity.

III. Results

A. Behavioural Data

Figure 6 depicts the psychometric functions fitted to the detection rate of the subjects at different stimulus levels.

The stimulus levels selected in the pretest were identical for the majority of subjects (see Table I). In addition, the corresponding psychometric functions are roughly consistent across subjects.

B. Event-Related Potentials

The ERP waveform is dominated by the P3 component that is evident from 400–600 ms following the quality change (Figure 7). Its peak is broadly distributed over the scalp with a center at central-parietal electrode sites (channel CPz). For QCmax, the P3 was present for all subjects. We found that P3 amplitude increases with the magnitude of quality change,

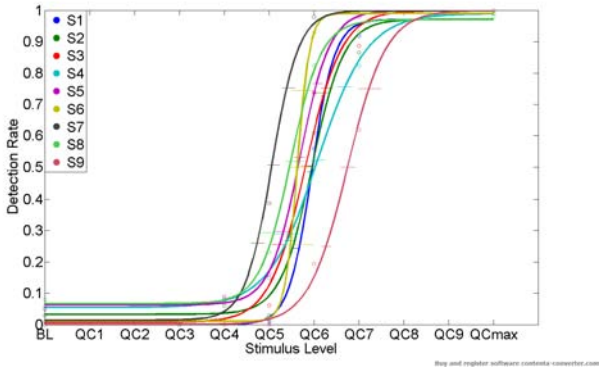

Fig. 6. Psychometric function fits from the psychophysical data (small circles); each function represents one subject. Horizontal lines depict the 95% confidence intervals of the fit.

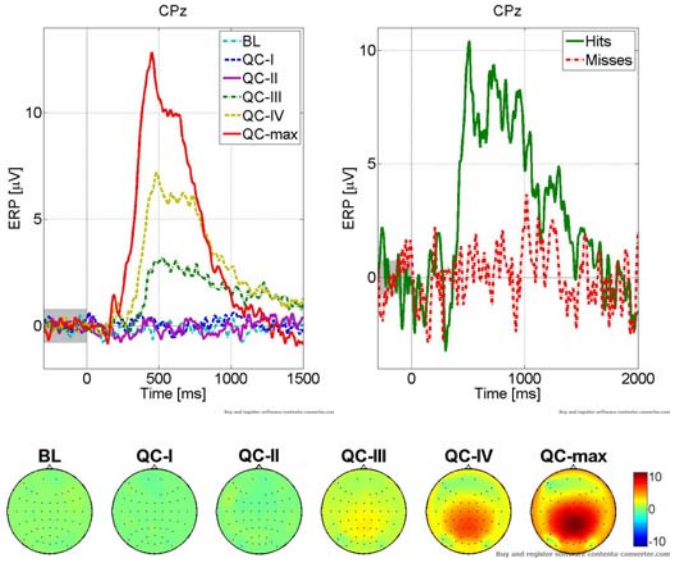

Fig. 7. Grand average ERP plots for the different stimulus levels. **Top left:** ERP for undistorted trials and the different quality changes at channel CPz. **Top right:** ERP for a selected stimulus level (QC-III) for subject S1, subdivided in hits (wherein the quality change was perceived) and misses (wherein the quality change was not perceived). **Bottom:** Scalp topographies for all channels. Each circle depicts a top view of the head, with the noise pointing upwards. Colors code the mean voltage for the time interval from 400-700 ms after quality change. ERP plots for single subjects can be found in the supplementary material.

which implies that it is a good index of stimulus intensity. This is supported by the fact that there is a strong correlation (r=0.89 on average) between the detection rate of the subjects for the different stimulus levels and the corresponding P3 amplitude (Figure 8). In other words, the average neuronal response directly reflects task difficulty: If the quality change is easy to detect, P3 amplitude is high [24], [25]. For quality changes near the threshold of perception, amplitude is low. For stimulus levels below the perceptual threshold (QC-I/QC-II), the grand average shows no P3 component.

Separate ERPs over reported and unreported quality changes differ considerably for all subjects, even within the same stimulus class (Figure 7, top right), which further indicates that the P3 reflects neuronal processing of the quality change.

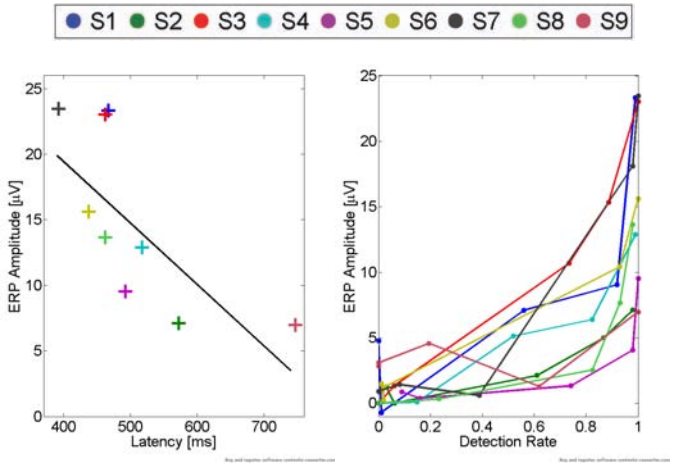

Fig. 8. Relationship between neurophysiological and behavioral measures. **Left:** Amplitude and latency of the P3 component for QCmax shows a significant linear correlation across subjects (r=-0.72, $p<0.05$, left plot). **Right:** Within subjects and across stimulus levels, amplitude and detection rate are positively correlated (r=0.84±0.15; right plot).

C. Classification

Classification results for all subjects are depicted in Figure 9. Single-trial classification performance of hits against undistorted trials is highly significant ($p<0.01$) for all subjects, with AUC values close to 1 for QCmax in most cases. For lower stimulus levels, the AUC drops slightly to 0.8–0.9 for most subjects. An increase in classification accuracy with higher stimulus levels was observable for all subjects. Since ERP amplitude increases with the stimulus level, classification performance is tightly coupled to ERP amplitude. Across subjects, a comparably small ERP amplitude for QCmax also tends to lead to lower performance (cf. Figure 9; subjects S4, S5, S9) and a high amplitude for QCmax leads to a high performance irrespective of the stimulus level (subjects S1, S3, S7).

Classifying misses against undistorted trials yielded significant results for three subjects, with AUC values around 0.65. For all other subjects, classification did not exceed chance level. In line with the hits vs. undistorted classification, there was a clear tendency for classification performance to increase with the stimulus level (e.g. all three statistically significant classifications were measured for the highest stimulus class with a sufficient number of misses).

However, classification performance alone is not a sufficient measure when the objective is to detect conservative behaviour, that is, trials that were reported as "not perceived" although the subject faintly did. Since the fraction of these trials within all miss trials within a stimulus level may be small and may differ across subjects, classification performance does not reflect how well these trials have been detected.

An alternative approach can be motivated as follows. If a quality change was perceived residually but not reported, this should be reflected by a (possibly low-amplitude) P3 component in the EEG. If not, the ERP waveform should be indistinguishable from the undistorted trials. Thus, if the subjects answered conservatively, a P3 component should be detectable by averaging over trials which were not labeled as undistorted trials by the classifier ("classifier-hits"). Generally, the P3 itself is characterized by its spatial and its temporal form. However, for LDA-prefiltered classifier-hits, the spatial distribution is roughly fixed by the LDA filter, i.e. the classifier will only label those trials as positive that have a P3-like spatial distribution in the classification time interval. However, as a P3-like spatial distribution might also arise due to random fluctuations in the EEG, the temporal time course is important to tell apart a random distortion from a true P3.

For Figure 10, three different classification tasks were performed corresponding to the three rows. In each column, the results of the same three example subjects are shown. The curves correspond to the ERP averages of the EEG data after it has been projected onto a single channel using the LDA-filter. For the classification of hits against undistorted trials (top row), a clear P3-like component is present for all three example subjects. This component is visible for the QC-III$^+$ trials, i.e. the distorted QC-III trials that have been detected by the classifier (green line), but it is also visible when one averages over all QC-III trials, including those that have not been detected by the classifier (blue dotted line). For QC-III$^-$ trials (undetected distorted trials) and undistorted trials, no P3 component is visible. For misses against undistorted trials (middle row), a P3 component with a similar latency but a smaller amplitude can be seen in at least two subjects (S2, S3). Note that this positive peak is not only visible for trials classified as hits but also for the average over all miss trials. For lower stimulus levels in which the classification was not significant (bottom row), there are no apparent differences in the average waveform between miss trials and undistorted trials. For these cases, some trials (classifier-hits) still seem to have a P3-like shape, but as classifier-misses have a negative peak during the classification interval, this effect can be attributed to random fluctuations within the EEG signal.

IV. Discussion

A. General Pattern of Activation

A P3 component related to the quality change was detected in all subjects. This component shows a graded response, i.e. its amplitude scales with the magnitude of the quality distortions for all subjects (mean correlation: r=0.89). The P3 has long been known to vary with stimulus probability and with stimulus intensity in both the auditory and visual modality [26]. In this experiment, the probability of a large change as in QCmax is low (17%) as all other quality changes are clearly more subtle, which is reflected neuronally by a very large P3 amplitude for QCmax compared to the other stimulus levels.

Across subjects, there is also a large variability in P3 amplitude (7-23 μV). However, amplitude differences of the P3 between subjects have been recognized for decades and are understood to depend on a variety of psychological and biological factors [27] rather than on a different processing of the stimuli.

For stimulus levels below the perceptual threshold (QC-I/QC-II) for which detection rates are below 0.15, we did not find a difference between the ERPs.

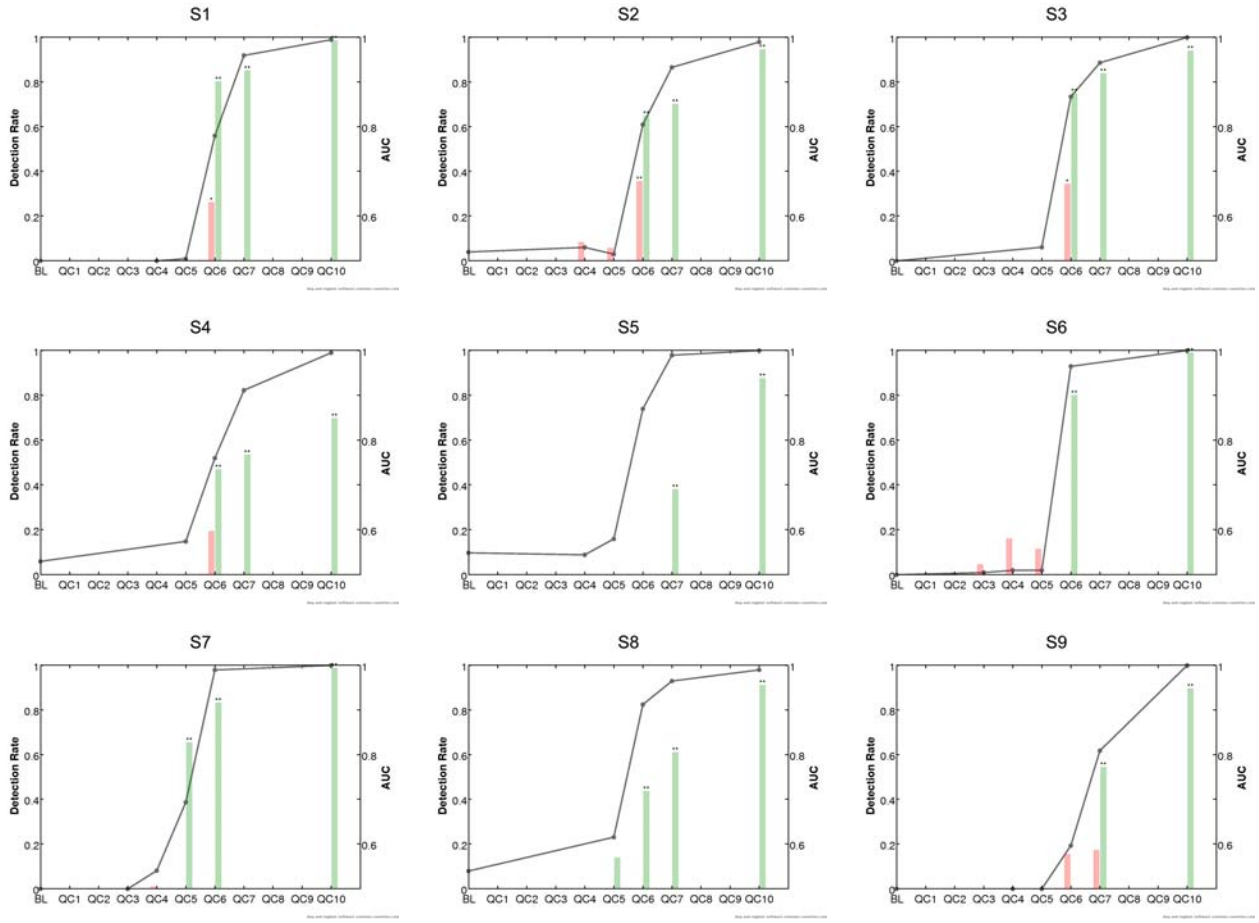

Fig. 9. Classification results for all subjects (S1-S8). Green bars show the classification performance (AUC value) of hits against QC0$^+$; red bars depict misses against QC0$^+$. One or two asterisks denote the significance level of the classification outcome in a Wilcoxon rank-sum test ($p<0.05$ or $p<0.01$, respectively). The grey curve depicts the detection rate over stimulus levels. Note that the detection rate and the classification performance have no direct connection as the classification is done separately on hit and miss trials.

B. Classification

Psychophysical experiments suffer from lapses of subjects, in which clearly perceivable stimuli are not reported (e.g. due to inattentiveness or a wrong button press) or stimuli far below the perceptual threshold are reported as perceived. For fitting a psychometric function to the behavioural data, methods have been developed to deal with these lapses [28]. However, in a yes/no-task, methods purely based on the behavioural data cannot detect conservative response behaviour, i.e. trials in which subjects reported to have seen no change in quality although they faintly did. We showed that the EEG classification has the potential to detect these "mislabeled" trials. For those cases where the classification is significant, the ERP of miss trials classified as hits shows a P3-like temporal (Figure 10) and spatial topography (positive peak centered over central-parietal areas, unpublished data), indicating that the classification is also reasonable neurophysiologically. That is, we can assume that the subject either consciously perceived the change, or that the brain processed it to some degree without conscious awareness.

In some cases, single QC0$^+$ trials were classified as hits. This is possibly due to random fluctuations in the EEG that have a similar spatial distribution as the P3 component. Although these events are rare, these fluctuations impede an error free classification and further investigations have to be made on how to reduce their influence. Overall, for subjects with a high P3 amplitude, the classification tends to be better and more stable over stimulus levels. The reason is that on single trial level, low-amplitude P3 components are hard to detect since the background EEG noise is high, especially in the low frequency range from which the P3 originates [4]. Subjects with a high P3 amplitude are thus less prone to random fluctuations in the EEG and therefore more suited for our approach than subjects with a low P3 amplitude.

Subjects did not have to press the button instantly when they perceived a quality change since this would lead to neuronal motor activity that might interfere with classification. As a result, we cannot determine whether subjects perceived the quality distortion immediately or with a time lag. Especially for stimuli around the perceptual threshold, subjects might not notice the change exactly at its onset. In the ERP waveform, a lower amplitude and slower decay can be observed for more subtle quality changes (Figure 7), which seems at least partially to be due to a temporal jitter of the P3 component.

[4]The power spectrum of EEG background noise has approximately a $\frac{1}{f}$ distribution.

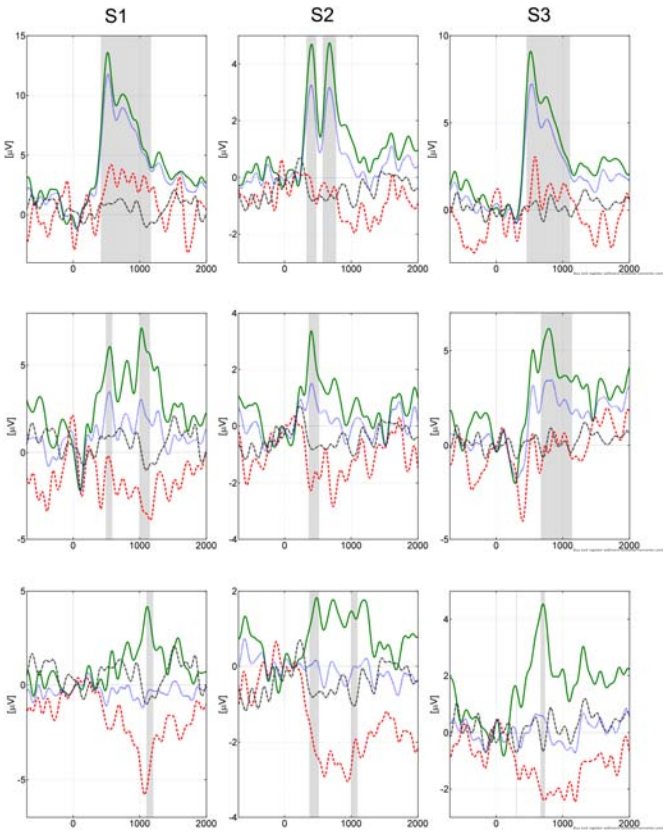

Fig. 10. ERPs of LDA-prefiltered data for subjects S1, S2, and S3. The blue dotted line shows the ERP over all trials of the stimulus level, the black dash-dotted line the ERP of the QC0$^+$ undistorted trials. The green and red line gives the ERP over the trials of the stimulus level, subdivided into trials that are classified as hits (green) and misses (red) by the classifier. The grey shaded areas depict the time intervals from which classification features are computed. Rows are the results for different classification runs: **Top row:** Hits vs. QC0$^+$ for QC-III. **Middle row:** Misses vs. QC0$^+$ for QC-III. These are the three classifications on miss trials that are statistically significant. **Bottom row:** Misses vs. QC0$^+$ for the lowest QC-level of the subject. Note the different scaling of the plots.

Although LDA-prefiltering is spatial and therefore time invariant, the discriminative intervals for classification are selected using all trials in the training set. If the P3 components of these all have a similar latency, this will lead to a peak at this latency for the sgn-r^2 discrimination function. If the latency is variable, this discriminative peak will be smoothed out and an interval will be chosen that may not be optimal for all trials.

EEG data is highly noisy, non-stationary and easily affected by psychological factors such as attentiveness, motivation, or fatigue. As a result, EEG varies considerably over the time course of an experiment, and methods that make the signals more robust against such fluctuations are important for a good single-trial classification performance. For instance, adaptive LDA classifiers [29] and stationary subspace analysis (SSA) [30] are two methods that enhance single-trial classification in brain-computer interfaces [31], [32]. Future work has to examine to what extent these and other methods prove useful in video quality experiments.

C. Towards neurotechnology in video quality assessment

A full assessment of the human perception of video quality is beyond the reach of conventional experimental methods. The purpose of this work is to pave the way for a neurotechnology-based approach to video quality assessment by giving a proof-of-concept. While we do not present a full-fleshed solution for objective, robust, and reliable assessment of video quality perception, we believe that our method shows that neurotechnology can be a useful complement to, and an extension of, established behavioral methods. To substantiate this, it seems feasible to enumerate its potential merits in quality assessment.

- Overt response. The direct monitoring of brain activity releases one from the necessity to record overt responses such as button presses. Overt responses not only interrupt the experimental flow, they can interfere with the subject's evaluation of a stimulus.
- Real time monitoring. Brain activity can be monitored and processed in real-time, potentially giving an impression of the user's assessment of video material while it is being viewed.
- Objectivity. Behavioral methods often suffer from response bias, that is, some subjects are more inclined than others to give a particular response or rating. Tapping the brain response directly promises a more objective account on the perception and assessment of a stimulus than behavioral methods.
- Sensitivity. EEG is sensitive to stimuli that are at or even below the threshold of conscious perception. In the present study, we have some indications for that from subject S6 (see Figure 9), who showed a brain response to low-distortion stimuli even though he did not report perceiving them. More robust evidence stems from a recent study on visual flicker, which using machine learning showed that the brain can respond to flicker even when it is not reported by the subject [33].

D. Caveats

Before an out-of-the-box solution to video quality assessment is in reach, several more hurdles have to be overcome. Our study is limited in the following respects.

First, the coding artifacts arising in our stimulus video do not cover the full range of possible artifacts and those artifacts appear in a sudden change of quality. Thus, the quality characteristics of the used stimuli are similar to the quality deviation in the special case of packet loss in fidelity scalable video coding [34]. For a full assessment, several experiments have to be conducted to examine the different types of codec artifacts and typical artifact combinations.

Second, the EEG approach yielded significant results only with a subset of subjects. In particular, our approach at present requires a sufficiently large P3 amplitude of the subject. At present, two routes are possibly to remedy this limitation. One route is to increase the sensitivity and robustness of EEG using advanced signal processing methods such as those mentioned above. Another route is to have a method to quickly identify the subjects that are suitable for an EEG-based approach. Such screening methods have already been explored in the context

of brain-computer interfaces [35], [36]. In the present context, one might envisage a short measurement using a classical oddball paradigm based on which the P3 amplitude can be estimated. P3 amplitude could serve as a good predictor of classification performance due to its strong correlation.

Third, it seems that the extra preparation time required by an EEG setup (>1/2 hour) may counteract the benefits of EEG measurements. However, a new generation of EEG caps using dry electrodes [37], [38] eliminates the time-expensive preparation of the EEG cap. This could substantially increase the practical applicability of our approach in quality assessment experiments.

Fourth, the present proof-of-concept study used artificially generated movie clips featuring a homogeneous 10 sec video stream. Thus, one might wonder how this approach would work for real videos, in particular given the regular occurrence of scene-changes, which could be perceived as unexpected changes and might thus trigger themselves a P3 component. In this respect, it is important to note that scene-changes may be considered not an interfering nuisance, rather they could serve as a kind of calibration normal for each subject under study: Scene-changes (which could be automatically detected from the video stream) define instances where the individual spatio-temporal profile of the P3 voltage maps can be obtained against which the P3 triggered by artificially inserted quality-changes can be compared. Thereby, one could obtain a natural metric for characterising the extent of quality-change related EEG data across subjects.

V. Conclusion

We proposed a novel approach based on neural correlates from EEG. This approach holds the promise to provide a less-biased, and therefore more objective, account of quality perception than obtained with behavioral methods.

Clearly, the method presented here needs further improvement, but we conjecture that it can be a valuable complement in psychophysical experiments wherein the number of trials is limited by relabeling mislabeled trials. Furthermore, since its focus is the perception of quality *change*, not quality perception per se, it forms only the first step of a neurotechnology-based approach to video quality assessment. Future work will aim towards a direct measure of quality perception using tonic EEG features such as oscillatory components (e.g., alpha rhythm). A direct neural index would allow for the real-time assessment of perceived image quality during the observations of videos and it would relinquish the need for an overt response by the subject.

VI. Acknowledgements

We would like to thank Chad Fogg for helpful comments on the manuscript. KRM and TW acknowledge the Falling Walls Conference 2009 for triggering the process that ultimately lead to this work.

References

[1] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.

[2] A. B. Watson and J. Malo, "Video quality measures based on the standard spatial observer," in *ICIP Proceedings, III*, Rochester, NY, 2002.

[3] Rec. ITU-R BT.500-11, "Methodology for the Subjective Assessment of the Quality of Television Pictures," 2002.

[4] Rec. ITU-T P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," 2008.

[5] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[6] P. Nunez, R. Srinivasan, A. Westdorp, R. Wijesinghe, D. Tucker, R. Silberstein, and P. Cadusch, "EEG coherency:: I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalography and clinical Neurophysiology*, vol. 103, no. 5, pp. 499–515, 1997.

[7] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr Clin Neurophysiol*, vol. 70, pp. 510–523, 1988.

[8] M. S. Treder, N. M. Schmidt, and B. Blankertz, "Gaze-independent brain-computer interfaces based on covert attention and feature attention," *J Neural Eng*, vol. 8, no. 6, p. 066003, 2011, open Access. [Online]. Available: http://dx.doi.org/10.1088/1741-2560/8/6/066003

[9] M. Schreuder, B. Blankertz, and M. Tangermann, "A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue," *PLoS One*, vol. 5, no. 4, p. e9813, 2010.

[10] J. Höhne, M. Schreuder, B. Blankertz, and M. Tangermann, "Two-dimensional auditory P300 speller with predictive text system," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 4185–4188.

[11] A. Porbadnigk, J. Antons, B. Blankertz, M. S. Treder, R. Schleicher, S. Möller, and G. Curio, "Using ERPs for assessing the (sub) conscious perception of noise," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 2690–2693.

[12] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects," *Neuroimage*, vol. 37, no. 2, pp. 539–550, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2007.01.051

[13] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, Eds., *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[14] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process Mag*, vol. 25, no. 1, pp. 41–56, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1109/MSP.2008.4408441

[15] C. Gerald and P. Wheatley, *Applied numerical analysis*, 1989.

[16] D. Marpe, H. Schwarz, and al., "Video Compression Using Nested Quadtree Structures, Leaf Merging, and Improved Techniques for Motion Representation and Entropy Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 12, pp. 1676–1687, Dec. 2010.

[17] JCTVC, T. Wiegand, W.-J. Han, B. Bross, J.-R. Ohm, and G.J. Sullivan, "WD3: Working Draft 3 of High-Efficiency Video Coding," Mar. 2011, download via anonymous ftp to: standard.pictel.com/video-site/9712_Eib/q15c11.doc.

[18] I. Fründ, N. Haenel, and F. Wichmann, "Inference for psychometric functions in the presence of nonstationary behavior," *Journal of Vision*, vol. 11, no. 6, 2011.

[19] R. Duda, P. Hart, and D. Stork, *Pattern classification*. Wiley New York, 2001, vol. 2.

[20] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[21] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components–a tutorial," *Neuroimage*, vol. 56, no. 2, pp. 814–825, 2011.

[22] K.-R. Müller, C. Anderson, and G. Birch, "Linear and nonlinear methods for brain-computer interfaces," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 11, no. 2, pp. 165–169, 2003.

[23] D. Green and J. Swets, *Signal detection theory and psychophysics*. Robert E. Krieger, 1974.

[24] G. Hagen, J. Gatherwright, B. Lopez, and J. Polich, "P3a from visual stimuli: Task difficulty effects," *International journal of psychophysiology*, vol. 59, no. 1, pp. 8–14, 2006.

[25] K. Kim, J. Kim, J. Yoon, and K. Jung, "Influence of task difficulty on the features of event-related potential during visual oddball task," *Neuroscience letters*, vol. 445, no. 2, pp. 179–183, 2008.

[26] J. Polich, P. Ellerson, and J. Cohen, "P300, stimulus intensity, modality, and probability," *International Journal of Psychophysiology*, vol. 23, no. 1-2, pp. 55–62, 1996.

[27] J. Polich and A. Kok, "Cognitive and biological determinants of P300: an integrative review," *Biological psychology*, vol. 41, no. 2, pp. 103–146, 1995.

[28] F. Wichmann and N. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, p. 1293, 2001.

[29] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz, "Machine-learning-based coadaptive calibration for brain-computer interfaces," *Neural Computation*, pp. 1–28, 2010.

[30] P. von Bünau, F. Meinecke, F. Király, and K.-R. Müller, "Finding stationary subspaces in multivariate time series," *Physical review letters*, vol. 103, no. 21, p. 214101, 2009.

[31] P. von Bünau, F. Meinecke, S. Scholler, and K.-R. Müller, "Finding stationary brain sources in EEG data," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 2810–2813.

[32] C. Vidaurre and B. Blankertz, "Towards a cure for BCI illiteracy," *Brain topography*, vol. 23, no. 2, pp. 194–198, 2010.

[33] A. K. Porbadnigk, S. Scholler, B. Blankertz, A. Ritz, M. Born, R. Scholl, K.-R. Müller, G. Curio, and M. S. Treder, "Revealing the neural response to imperceptible peripheral flicker with machine learning," in *Conf Proc IEEE Eng Med Biol Soc*, vol. 2011, 2011, pp. 3692–3695.

[34] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[35] B. Blankertz, C. Sannelli, S. Halder, E. Hammer, A. Kübler, K.-R. Müller, G. Curio, and T. Dickhaus, "Neurophysiological predictor of SMR-based BCI performance," *NeuroImage*, vol. 51, no. 4, pp. 1303–1309, 2010.

[36] M. S. Treder, A. Bahramisharif, N. M. Schmidt, M. van Gerven, and B. Blankertz, "Brain-computer interfacing using modulations of alpha activity induced by covert shifts of attention," *J Neuroeng Rehabil*, vol. 8, p. 24, 2011. [Online]. Available: http://www.jneuroengrehab.com/content/8/1/24/abstract

[37] F. Popescu, S. Fazli, Y. Badower, B. Blankertz, and K.-R. Müller, "Single trial classification of motor imagination using 6 dry EEG electrodes," *PloS one*, vol. 2, no. 7, p. e637, 2007.

[38] C. Grozea, C. Voinescu, and S. Fazli, "Bristle-sensors – low-cost flexible passive dry EEG electrodes for neurofeedback and BCI applications," *Journal of Neural Engineering*, vol. 8, no. 2, p. 5008, 2011.