Crowdsourced Estimation of Collective Just Noticeable Difference for Compressed Video with Flicker Test and QUEST+

Mohsen Jenadeleh [1], Raouf Hamzaoui [2], Ulf-Dietrich Reips [2], and Dietmar Saupe [2]

[1]University of Konstanz
[2]Affiliation not available

October 31, 2023

Abstract

The concept of video-wise just noticeable difference (JND) was recently proposed to determine the lowest bitrate at which a source video can be compressed without perceptible quality loss with a given probability.

This bitrate is usually obtained from an estimate of the satisfied used ratio (SUR) at each bitrate, respectively encoding quality parameter. The SUR is the probability that the distortion corresponding to this bitrate is not noticeable. Commonly, the SUR is computed experimentally by estimating the subjective JND threshold of each subject using binary search, fitting a distribution model to the collected data, and creating the complementary cumulative distribution function of the distribution. The subjective tests consist of paired comparisons between the source video and compressed versions. However, we show that this approach typically over- or underestimates the SUR.

To address this shortcoming, we directly estimate the SUR function by considering the entire population as a collective observer. Our method randomly chooses the subject for each paired comparison and uses a state-of-the-art Bayesian adaptive psychometric method (QUEST+) to select the compressed video in the paired comparison.

Our simulations show that this collective method yields more accurate SUR results with fewer comparisons.

We also provide a subjective experiment to assess the JND and SUR for compressed video. In the paired comparisons, we apply a flicker test that compares a video that interleaves the source video and its compressed version with the source video. Analysis of the subjective data revealed that the flicker test provides on average higher sensitivity and precision in the assessment of the JND threshold than the usual test that compares compressed versions with the source video.

Using crowdsourcing and the proposed approach, we build a JND dataset for 45 source video sequences that are encoded with both advanced video coding (AVC) and versatile video coding (VVC) at all available quantization parameters. Our dataset is available at http://database.mmsp-kn.de/flickervidset-database.html.

Crowdsourced Estimation of Collective Just Noticeable Difference for Compressed Video with Flicker Test and QUEST+

Mohsen Jenadeleh, *Member, IEEE* , Raouf Hamzaoui, *Senior Member, IEEE* , Ulf-Dietrich Reips, and Dietmar Saupe

***Abstract*—The concept of video-wise just noticeable difference (JND) was recently proposed to determine the lowest bitrate at which a source video can be compressed without perceptible quality loss with a given probability. This bitrate is usually obtained from an estimate of the satisfied used ratio (SUR) at each bitrate, respectively encoding quality parameter. The SUR is the probability that the distortion corresponding to this bitrate is not noticeable. Commonly, the SUR is computed experimentally by estimating the subjective JND threshold of each subject using binary search, fitting a distribution model to the collected data, and creating the complementary cumulative distribution function of the distribution. The subjective tests consist of paired comparisons between the source video and compressed versions. However, we show that this approach typically over- or underestimates the SUR. To address this shortcoming, we directly estimate the SUR function by considering the entire population as a collective observer. Our method randomly chooses the subject for each paired comparison and uses a state-of-the-art Bayesian adaptive psychometric method (QUEST+) to select the compressed video in the paired comparison. Our simulations show that this collective method yields more accurate SUR results with fewer comparisons. We also provide a subjective experiment to assess the JND and SUR for compressed video. In the paired comparisons, we apply a flicker test that compares a video that interleaves the source video and its compressed version with the source video. Analysis of the subjective data revealed that the flicker test provides on average higher sensitivity and precision in the assessment of the JND threshold than the usual test that compares compressed versions with the source video. Using crowdsourcing and the proposed approach, we build a JND dataset for 45 source video sequences that are encoded with both advanced video coding (AVC) and versatile video coding (VVC) at all available quantization parameters. Our dataset is available at http://database.mmsp-kn.de/flickervidset-database.html.**

***Index Terms*—Just noticeable difference, psychometric function, satisfied user ratio, flicker test, subjective quality assessment, Bayesian adaptive psychometric testing method, AVC, VVC.**

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 496858717 and Project ID 251654672 – TRR 161 (Project A05). This research was also kindly supported by the Zukunftskolleg, the University of Konstanz, with funding from the Excellence Strategy of the German Federal and State Governments.

Mohsen Jenadeleh (corresponding author) and Dietmar Saupe are with the Department of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany (e-mail: mohsen.jenadeleh@uni-konstanz.de; dietmar.saupe@uni-konstanz.de).

Raouf Hamzaoui is with the School of Engineering and Sustainable Development, De Montfort University, LE1 9BH Leicester, UK (e-mail: rhamzaoui@dmu.ac.uk).

Ulf-Dietrich Reips is with the Department of Psychology, University of Konstanz, 78464 Konstanz, Germany (e-mail: reips@uni-konstanz.de).

I. Introduction

COMPRESSION is the main tool to achieve a target video bitrate required to meet transmission bandwidth and storage constraints. However, as the bitrate decreases, more compression artifacts are introduced, eventually becoming noticeable and even annoying to human consumers. Therefore, methods for assessing video quality are being explored to determine the lowest bitrate at which perceived visual quality is at a level the video content provider deems appropriate for delivery.

In the context of high-quality media streaming, the following challenges arise in automatic video quality assessment [1]: (1) Quality assurance for the re-encodings of media submitted by the original producers, (2) quality monitoring of the delivered video sequences to characterize the general satisfaction requirements of subscribers, (3) optimizing encoding parameters such that the bitrate is minimized for each targeted visual quality level, (4) optimizing streaming bitrate selection based on the speed of the consumer's network and the perceptual qualities of the upcoming video segments within some time horizon, and (5) codec and processing technology evaluation for the purpose of helping to select and update the methods for deployment that yield the best perceptual qualities.

In streaming applications, consumers pay for services and expect to receive content that does not exhibit any annoying impairments. Thus, in this context only the top quality levels of streaming around the near-lossless range are relevant. To provide a fine granularity for such high quality stimuli, a new evaluation approach was introduced, based on the concept of just noticeable difference (JND). The JND goes back to the 19th-century psychologist Ernst Weber, who defined it as the "minimum amount by which stimulus intensity must be changed in order to produce a noticeable variation in sensory experience".

Without loss of generality, let us consider a video codec parameterized by a distortion level $x \in [0,1]$. When $x = 0$, the coding is lossless, i.e., the reconstructed video is identical to the source, so no distortion occurs. However, as x increases, the bitrate of the compressed video decreases, which implies that the likelihood increases that distortions can be perceived. Thus, when a video is compressed, the smallest distortion level x at which an observer can perceive visual distortion is the JND.

However, since physiological and visual attention mechanisms vary and involve many subjective factors, the JND is a quantity that depends on such indeterminate circumstances. In mathematical terms, the JND is a random variable. A discrete

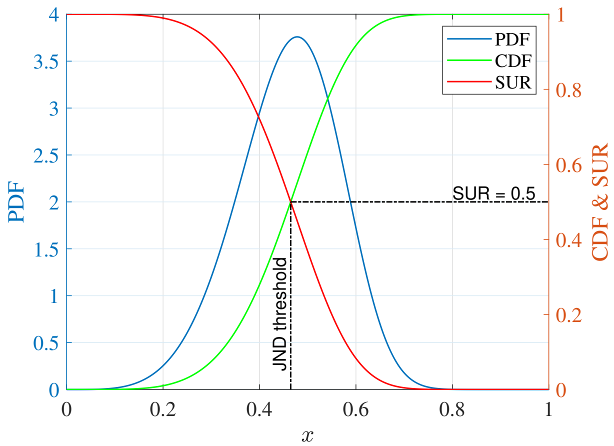


Fig. 1: Probability density function (PDF), cumulative distribution function (CDF) or psychometric function, and SUR function for a continuous JND random variable. The JND threshold is the distortion level where the psychometric function is equal to 1/2.

random variable is characterized by its probability distribution function and a continuous random variable by its probability density function. When giving a number for the JND, it usually refers to the JND threshold, namely the smallest distortion level for which the observer will notice a degradation of quality with a probability of 1/2.

The cumulative distribution of the JND random variable is a monotonically increasing function that specifies the probability of detecting distortion for each distortion level. In psychophysics, this function is called the psychometric function for the JND, and its graph typically is an S-shaped curve (Fig. 1).

For a whole population of subjects, the satisfied user ratio (SUR) denotes the probability that a randomly drawn subject cannot notice any distortion artifact when comparing a source video with its compressed version at a given distortion level. So the SUR is also a function of the distortion level and given by the complementary cumulative distribution function of the JND random variables, averaged over all subjects in the population. Modeling the SUR can help content providers minimize transmission costs while guaranteeing user satisfaction for any target proportion of their customers.

Since 2015, datasets of videos annotated with JND measurements have been created [2], [3], [4], [5]. These works apply a three-stage method which, for a given source stimulus, first determines the individual JND thresholds for all participating subjects of a lab study, then fits a distribution model to the threshold data, and finally computes the complementary cumulative distribution function as an estimate of the SUR curve.

This procedure models the individual psychometric functions as Heaviside step functions, each having the discontinuous jump at the individual JND threshold. However, this simplification cannot account for expected occurrences of detection of distortions at subthreshold distortion parameters. Consequently, the resulting SUR can be overestimated at subthreshold parameters. When applied in a system for rate-control, such an overestimate would make compression artifacts visible to a larger fraction of the viewer population than anticipated.

Instead of JND thresholds, entire psychometric functions must be attained for the correctness of the derived SUR curves. Models of psychometric functions come from different classes that describe the probability of distortion detection by the underlying sensory mechanism as a parameterized function of the distortion level [6]. There are several adaptive methods to estimate the parameters. For this purpose, we adopted a state-of-the-art Bayesian adaptive psychometric method, QUEST+ [7].

To estimate the SUR, we could use the adaptive psychometric method to sequentially estimate the cumulative JND distributions of many subjects to be averaged. However, alternatively, we may regard the entire population as a collective observer and apply QUEST+ to estimate its psychometric function directly. This can be implemented by newly choosing a random observer for each trial of the adaptive method. We demonstrate by simulation that this is more efficient than estimating and averaging individual JND distributions of subjects.

In 2014, the flicker test was introduced [8] to compare a source image and a distorted version. In this method, the test image is temporally interleaved with the source image, and artifacts in the test image may or may not appear to the viewer as a flicker effect. In the experiments, the flickering image is randomly displayed on the left or right side and the source image on the other side. The test subjects have to judge on which side a flicker effect is seen.

We show that the flicker test can also be applied for video quality assessment in the near-lossless quality range. The flicker test provides increased sensitivity to distortion for the human visual system. Moreover, it increases the precision compared to the plain test, i.e., side-by-side comparisons between a distorted video sequence and its source.

Evaluating video quality in controlled and standardized lab experiments is often time-consuming and may not fully capture the range of viewing conditions experienced in real-world settings. In contrast, crowdsourcing experiments provide various advantages, including a diverse participant pool, realistic hardware setups, and viewing environments that resemble typical users [9], [10]. Recent research [9], [10], [11], [12], [13] has shown that video quality ratings obtained through crowdsourcing are comparable to those obtained in lab settings.

In [14], a large JND-based dataset of JPEG and BPG compressed images was generated using crowdsourcing experiments. In this paper, we propose crowdsourcing and the flicker test to build a JND dataset for compressed videos.

There is an ongoing quest in research to develop the most accurate, efficient, and general models for predicting visual quality of compressed video. Examples are [15], [16], [17], [18] for prediction of mean opinion scores and [19], [20], [21], [22], [23], [24], [25], [26] for prediction of the JND, respectively the SUR. The most successful approaches are based on machine learning, especially deep learning models. However, deep learning is data hungry, and the datasets for video JND are lacking in size and quality. All current video JND datasets were produced in laboratory studies in which individual JNDs were obtained by a time-costly method, the binary search. This severely limited the number of video sequences and their JNDs. The data in some previous studies are also inconsistent, showing very large intersubject variability. Therefore, and as noted in recent papers [25], [27], [19], there is a need for larger and more consistent datasets for video JND to facilitate further advances in the state-of-the-art video JND prediction methods, in particular deep learning models.

Our paper makes four main contributions to help build such large and reliable video JND datasets.

- We show that the common approach of estimating the SUR function from the distribution of individual JND thresholds of a group of subjects causes a bias that may lead to an overestimation or underestimation of the SUR.
- We propose a new method to estimate the SUR. Our method, which we call the collective observer, randomly selects subjects to collect responses to paired comparisons. The distortion level for each paired comparison is decided by an adaptive psychometric testing procedure (QUEST+ [7]). We show by simulations that our method estimates the SUR more accurately and efficiently than the common approach.
- We use a flicker test to estimate the JND in encoded video sequences. Our experimental results show that the flicker test increases the sensitivity and precision of the estimation of the JND threshold.
- We conduct a within-subjects design subjective study to compare the estimated SUR using the flicker and plain tests. We use our QUEST+-based collective psychometric function method and crowdsourcing to build a JND dataset for 45 source videos encoded with advanced video coding (AVC) (a.k.a. H.264 and MPEG-4 Part 10) [28], [29] and versatile video coding (VVC) (a.k.a. H.266 and MPEG-I Part 3) [30], [31].

The remainder of this paper is organized as follows. Section II presents the state-of-the-art related to our contributions. We review the past attempts to estimate JND thresholds, fitted distributions, and corresponding SUR curves. Then we review the past work with the flicker test and present an overview of the current JND-based video quality sets. Section III gives formal definitions of the main terminology such as the individual and collective JND, their thresholds, psychometric functions, and the SUR. For Gaussian models of the JND, we show that a percentile of JND thresholds in an observer population is an overestimate of SUR at subthreshold distortion parameters. We introduce the concept of the collective observer and show by simulation that it provides a more efficient way to estimate the collective psychometric function and the SUR curve than averaging a set of individual JND distributions. Section IV explains how we adapted the flicker test for video quality assessment. Section V describes how to crowdsource JNDs for videos. Section VI compares the efficiency of the flicker test with conventional paired comparisons. For this purpose, we conducted a crowdsourcing experiment using a within-subject study design to evaluate the SUR curves for these source videos using QUEST+, with both the flicker test and the plain test.

II. Related work

In this section, we review the state-of-the-art with respect to the following main contributions of the paper: JND and SUR estimation, flicker test for quality assessment, and construction of video-based JND datasets.

A. JND and SUR estimation

In previous JND studies [32], [33], [4], [3], [34], [35], [36], [37], [14], the SUR function is built by aggregating the psychometric functions of a group of subjects. Each subject's psychometric function is estimated as a Heaviside function, with the step at the JND threshold and neglecting the subject's uncertainty in determining the JND. A paired comparison is typically used to estimate the JND threshold for a given subject. For example, when comparing a pair of videos, a high-quality source video and a distorted version are played side-by-side, one after the other or simultaneously. The subject indicates which video is of lower or higher quality depending on the subjective test question. In principle, the source video should be compared to the whole set of distorted versions. However, in practice, the search for the JND threshold is accelerated with binary search [5], [2], [4] or relaxed binary search [3].

Note that modern psychophysics approaches [38], [39], such as signal detection theory, argue that the observed JND is not an absolute quantity but depends on motivational and perceptual parameters. Thus, the JND for a given individual is a statistical rather than an exact quantity. We show in Section III that this implies that fitting a distribution to the obtained JND thresholds does not properly represent the overall, population-wise JND distribution and may result in an overestimation of the SUR. In this contribution, we propose a solution to overcome this limitation.

B. Flicker test for quality assessment

The flicker test was introduced by Hoffman and Stolitzka in 2014 [8] for image quality assessment to increase the sensitivity of the human visual system in detecting image artifacts in near-lossless image compression. The method proved effective and later was adopted by the JPEG AIC standard [40].

In [41], it was shown that the flicker test provides higher sensitivity for reconstructing impairment scales of distorted images than the plain test. Furthermore, it was shown that the flicker test achieves the same correlation with ground truth scores with a smaller number of required paired comparisons than the plain test. In [37], the flicker test was used for subjective picture-wise JND assessment to compare a slider-based method, a keystroke-based method, and the paired comparison with the relaxed binary search method. The flicker test was shown to provide about twice the sensitivity of a conventional side-by-side comparison for estimating the JND for JPEG compressed images. We attribute the benefits of the flicker test compared to the conventional side-by-side comparison to a reduction in the required visual short-term memory to make a judgment [42] and to the high sensitivity of the human visual system to detect a temporal contrast [43]. In [14], the flicker test was used in large-scale crowdsourcing experiments to determine the JND for JPEG and BPG compressed images.

First attempts to extend the flicker test to video sequences were briefly mentioned in [44]. The results were not regarded as promising, but no detailed report was given. In [40], the decision not to include the flicker test for video quality assessment in the JPEG XS standard was motivated by the fact that motion could mask the flicker, an effect called motion-silencing.

TABLE I: Comparison of the state-of-the-art JND-based video quality datasets

Datasets	Lin *et al.* [5]	MCL-JCV [2]	Huang *et al.* [4]	VideoSet [3]	our dataset
Publication year	2015	2016	2017	2017	2022
Number of source videos	5	30	40	220	45
Resolution of source videos	1920×1080	1920×1080	1920×1080	1920×1080[a]	640×480
Distortion type	AVC/HEVC	AVC	HEVC	AVC	AVC/VVC
Distortion levels per each stimulus	51/51	51	51	51	51/63
Test environment	lab	lab	lab	lab	online
Subjective assessment method	PC[b]	PC[b]	PC[c]	PC[b]	FT[d]/PC[c]
Search algorithm	binary search	binary search	binary search	relaxed binary search	QUEST+

[a] Three lower resolutions of the same video were also used: 1280×720, 960×540, and 640×360.
[b] paired comparison (PC): the two videos are displayed sequentially.
[c] PC: the two videos are displayed side-by-side.
[d] Flicker test (FT): a source high quality video and a flickering version are displayed side-by-side. Subjects determine which video is flickering.

In this contribution, we revisit and examine the flicker test for video quality assessment in a crowdsourcing study.

C. *JND-based video quality datasets*

The current JND datasets [2], [5], [4], [3] for compressed video differ in the number of source videos, resolution, compression type, subjective assessment method, and search algorithm. Table I summarizes these datasets. In the following, we briefly describe how each dataset was built.

The study by Lin *et al.* [5] involved five video sequences of resolution 1920×1080. The videos were displayed on a 65-inch TV with a resolution of 3840×2160. The viewing distance was 2 m from the center of the monitor. The video sequences were encoded with AVC and high efficiency video coding (HEVC). The video sequences were compressed by varying the value of the quantization parameter (QP) of the video codec from 1 to 51. A subjective study was conducted to determine the number of quality levels that a subject can distinguish. A bisection search method was used to determine these quality levels. In this search two videos are displayed sequentially, and the subject has to assess whether they are noticeably different. 20 subjects participated in the study.

Wang *et al.* [2] considered 30 source video sequences. More than 150 people participated in the study. JND samples were collected from 50 subjects. The other settings were as in [5]. The resulting JND dataset was called MCL-JCV.

Huang *et al.* [4] generated a JND-based dataset for HEVC. The dataset contains 40 high-definition (HD) source video clips with a frame rate of 30 fps and a duration of 5 s. All source videos were encoded using the HM 16.0 HEVC reference software, with QP values ranging from 0 to 51. To estimate the JND threshold for each source and its 51 encoded versions, a subjective test was conducted with 30 subjects. The source video and an encoded version were played side-by-side time-synchronously on a 65-inch 4K UHD TV display in a laboratory environment. The standard binary search was used to speed up the search for the JND threshold. Outliers were excluded with the three-sigma rule.

Wang *et al.* [3] built a large-scale JND video dataset called VideoSet for 220 source videos of 5 s in four resolutions (1080p, 720p, 540p, 360p). Each source video was compressed with AVC using QP values from 1 to 51. The viewing distance was set according to the ITU-R BT.2022 recommendation. The source video and a distorted version were displayed sequentially. A relaxed binary search was used to estimate the personal JND threshold for each subject. At least 30 subjects were involved in the JND estimation for each video sequence. Data from unreliable subjects and outliers were removed.

All of the current video JND datasets were built in laboratory experiments and are limited in size.

III. The Collective Psychometric Function

This section presents a formal definition of the collective psychometric function that underlies the SUR curve. For this purpose, we follow and adapt the notations introduced in [45]. Then we consider estimating the SUR and show by simulation that it is more efficient to directly estimate the collective psychometric function rather than averaging a number of individual psychometric functions which are estimated individually.

A. *Definitions*

The psychometric function models the relationship between the level of distortion of the stimulus and an observer's performance in detecting the distortion or discriminating between the distorted and the source stimulus. The SUR assumes a total population of observers and estimates the proportion of those who cannot detect the distortion.

We consider a lossy image or video compression scheme that produces monotonically increasing distortion magnitudes. The distortion depends on an encoding parameter that can take only a finite number of values. For example, for AVC and VVC, we use QP to control the quality of the encoded video. The range of QP values is $1, \ldots, 51$ for AVC and $1, \ldots, 63$ for VVC. Increasing QP decreases the bitrate and reduces the visual quality.

Definition 1 (Individual JND). For a given observer and a pristine source stimulus $S[0]$, we associate distorted stimuli $S[n],\ n = 1, \ldots, N$ corresponding to distortion levels $n = 1, \ldots, N$. The *individual just noticeable difference*, which we denote by JND, is a random variable whose value is the smallest distortion level n that can be perceived by the observer when the stimulus $S[n]$ is compared to the source stimulus $S[0]$.

We give an interpretation of this model in the context of the detection task, judging a test stimulus at distortion level n with respect to the corresponding source. According to the above definition, the observer will detect the distortion in the test stimulus with the probability that JND $\leq n$. Equivalently, we may state that the observer notices the distortion if a randomly drawn sample of JND has a value less than or equal to n.

In a paired comparison in the 2-alternative forced-choice (2AFC) setting, the two stimuli are presented in random order, and the task is to identify the one with distortion. In this case, an attentive observer will identify the distorted stimulus with probability $\frac{1}{2} + \frac{1}{2}\text{Prob}(\text{JND} \leq n)$.

To define the SUR, we consider a whole population of observers, each of whom has an individual JND distribution.

Definition 2 (Collective JND). Assume a population of observers and a pristine source stimulus $S[0]$ with distorted stimuli $S[n],\ n = 1, \ldots, N$. The *collective just noticeable difference*, which we denote by JND, is a random variable whose value is the smallest distortion level n that can be perceived by a random observer of the population when the stimulus $S[n]$ is compared to the source stimulus $S[0]$.

For a finite population of observers, the distribution of the collective JND is just the average of the distributions of the individual JNDs of all observers.

Definition 3 (JND threshold). The median of the individual (resp. collective) JND random variable is called the individual (resp. collective) *JND threshold*.

Definition 4 (Psychometric functions and SUR). The *individual and collective psychometric functions associated with the JND* are the cumulative distribution functions of the corresponding JND random variables. The SUR is the complementary cumulative distribution function of the collective JND random variable.

In the above definitions, discrete distortion levels $n = 1, \ldots, N$ are used, which is the case for the main application, i.e., JND-based quality assessment of compressed video sequences. It is straightforward, however, to rewrite and apply these definitions for continuous distortion levels such as additive Gaussian noise, parameterized by the noise amplitude as the distortion level. In the next subsection, for simplicity of notation, we assume that the distortion level is continuous.

B. Analysis of common SUR estimation

In previous work, the SUR is estimated by the complementary cumulative distribution function of a set of estimated individual JND thresholds. Several sources of error may affect the accuracy of this estimation. (1) The sample of JND thresholds may stem from a set of subjects that is not representative of the population, (2) JND thresholds estimated by approximation methods like bisection may be inaccurate, and (3) the fitting procedure to approximate the distribution of JND thresholds may also introduce errors.

We show that even if we could rule out all these sources of error, there still would be a systematic bias in the method. Typically, this bias will lead to an overestimation of the SUR, as shown by the following general counterexample.

Proposition 1. *Assume that the individual JND random variables are normally distributed with variance σ^2, and the individual JND thresholds are normally distributed with mean $\bar{\mu}$ and variance σ_0^2. Then the complementary cumulative distribution function of the individual JND thresholds overestimates the SUR for all distortion levels smaller than $\bar{\mu}$ and underestimates it for all distortion levels greater than $\bar{\mu}$.*

Proof. Let $\widehat{\text{SUR}}(x)$ denote the estimate of the satisfied user ratio at x given by the complementary cumulative distribution function of the individual JND thresholds. We have

$$\begin{aligned} \widehat{\text{SUR}}(x) &= \int_x^\infty \phi_{\bar{\mu},\sigma_0^2}(\mu)\, d\mu \\ &= 1 - \Phi\left(\frac{x - \bar{\mu}}{\sigma_0}\right) \end{aligned} \quad (1)$$

where ϕ_{a,b^2} denotes the probability density function of the normal distribution with mean a and variance b^2, and Φ is the cumulative distribution function corresponding to $\phi_{0,1}$.

Let $f(s)$ denote the probability density function of the collective JND random variable at distortion level s. From Definition 2, we have $f(s) = E[\phi_{M,\sigma^2}(s)]$, where E denotes the expectation operator and M is the random variable that gives the mean μ of the JND of a random observer. Thus,

$$f(s) = \int_{-\infty}^{\infty} \phi_{\mu,\sigma^2}(s)\phi_{\bar{\mu},\sigma_0^2}(\mu)\, d\mu.$$

Noting that $\phi_{\mu,\sigma^2}(s) = \phi_{0,\sigma^2}(s - \mu)$ and that the convolution of two Gaussians is a Gaussian, we obtain

$$\begin{aligned} f(s) &= \int_{-\infty}^{\infty} \phi_{0,\sigma^2}(s-\mu)\phi_{\bar{\mu},\sigma_0^2}(\mu)\, d\mu \\ &= (\phi_{0,\sigma^2} * \phi_{\bar{\mu},\sigma_0^2})(s) \\ &= \phi_{\bar{\mu},\sigma^2+\sigma_0^2}(s). \end{aligned}$$

The SUR at distortion level x is the complementary cumulative distribution function of $f(s)$,

$$\begin{aligned} \text{SUR}(x) &= \int_x^\infty \phi_{\bar{\mu},\sigma^2+\sigma_0^2}(s)\, ds \\ &= 1 - \Phi\left(\frac{x-\bar{\mu}}{\sqrt{\sigma^2+\sigma_0^2}}\right). \end{aligned} \quad (2)$$

If $x < \bar{\mu}$, then $(x - \bar{\mu})/\sqrt{\sigma^2 + \sigma_0^2} > (x - \bar{\mu})/\sigma_0$. Since Φ is strictly increasing, (1) and (2) give $\widehat{\text{SUR}}(x) > \text{SUR}(x)$. Similarly, if $x > \bar{\mu}$, then $\widehat{\text{SUR}}(x) < \text{SUR}(x)$. □

The SUR takes the variance σ^2 of the individual JND distributions into account, which is ignored by the estimate $\widehat{\text{SUR}}(x)$.

Using an example from the VideoSet data [3], we estimated the parameters $\bar{\mu}$ and σ_0 as 30 and 5, respectively. If we assume

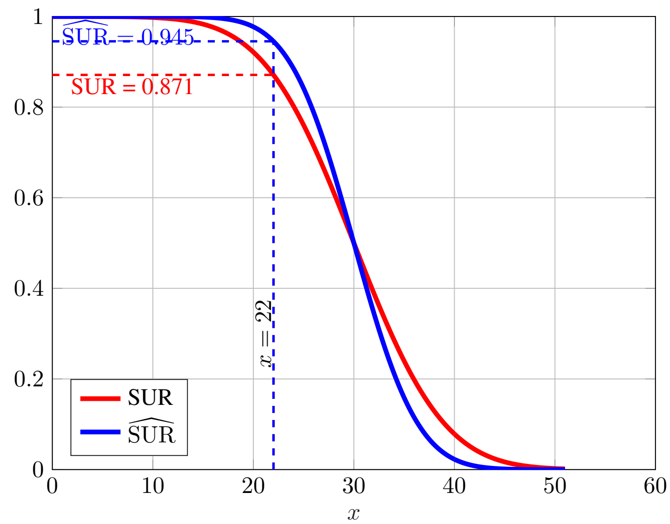


Fig. 2: Comparing the SUR curve of the collective JND distribution vs. its estimate using the common approach as discussed in Section III-B. In this plot, for the collective JND random variable X, we assumed each observer i has a JND random variable $X_i \sim \mathcal{N}(\mu_i, 5)$ where $\mu_i \sim \mathcal{N}(30, 25)$.

all participants have $\sigma = 5$, then for $x = 22 < \bar{\mu}$, we get the true value $\text{SUR}(x) = 0.871$, which is overestimated as $\widehat{\text{SUR}}(x) = 0.945$ as shown in Fig. 2.

In summary, we have provided a mathematical analysis of a general model of a population of observers with normal distribution functions of the individual JND random variables with different means and equal variance. Our result proves that fitting a distribution model to JND thresholds may over- or underestimate the SUR.

When the SUR is overestimated, media are transmitted at compression levels that are too strong. The fraction of viewers that will notice compression artifacts will be larger than anticipated. Similarly, when the SUR is underestimated, media are transmitted at higher bitrates than necessary. Our analysis will help to prevent such undesirable effects in practice.

C. The collective observer

To overcome the above limitations of the common method of SUR estimation, one should compute the collective JND distribution not from the individual JND thresholds alone, but from the average of the entire individual JND distributions. For this purpose, each participating subject in an empirical study could compare each source stimulus with a number of compressed versions of it and report whether a distortion can be detected. Different implementations of this process are possible. From such data, individual JND distribution functions can be fitted and averaged for each source stimulus. The complementary cumulative distribution functions serve as estimates of the corresponding SUR functions. We call this approach the “average observer”.

However, considering a fixed budget for a number of comparisons of one source stimulus with its distorted versions, we claim that it is inefficient to have a few subjects carry out multiple comparisons to assess each individual JND distribution, all of which then are averaged to estimate the collective JND distribution. Instead, it is better to have the comparisons done by many different subjects to directly estimate the collective JND distribution, as described in the following.

Definition 5 (Collective observer). A JND/SUR assessment method is called a “collective observer” if it directly estimates the collective JND distribution and the SUR function as follows. Responses to a number of paired comparisons between distorted stimuli and the source stimulus are collected where each response is obtained from a randomly selected observer. The collective JND distribution function is estimated by a fitting procedure applied to the collected responses.

In Subsection III-E, we present simulations that compare the accuracy, precision, and efficiency of the collective and average observer, as well as the common method of SUR estimation.

D. Modelling and estimation of the psychometric function

Psychometric functions for detection tasks like the one considered here are often modeled by S-shaped cumulative distributions functions $F(x; \alpha, \beta)$, where α and β are related to the threshold and the slope at the threshold, respectively. For example, in this section we use the Gaussian CDF

$$F(x; \alpha, \beta) = \Phi\left(\frac{x - \alpha}{\beta}\right). \tag{3}$$

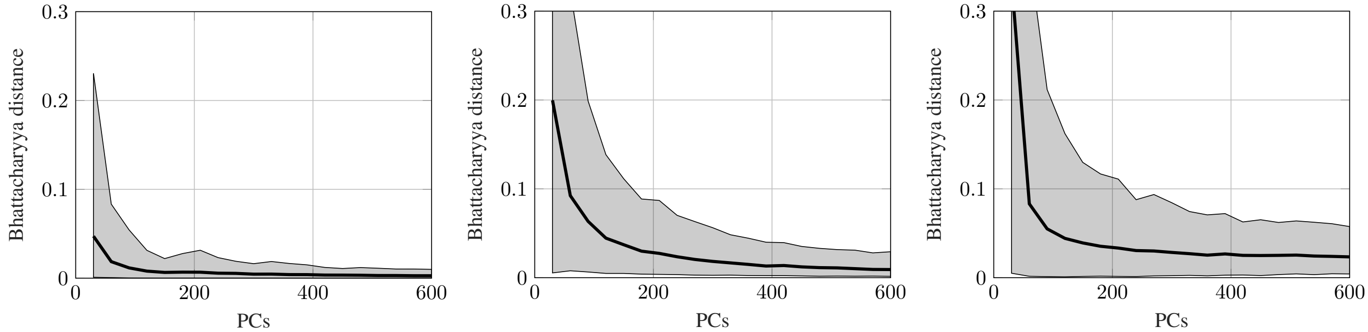
In psychophysics, it is common practice to use 2AFC questions in paired comparisons for JND threshold estimation. A pristine and a distorted stimulus are presented in random order, and the task is to identify the one with distortion. Just by guessing, the correct response comes up with a probability of 1/2. To account for that, the psychometric function w.r.t. the 2AFC setting, giving the probability of a correct response, is typically expressed as

$$\psi(x; \alpha, \beta, \lambda) = \frac{1}{2} + \left(\frac{1}{2} - \lambda\right) F(x; \alpha, \beta). \tag{4}$$

This includes a lapse rate λ, indicating a probability that the distorted stimulus is not identified, regardless of how strong the artifacts are. The lapse rate accounts, for example, for moments when the observer was inattentive or the view of the stimulus was obscured. From the cumulative distribution function $F(x; \alpha, \beta)$, the SUR is obtained as $1 - F(x; \alpha, \beta)$.

There are many methods to estimate psychometric functions empirically. Most are designed to just estimate the threshold α as this is often the most important and only parameter of interest. To quantify the SUR, however, we need the whole psychometric functions, including the slope parameter β.

For several decades, adaptive methods have been researched, in which an algorithm decides the next stimulus for a paired comparison, based on the observer responses for the previous comparisons. Two of the most prominent ones are PEST [46] and QUEST+ [7]. In our work, we chose QUEST+, a recent Bayesian method, which offers a very large palette of application scenarios.



(a) Collective observer (b) Average observer (c) Common approach

Fig. 3: Average Bhattacharyya distance of estimated JND distributions to the ground truth for different budgets of paired comparisons, derived from 1000 runs. The shaded areas are the 95% confidence intervals.

TABLE II: Parameters of 10000 simulated observers were selected randomly from truncated normal distributions.

Parameter	Mean	Variance	Lower bound	Upper bound
Threshold α	26	36	1	51
Standard deviation β	5.5	1.12	1	10
Lapse rate λ	0.02	0.00002	0	0.04

E. Comparison of the collective observer with other approaches

We compared the proposed collective observer for estimating the collective psychometric function $F(x;\alpha,\beta)$ with the average observer and the common SUR estimation approach. For this purpose, we used a simulation. In a simulation, the ground truth collective JND distribution of a population is available, allowing to study accuracy and efficiency of competing methods.

1) Simulated population of observers: We generated a population of 10000 simulated observers, each represented by an individual psychometric function with a Gaussian CDF from Equation (4). Gaussian CDFs have also been used in prior research for the modeling of the JND [2], [3]. For each observer, we randomly selected the parameters α, β in $F(x;\alpha,\beta)$ and the lapse rate λ from truncated normally distributed values. Table II summarizes these normal distributions.

The average of all individual psychometric functions $F(x;\alpha,\beta)$ is the ground truth collective psychometric function in our simulation. To numerically compare estimates with the ground truth, we represented the collective psychometric function and the SUR by samples at equally spaced distortion levels $x_m = 1 + \frac{m}{100} \in [1, 51], m = 0, \ldots, 5000$.

In the following we present the implementation details for the simulation of the collective and average observer as well as the common estimation method. Simulations were then run with a fixed budget of n paired comparisons.

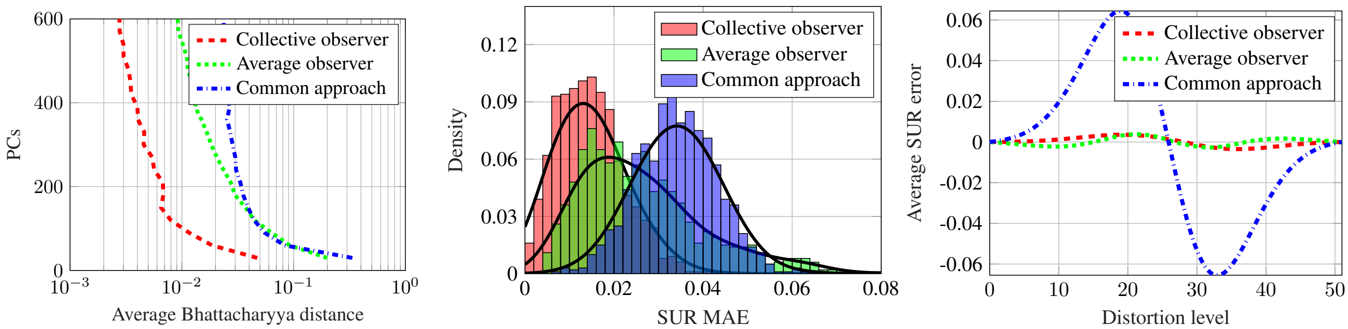
2) Collective observer: We estimated the collective psychometric function using the adaptive psychometric procedure QUEST+. We simulated 2AFC paired comparisons between distorted stimuli at levels x and the undistorted stimulus. Following the principle of the collective observer, we randomly selected one of the 10000 observers for each comparison and evaluated the corresponding individual psychometric function (4), giving the probability of a correct response. The simulated response was then drawn according to this probability. The distortion levels for these comparisons were adaptively decided by QUEST+ until the given budget was used up.

3) Average observer: In the average observer, individual psychometric functions $F(x;\alpha,\beta)$ and corresponding JND distributions are estimated separately for up to 20 randomly selected subjects from the population and then averaged. For each of the chosen subjects, the individual psychometric function was obtained by simulation of 2AFC paired comparisons using QUEST+ as above, however, applying only the particular psychometric function $\psi(x;\alpha,\beta,\lambda)$ of the corresponding subject. The number of comparisons per subject was set to 30, which is the recommended number for QUEST+ [7]. The number of subjects was determined by the given budget.

4) Common SUR estimation: Commonly, the SUR function is estimated by the complementary cumulative distribution of a Gaussian density function that is fitted to a set of estimated individual JND thresholds [5], [2], [4], [3]. To assess these thresholds, we followed the latest of the above works, which proposed a relaxed binary search method. For details, see [3]. The search space for the thresholds consisted of the integer range from 1 to 51. Note that the number of comparisons for each threshold estimate may vary between 10 and 11. We estimated JND thresholds from randomly drawn subjects until the budget was used up. For the set of estimated thresholds, a Gaussian density function was fitted by maximum likelihood estimation.

5) Accuracy of estimated collective JND distribution and SUR: We compared the estimated collective JND distributions, evaluated at the chosen distortion levels x_m. The sampled values were scaled to yield probability distributions. To compare an estimated distribution with the ground truth, we computed the Bhattacharyya distance measuring the magnitude of the difference between two probability distributions. Smaller distances indicate better approximation.

We compare estimates of the SUR function using the mean absolute error (MAE), $\frac{1}{5001}\sum_{m=0}^{5000} |\widehat{\mathrm{SUR}}(x_m) - \mathrm{SUR}(x_m)|$. To check for the bias of the SUR estimate that is expected from Proposition 1 for the common approach, we computed the signed error at each distortion level, $\widehat{\mathrm{SUR}}(x_m) - \mathrm{SUR}(x_m)$, averaged over all simulation runs.



(a) Number of paired comparisons to meet accuracy constraint.

(b) Histograms of MAE of SUR estimates from 600 PCs over 51 distortion levels.

(c) Signed SUR error, averaged over all simulation runs with 600 PCs.

Fig. 4: Comparison of the collective observer, the average observer, and the common approach from 1000 simulation runs each. Shown are (a) the computational complexities to yield an approximation of the collective JND distribution, (b) the achieved mean absolute errors (MAE) for the estimated SUR functions using 600 PCs, and (c) the bias for each method at all distortion levels, also with 600 PCs each.

6) Results: After generating the simulated observers and computing the ground truth, we carried out 1000 simulation runs for each assessment method and each budget of $n = 30k$, $k = 1, \ldots, 20$ paired comparisons.

Fig. 3 shows the average Bhattacharyya distances to the ground truth distribution with 95% confidence intervals (CIs). The mean distance for the collective observer clearly is much smaller than that of the average observer and the common approach. To better compare the efficiency of the methods, Fig. 4a plots the required number of paired comparisons to achieve a specified accuracy in terms of the average Bhattacharyya distance. For instance, let us consider 30 individual JND thresholds as assessed in VideoSet [3]. This requires approximately 330 paired comparisons using the relaxed binary search method to estimate the SUR for each source video. In our simulation, 330 paired comparisons yielded a Bhattacharyya distance of 0.027. By employing linear interpolation of the data presented in the figure, we estimate that the collective observer would only need 51 paired comparisons to achieve the same mean distance, while the average observer would require 277.

In Figs. 4b and 4c we study the accuracy and bias of the estimates of the SUR function. The histograms and density plots for the SUR MAE clearly show the superiority of the collective observer regarding accuracy. Fig. 4c reveals that in the simulations the common approach overestimates the SUR at substhreshold distortion parameters and underestimates it at suprathreshold parameters. In comparison, the magnitude of the corresponding bias for the collective and the average observers appear negligible.

In summary, we have shown by simulation that the average observer as well as the collective observer are bias-free and approximate the collective JND and SUR functions with decreasing error when the number of paired comparisons is increasing. The collective observer is more efficient than the average observer and much more efficient than the current state-of-the-art method that uses relaxed binary search to estimate JND thresholds to be fitted by a Gaussian distribution.

We conclude this section by listing its main contributions in the following:

- We provided a clear framework of definitions for individual and collective just noticeable difference.
- We showed that the current state-of-the-art method to estimate satisfied user ratio curves, based on fitting a model to JND thresholds, suffers from a systematic bias, causing an overestimation for subthreshold distortion levels and underestimation for suprathreshold parameters.
- We introduced a new approach for the estimation of the distribution of the just noticeable difference and satisfied user ratio curves, based on the collective observer.
- We showed by a large-scale simulation that the collective observer is much more accurate and efficient than the common state-of-the-art method.
- The simulation confirmed the systematic bias that was derived theoretically for the common method, while there was hardly any such bias for the collective observer.

IV. Flicker test for video-wise JND assessment

In this section, we describe how we adapted the flicker test for JND-based video quality assessment. Also, we explain how we re-encoded the videos for transmission to the users' computers in the crowdsourcing experiment.

A. *Test videos*

1) Encoding source video sequences: We selected 45 source videos with diverse content from VideoSet [3]. The source video clips have a duration of 5 s, a spatial resolution of 1920×1080 pixels (full high-definition resolution), frame rates of 24 and 30 fps, and color format YUV420p. The videos do not include audio.

For encoding with H.264/AVC, we used FFmpeg with the libx264 implementation of H.264/AVC with the "high" profile. We disabled adaptive scene cut detection, used the quantization parameter (QP) as the primary bit rate control method with $\text{QP} = 1, \ldots, 51$, disabled adaptive quantization, and set the GOP size to the frame rate of the input video (24 or 30).

For encoding with H.266/VVC, we used the Fraunhofer Versatile Video Encoder (VVenC) software with the expert mode

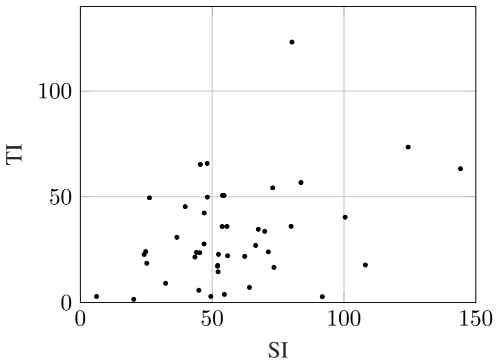


Fig. 5: TI vs. SI of the 45 source videos.

encoder (vvencFFapp) implementation of H.266/VVC [47]. We applied the "medium" preset, QP $= 1, \ldots, 63$, disabled adaptive quantization, and set an intra-frame period of length 64.

Fig. 5 illustrates the diversity of the selected 45 cropped source videos by showing temporal information (TI) against spatial information (SI). TI and SI were calculated as in [48].

B. Generating flickering versions of the compressed test videos

For assessment of the QP at the JND in a sequence of compressed versions of a source video, observers compare compressed videos with the source video and report whether they can perceive a difference between them. More precisely, the order of the two videos to be compared is random, and the observer must choose each time the distorted video. The videos in the comparison can be played side-by-side or sequentially. As an alternative to the playback of the compressed video, we adapted the flicker test that was proposed by the ISO/IEC standard 29170-2 for image quality assessment [40] to image sequences as follows.

The frames of the source video alternate with the frames of a compressed version with a temporal frequency of 8 Hz. For example, suppose a source video has a frame rate of 24 fps. Then, the first three frames of the source video are played, followed by the next three frames, taken from the encoded version. Next, the subsequent three frames of the source video are played, and so on. We decoded the source and compressed videos into RGB frames to create the flickering videos. The observers are presented with such flickering videos side-by-side with the source video and asked on which side they can notice any flicker.

The generated flicker videos have a resolution of 1920×1080 pixels, which is not practical for crowdsourcing. Therefore, we cropped the sources, the flickering, and the compressed videos to 640×480 pixels.

C. Video transmission to crowdworkers

In our empirical study, we compared side-by-side compressed and source videos, as well as flickering and source videos. For our online study, several such comparisons were grouped in batches to be processed by observers. To avoid stalling and other network and bandwidth-related issues, all videos that were scheduled in a batch were loaded onto the participant's computer before playback. A flickering video generally has a high bitrate, roughly the same as the bitrate of the source video because it contains frames from the source. Therefore, we encoded the videos for transmission in a visually lossless setting to reduce the download time in crowdsourcing.

In previous work using videos in crowdsourcing, authors kept file sizes manageable by compressing videos with a constant rate factor (CRF) of 18 in [49], an average file size of 1.23 MB in [50], or cropped the high-resolution videos to 540p in [10]. In our work, we compressed the videos with a CRF of 12 to ensure perceptually lossless compression. We conducted a small study to compare the JND location assessed by a small group (10 people) for 10 source videos at CRF of 5, 10, and 12 using the flickering test. We found that there was no statistically significant difference between the JND locations when using these CRFs. Therefore we set the CRF to 12 to encode the video sequences for transmission in our main crowdsourcing experiment.

For visually lossless compression of the videos, we used the x264 implementation of H.264/AVC with the "high" profile and "veryslow" preset. We used CRF rate control and a GOP size equal to the frame rate.

We used a commercial global content delivery network service for fast, low-latency delivery of the test videos. Prior to playback, the videos were preloaded onto the users' machines.

V. CROWDSOURCING STUDY FOR JND ASSESSMENT

In our online experiment, we collected paired comparison responses to estimate SUR curves for all combinations of source videos, codecs, and test modalities. Thus, for each of the 45 source videos, we estimated the collective psychometric functions for the following sequences of stimuli:

- Compressed with AVC (QP $= 0, \ldots, 51$),
- Compressed with AVC (QP $= 0, \ldots, 51$) and interleaved with the source,
- Compressed with VVC (QP $= 0, \ldots, 63$),
- Compressed with VVC (QP $= 0, \ldots, 63$) and interleaved with the source.

This resulted in the estimation of 180 SUR curves.

We defined a *question* as showing a participant a side-by-side paired comparison and asking the participant to select the side of the flickering video in the flicker test or to identify the side of the video with the lower quality in the plain test.

We used the freelancer.com platform (www.freelancer.com), an online job marketplace that allows *clients* and *freelancers* to collaborate. On this platform, *clients* create and submit projects and *freelancers* bid to carry out the work. *Freelancers* can communicate and chat with the *client*. *Clients* can also contact their *freelancers*.

A. Overview

Fig. 6 shows the workflow of our study. To be eligible to participate in the main study, subjects signed a consent form and obtained a qualification label by passing a quiz. The main study was divided into many small tasks. Once a task was completed, a subject could take a break before proceeding to the next task.

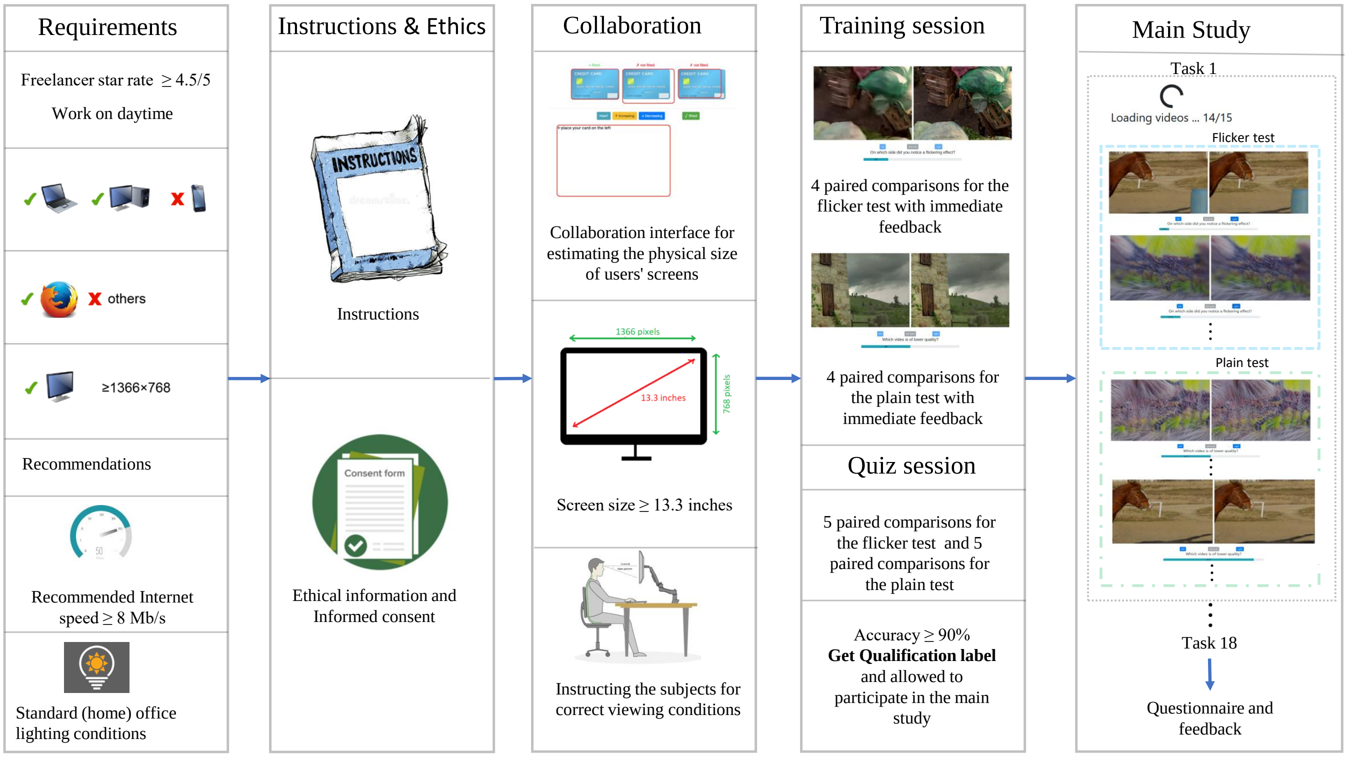


Fig. 6: Workflow of the online subjective JND-based video quality assessment study.

B. Requirements

To ensure quality, we invited freelancers with a rating of at least 4.5/5 from their previous jobs. The following requirements were automatically checked.

- Desktops and laptops were allowed, while cell phones and tablets were not.
- Firefox browser.
- Minimum logical resolution of 1366×768 pixels.
- Work was allowed only during the daytime of the freelancer time zone.
- Minimum screen size of 13.3 inches.

We used the JavaScript "navigator" object to access the user agent information and determine the type of device being used by the participants. To get the logical resolution of the screen used by a browser, we used JavaScript and the "window" object's properties. Participants were instructed to maximize their browser window. If the browser was not maximized, participants were prompted to maximize the browser, and the experiment was paused until compliance was ensured. The method used to estimate the physical screen size of the users is described in Section V-E.

To display the entire graphical user interface (GUI) of our subjective experiment, participants were required to have a display with a minimum logical resolution of 1366×768 pixels and a minimum physical size of 13.3 inches. The pixel density for screens with this resolution and size is 117.8 pixels per inch (PPI). In this case, the stimuli were displayed with a one-by-one pixel ratio. Up-sampling was carried out for screens with higher pixel density, while screens with lower pixel density required down-sampling. However, up-sampling does not cause any information loss and can be expected to have a negligible effect on the perceived quality. On the other hand, the down-sampling in our experiment resulted in only a slight reduction of the logical image pixel density from 117.8 PPI to 102.46 PPI in the worst case. Therefore, we do not expect that this process significantly affected the results of the paired comparisons.

Moreover, freelancers were advised to check that their Internet download speed was at least 8 Mb/s.

C. Instructions

Participants were given instructions in four steps. In the first step, they were familiarized with the flickering videos and the flicker effect by showing a source video and two flickering versions, one with barely perceptible distortion (flickering effect) and one with a strong flicker effect, one after the other. The second step showed a paired comparison between a source video and a flickering version. The paired comparison was shown for 5 s, followed by a 3 s phase, in which an interface gave the subjects the option of choosing the side of the flickering video or pressing the "not sure" button. The "not sure" option was used in [51] for paired comparisons to reduce stress and fatigue. In [52], we demonstrated through a subjective study that the inclusion of the "not sure" response option in the forced-choice method reduces mental load and results in models with improved goodness of fit. In the third step, we familiarized the participants with the compressed videos. Another source video and two compressed versions were shown, one with barely noticeable distortion and one with heavy compression artifacts. The fourth step showed a paired comparison between a source video and its compressed version for the plain test. Finally, the study and the payment method were explained.

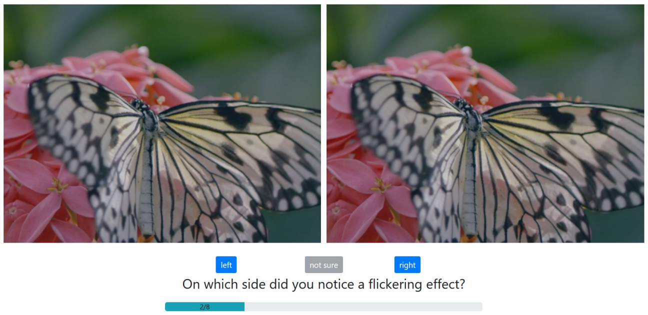


Fig. 7: User interface for the flicker test.

D. Ethics

Ethical approval of the experimental procedures and protocols was granted by the Institutional Review Board of the University of Konstanz. The participants were given the study instructions with further information, including the purpose and benefits of the study and legal rights. Participants were given the opportunity to ask questions and were requested to sign an informed consent form.

E. Calibration

In an online assessment of perceived video quality, it is challenging to ensure uniform experimental conditions for all participants [53]. For example, the stimuli would have different physical sizes on screens with different sizes and resolutions, which would affect the perceived quality. Therefore, we displayed the videos with a fixed physical size on all participants' screens and recommended a fixed viewing distance. For this purpose, we implemented the calibration method described in [14] to estimate the physical size of the screens of the participants. In this method, after imposing the minimum resolution described in Section V-B, the screen size is estimated with the virtual chinrest method [54].

As a result, each video was rendered on all participants' screens with a fixed physical size of $138\,\mathrm{mm} \times 103.5\,\mathrm{mm}$. Therefore, for the side-by-side comparison, a fixed physical size of $281\,\mathrm{mm} \times 103.5\,\mathrm{mm}$ was used, including $5\,\mathrm{mm}$ of white space between the paired videos. The corresponding calibration parameters were stored in the local memory of the participants' browsers. If the browser zoom level was changed after calibration, the participant was asked to restore it or redo the calibration.

We also asked the participants to adjust their viewing distance to $60\,\mathrm{cm}$. This distance was derived by the trigonometric calculation described in [54] and the ISO standard [55] for two videos of width $138\,\mathrm{mm}$ with a $5\,\mathrm{mm}$ blank space in between.

F. Training session

To guide the participants in using our user interface and familiarize them with the subjective task, we asked them eight questions. The first four questions were presented using the flicker test. The participants were shown a flickering video with its high-quality source video and asked to choose "right", "left", or "not sure" in response to the question "On which side did you notice a flicker effect?" (Fig. 7). These stimuli were manually selected. The stimuli in two questions were compressed using AVC, and the stimuli in the other two questions were compressed using VVC. The next four questions were about the plain test. The test question was "Which video is of lower quality?"

The order of the questions was randomized at the beginning of each training session for each participant. Participants were not allowed to work on a question until all required videos for all the training questions were loaded. In one training question of both the flicker test and plain test, two identical videos had to be compared side-by-side. Only the answer "not sure" was correct for this question. In another training question, the source video was compared to a highly compressed video or a video with a strong flicker effect. The only correct answer was to choose the side of the compressed/flickering video. In two other questions, a source video was compared to its compressed or flicker version with barely perceptible distortion. Choosing the side of the compressed video or "not sure" would be the correct answer. If the answer was correct, a message confirmed that the participant made the right choice. If the answer was incorrect, a message explained why the choice was incorrect and helped the participant on how to give the correct answer.

For each question, the source stimulus and its flickering or compressed version were presented side by side randomly on the left and right sides. During playback the "left" and "right" buttons were enabled but the "not sure" button was disabled. If the participant made a choice before the end of the video duration (5 s), the next question was shown. Otherwise, the participant was shown a decision-making interface for $3\,\mathrm{s}$. In this interface, the "left", "right", and "not sure" buttons were enabled. A progress bar was presented to visualize progress.

G. Quiz session

Participants were only allowed to take part in the main study if they passed a quiz. There were 10 quiz questions, with the first five for the flicker test, and the second five questions for the plain test. Although the content of the quiz videos was different from that of the training videos, the steps for answering a quiz question were the same as those for answering a training question, except that no feedback was provided to participants after each quiz question.

In the quiz session, for each test condition, i.e., flicker test or plain test, the compared videos were identical in one of the questions. For this question, the answer "not sure" was the correct response. In another question, the flicker effect or compression level was strong. Thus, the correct response was to select the side of the flickering, resp. compressed video. For the remaining three questions, the stimuli had barely perceptible distortions around the JND location assessed by the authors. Therefore, for these three questions, choosing the side with the flickering, resp. compressed video or pressing the "not sure" button was regarded as the correct response.

Once a participant completed the quiz, the result was sent to our server. Participants with a score of at least eight correct answers for the ten questions received a qualification label, and a new page with the link to the main study was displayed.

Participants with lower score received a message informing them of their failure and were not allowed to redo the training and quiz sessions.

H. Main study

When a participant with a valid qualification label opened the link to the main study, the requirements of Subsection V-B were checked. If these were satisfied, the local browser storage was checked to determine whether the calibration was completed. If there was such a record, the participant was allowed to start the main study. If the browser was not already maximized, the participant was asked to do so. On the other hand, if no record was saved in the browser's local storage, the participant had to repeat the calibration, training, and quiz.

For each sequence of stimuli, the sequence of QPs that define the requested paired comparisons was generated adaptively by QUEST+. After a participant answered the paired comparisons in a task, the results were sent to the corresponding QUEST+ objects in our server, and the next QPs were determined for delivery to other participants.

Each task of ten questions consisted of two parts. For the first part, five different sources were randomly picked together with one of the two encoders, AVC or VVC, and the flicker test condition. The second part used the same five source videos and encoders, but with the plain test condition. The QPs for all compressed resp. flickering versions were determined by QUEST+

This task structure provided a within-subject study design for comparing the flicker test to the plain test.

We used the collective observer method described in Section III to estimate each of the 180 collective psychometric functions. To emulate the collective observer, we restricted our subjects to provide only a single response to the paired comparisons per sequence of stimuli. Furthermore, because the paired comparisons were presented to participants in batches of 10, each participant could complete no more than 18 tasks with 10 comparisons each.

To correctly assemble the tasks for delivery to participants, our server maintained a table of available paired comparisons and, for each participant, a list of already issued comparisons. The table showed whether there was a next available QP for each of the 180 combinations of the source video, encoder, and test condition. If that was the case, the QP was given. Otherwise, a flag was set.

For each task, the server uploaded 15 videos to the computer of a freelancer: five source videos (used in both parts of the task as source videos), five compressed versions (for the plain test), and five compressed and interleaved versions (for the flicker test). After the download was completed, the participant could start to answer the corresponding 10 questions. As in the training session, video pairs were played for 5 s. Participants who did not make a decision during these 5 s were given an additional 3 s.

When the task was finished, the participant was asked to rest, reread the instruction, do the next task, or quit the experiment. Also, counters that showed the number of the already finished and the remaining open tasks were shown.

If a participant did not answer a paired comparison for one of the test conditions (flicker test or plain test) within the given 8 s, the answer for the question of the same source video in the other test condition was also discarded, and both paired comparisons were returned to the server to be reinserted in the table for future tasks.

VI. Experimental results

In total, 67 freelancers participated in the training session and took the quiz. Of the 57 freelancers who passed the quiz, 55 placed a bid. Then, through the freelancers.com chat tool, the first author discussed the project with them to ensure they fully understood the test. Of the 55 freelancers who submitted a bid, 51 took part in the study. Some demographic and experimental data collected through questionnaires from the freelancers are illustrated in Fig. 8.

The average time taken by freelancers to complete the initial tasks, including reading instructions, filling in the consent form, and performing calibration, was 5:21 min. The training session lasted approximately 1:45 min, followed by a quiz taking an average of 1:09 min.

Workers who successfully passed the quiz were eligible to participate in the main study, where they completed up to 18 assignments. Each assignment consisted of 10 paired comparisons (5 for each test condition). On average, workers spent 0:57 min answering the 10 questions per assignment. The duration for video preloading was not considered in these calculations.

A. Data filtering

In the subjective experiment, we used the relaxed forced-choice method. In addition to the "right" and "left" stimuli, this method also provides the "not sure" option. With the flicker test, a reliable subject would either select the flickering video correctly or press "not sure" if the flicker effect was below the perceptual threshold. Therefore, a subject who selected the source video was considered inattentive at this time, and we discarded these responses and their corresponding responses for the plain test to ensure the within-subjects design. For the plain test, we proceeded accordingly. As a result, we removed 5.24% of all responses. We then reconstructed the psychometric functions of the population based on the remaining responses.

B. Probability distribution fitting

For the psychometric function in (4), we used the Weibull cumulative distribution function

$$F(x;\alpha,\beta) = 1 - e^{-(x/\alpha)^\beta} \quad \text{for} \quad x \geq 0$$

where $\alpha > 0$ is the scale parameter and $\beta > 0$ is called the slope. The Weibull distribution is flexible and allows the fitting of non-symmetric curves.

The fitting is based on maximum likelihood estimation. Since each psychometric function can only specify the probability of correct detection, we needed to also define the probability of the "not sure" response. In the model of the 2AFC setting, the participant does not have the "not sure" option and would have

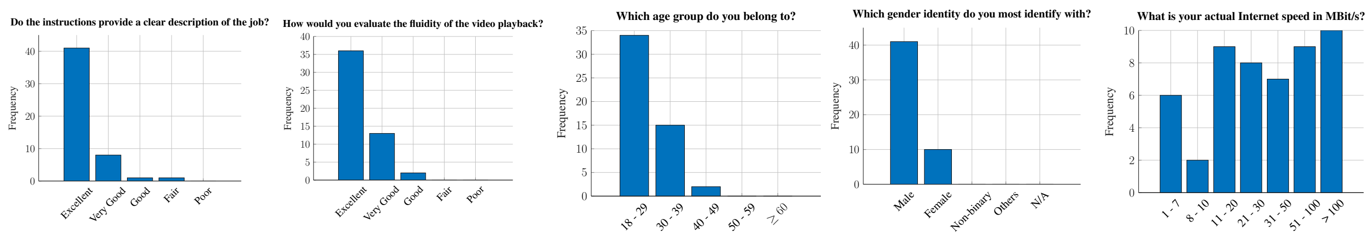


Fig. 8: Participant diversity and experimental insights: Data from 51 freelancers in the subjective study.

to select one of the sides “left” or “right” at random. This led us to treat a “not sure” response as two responses, one for “left” and one for “right”, each weighted by a factor of 1/2.

Using QUEST+, we also included the lapse rate parameter λ of the psychometric function (4) in the estimation besides the scale α and the slope β. Between 43 to 51 paired comparisons were collected to estimate the parameters of a psychometric function. Fig. 9 presents examples of the resulting SUR functions $\text{SUR}(x;\alpha,\beta) = 1 - F(x;\alpha,\beta)$ for some source videos using the flicker test and the plain test.

For further analysis of the results in terms of sensitivity and precision of the flicker test compared with the plain test, we calculated the JND threshold x_{JND} for each Weibull JND distribution with $F(x_{\text{JND}};\alpha,\beta) = 0.5$ and the variance $\sigma^2 = \alpha^2\left[\Gamma(1+2/\beta) - (\Gamma(1+1/\beta))^2\right]$, where Γ is the Gamma function.

C. Sensitivity

Fig. 9 shows estimated SUR curves for the flicker test and the plain test. Curves that are more to the left indicate a more sensitive JND assessment because differences to the source video are detected for smaller QPs. This shift of the JND caused by the flicker test can be quantified by Δ_{JND}, the JND threshold assessed with the flicker test minus the threshold assessed with the plain test for the same source video compressed with the same video codec. The figure shows that the flicker test provided a more sensitive JND assessment ($\Delta_{\text{JND}} < 0$) for the source videos SRC129, SRC193, and SRC009 and both codecs AVC and VVC. However, for the source video SRC059, the plain test was more sensitive ($\Delta_{\text{JND}} > 0$).

To summarize the comparison of the two test conditions for all 45 sources and both codecs, we show the boxplots of the JND thresholds for the videos compressed with AVC and VVC in Fig. 10 (a) and (c). We note that the mean thresholds derived with the flicker test are smaller than those from the plain test (33.5 and 36.4 for the flicker test vs. 34.1 and 38.5 for the plain test). Parts (b) and (d) of the figure show the differences between the JND thresholds estimated with the flicker test and those estimated with the plain test. The differences are negative in most cases. Overall, the result shows that, on average, the flicker test was more sensitive than the plain test.

To check whether this finding is statistically significant, we conducted a nonparametric paired samples Wilcoxon signed-rank test to examine the hypothesis that the flicker test has a higher sensitivity than the plain test.

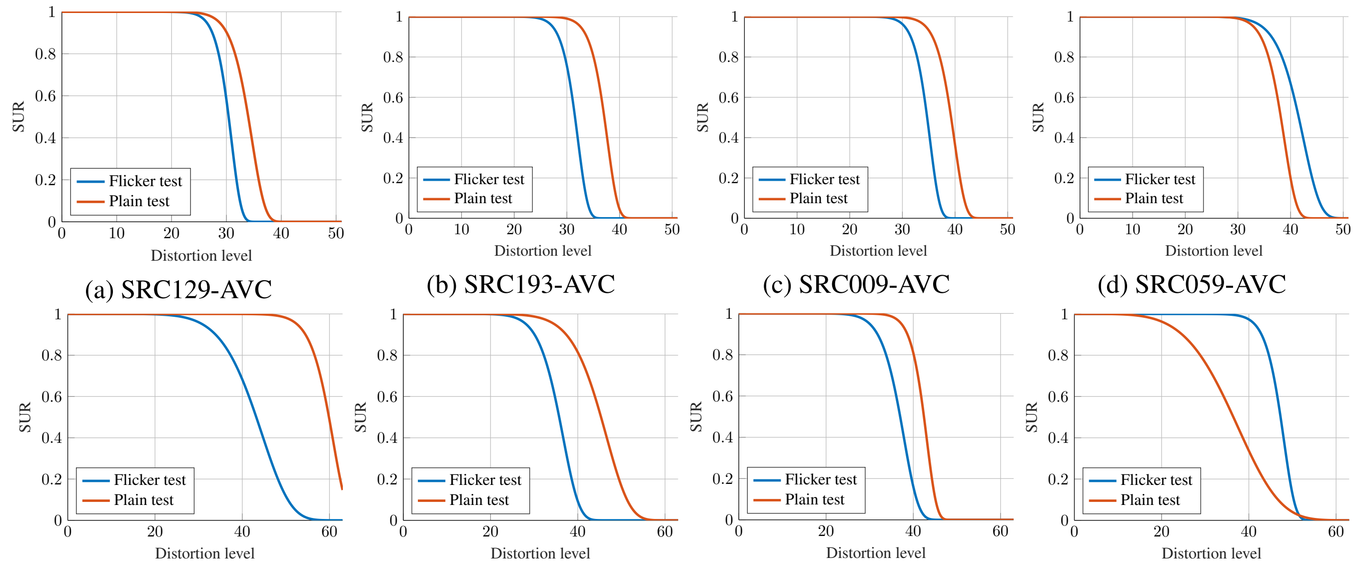
The null hypothesis was that the median of the JND differences for the same source videos and compression codec comes from a distribution with a median of zero against the alternate that the median is less than zero. The p-value for the test was 0.0004, which means that the test rejected the null hypothesis with 95% confidence level in favor of greater sensitivity of the flicker test.

In some cases, the flicker test was less sensitive than the plain test ($\Delta_{\text{JND}} > 0$), namely for 16 of 45 source videos for AVC compressed videos, and for 10 of 45 source videos for VVC compressed videos, see Figs. 9 (d,h) and 10 (b,d). This may be due to the strong motion in the videos, which may have masked the flicker effect. Videos with strong motion have large temporal information. Fig. 11 shows the temporal information versus Δ_{JND}. The plain test is more sensitive when $\Delta_{\text{JND}} > 0$, i.e., in the region on the right side of the vertical line. The source video sequences, determined to have high motion through a visual inspection conducted by the authors, were marked with crosses. For most of these videos, the flicker test did not yield higher sensitivity. This confirms that strong motion in the video may indeed mask the flicker effect, thereby reducing the sensitivity.

D. Precision

As common in statistics, we express the precision of a random variable as the reciprocal of its variance. A small variance of the collective JND results in high precision and typically corresponds to a larger (absolute) slope of the SUR curve at the JND threshold, as seen in the examples for the flicker test in Fig. 9. To compare the two test conditions for all 45 sources and both video codecs, we show the boxplots of the variances of the JND distributions and the pairwise differences of variance in Fig. 12. The variances derived from the flicker test typically are smaller than those from the plain test. Negative Δ_{Variance} was observed in 35 out of 45 source videos for AVC compression, and in 26 out of 45 source videos for VVC compression.

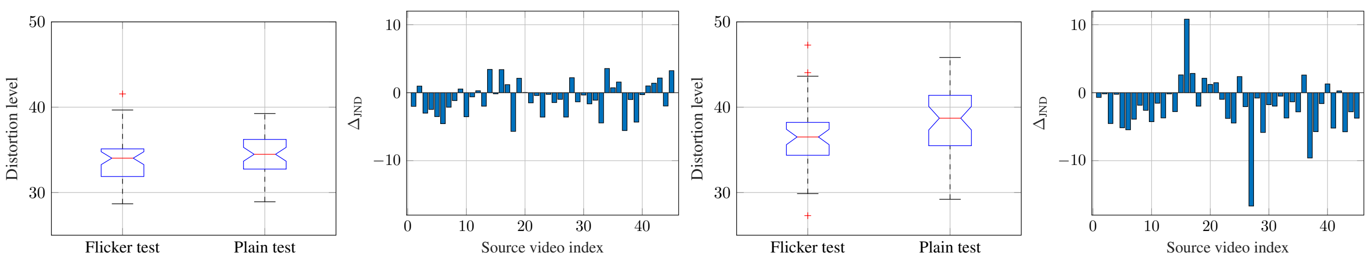
We applied the paired-samples Wilcoxon sign-rank test to examine the hypothesis that the flicker test yields smaller variances in the estimated JND distributions and thus provides greater precision in JND threshold assessment than the plain test. The null hypothesis of the test was that the differences between the variances for the same source videos and compression codec come from a distribution of zero median against the alternate hypothesis that the median is less than zero. The p-value of the test was 0.033, clearly rejecting the null hypothesis at the 5% significance level.



(a) SRC129-AVC (b) SRC193-AVC (c) SRC009-AVC (d) SRC059-AVC

(e) SRC129-VVC (f) SRC193-VVC (g) SRC009-VVC (h) SRC059-VVC

Fig. 9: SUR curves for the flicker test and the plain test. The first three columns show the SUR curves with and without the flicker test for the source videos with the smallest (negative) Δ_{JND} values, averaged for AVC and VVC, while the fourth column shows them for the source video with the largest average Δ_{JND}.



(a) AVC (b) AVC (c) VVC (d) VVC

Fig. 10: Comparison between the JND threshold estimation by the flicker test and the plain test. The boxplots in (a) and (c) show the summary statistics of the JND thresholds for the flicker test and the plain test, respectively. The bars in (b) and (d) show, for each of the 45 source videos, the difference Δ_{JND} between the JND threshold estimated with the flicker test and the one estimated with the plain test. A negative difference means that the flicker test is more sensitive.

E. Time complexity

The flicker test in crowdsourcing required more time than the plain test. The flickering videos have about the same bitrate as the source video. For the plain test, compressed videos are transmitted, which requires much less download time (Sec. IV-C). However, the waiting time for the complete upload of all videos required in a task with flicker tests was less than twice as large as with plain tests.

The response time is the duration from the start of the display of the stimuli to the time when the participant pressed one of the “left”, “right”, or “not sure” buttons. In our experiment, the response time for paired comparisons with the flicker test was slightly longer than that for the plain test, $5.0\,\text{s}$ for the flicker test vs. $4.7\,\text{s}$ for the plain test. However, this difference is very small and hardly relevant for subjective experiments.

Fig. 13 compares the cumulative distribution of the participants’ responses time for the flicker test and the plain test. For more than half of the comparisons, the response time was greater than $5\,\text{s}$, which is the duration of the videos. In these cases, participants viewed entire paired videos before making their decision within the $3\,\text{s}$ window following the video playback.

The cumulative response time curves for both tests sharply increase in steepness about 500 ms after the end of the video playback. This corresponds well to the expected reaction time required for a participant to press one of the buttons when the video playback is finished.

VII. Conclusion

When compressing video sequences to smaller bitrates, the probability that an observer cannot see any distortion in the reconstructed video is of interest in many applications. This paper presents improvements in methods for the estimation of these so-called satisfied user ratios. We showed that the common procedure of fitting a distribution model to JND thresholds suffers from a bias. This bias is removed by our proposed collective observer method, in which a randomly selected observer responds to each comparison of a compressed video with its

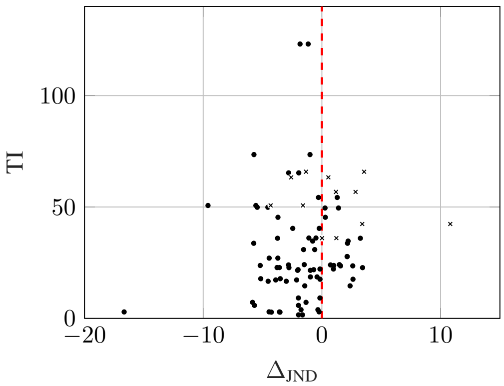
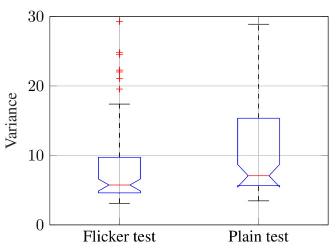
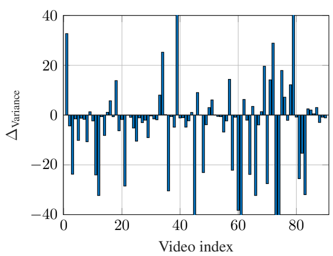


Fig. 11: Temporal information (TI) vs. Δ_{JND} for 45 source videos encoded with AVC and VVC. The videos marked with crosses exhibit high motion, as determined by visual inspection conducted by the authors.



(a) AVC/ VVC



(b) AVC/ VVC

Fig. 12: Comparison between the variances of the JND distributions estimated with the flicker test and the plain test. The boxplots in (a) show the summary statistics. The bars in (b) show the difference between the variances estimated with the flicker test and the plain test. Video indices 1 to 45 are for AVC and 46 to 90 for VVC compressed sources. A negative difference means that the flicker test was more precise.

source. This is also more efficient than estimating and averaging the psychometric functions of many individual observers. For the purpose of estimating the collective psychometric function, we applied an adaptive psychometric Bayes method, QUEST+, in a crowdsourcing environment.

For our experimental work, we adapted the flicker test for paired video comparisons, which had been developed for evaluating near-lossless image coding. We implemented a web-based user interface using a within-subject study design for evaluating video quality under the two test conditions, i.e., the flicker test and the plain side-by-side comparison. This web interface and the crowdsourcing environment were governed by our server application that ran multiple parallel instances of QUEST+ to adaptively determine the stimuli delivered in each task for each study participant. We estimated the SUR curves of 45 source video sequences encoded with AVC and VVC at all available QP values. The results show that the flicker test increased the sensitivity and precision of the JND-based video quality assessment compared to plain side-by-side presentation. Our dataset will be made available online at the time of publication.

Our approach paves the way for larger and better quality

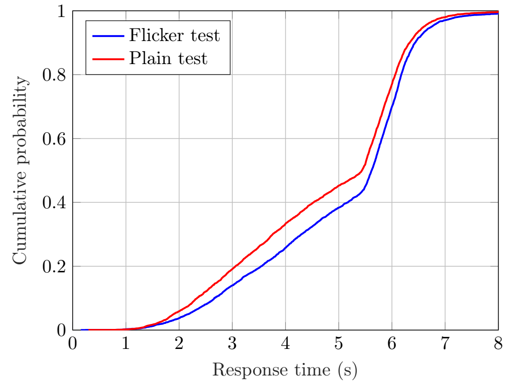


Fig. 13: Cumulative response time for paired comparisons.

datasets of JNDs in video compression and corresponding SUR curves. In our future work, we will conduct such a large-scale crowdsourcing campaign. Furthermore, in a laboratory study, we will evaluate the JND and SUR curves for high-resolution videos of 1920×1080 and 3840×2160 pixels.

References

[1] A. Aaron, Z. Li, M. Manohara, J. Y. Lin, E. C.-H. Wu, and C.-C. J. Kuo, "Challenges in cloud based ingest and encoding for high quality streaming media," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 1732–1736.

[2] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1509–1513.

[3] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 292–302, 2017.

[4] Q. Huang, H. Wang, S. C. Lim, H. Y. Kim, S. Y. Jeong, and C.-C. J. Kuo, "Measure and prediction of HEVC perceptually lossy/lossless boundary QP values," in *IEEE Data Compression Conference (DCC)*, 2017, pp. 42–51.

[5] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, "Experimental design and analysis of JND test on coded image/video," in *Applications of digital image processing XXXVIII*, vol. 9599. International Society for Optics and Photonics, 2015, p. 95990Z.

[6] F. Kingdom and N. Prins, *Psychophysics: A Practical Introduction, Second Edition*. Academic Press, 2016.

[7] A. B. Watson, "QUEST+: A general multidimensional bayesian adaptive psychometric method," *Journal of Vision*, vol. 17, no. 3, pp. 1–27, 2017.

[8] D. M. Hoffman and D. Stolitzka, "A new standard method of subjective assessment of barely visible image artifacts and a new public database," *Journal of the Society for Information Display*, vol. 22, no. 12, pp. 631–643, 2014.

[9] S. Göring, R. R. R. Rao, and A. Raake, "Quality assessment of higher resolution images and videos with remote testing," *Quality and User Experience*, vol. 8, no. 1, p. 2, 2023.

[10] R. R. R. Rao, S. Göring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2021, pp. 1–6.

[11] B. Naderi and R. Cutler, "A crowdsourcing approach to video quality assessment," *arXiv preprint arXiv:2204.06784v2*, 2022.

[12] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowd workers proven useful: A comparative study of subjective video quality assessment," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[13] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2013.

[14] H. Lin, G. Chen, M. Jenadeleh, V. Hosu, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Large-scale crowdsourced subjective assessment of picturewise just noticeable difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5859–5873, 2022.

[15] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Jun 2016. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[16] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2019.

[17] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: Full-reference video quality assessment with 3D convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4447–4451.

[18] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C. J. Kuo, "Objective video quality assessment based on perceptually weighted mean squared error," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1844–1855, 2016.

[19] Y. Zhang, H. Liu, Y. Yang, X. Fan, S. Kwong, and C. C. J. Kuo, "Deep learning based just noticeable difference and perceptual quality prediction models for compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1197–1212, 2022.

[20] X. Zhang, C. Yang, H. Wang, W. Xu, and C.-C. J. Kuo, "Satisfied-user-ratio modeling for compressed video," *IEEE Transactions on Image Processing*, vol. 29, pp. 3777–3789, 2020.

[21] J. Zhu, P. Le Callet, A.-F. Perrin, S. Sethuraman, and K. Rahul, "On the benefit of parameter-driven approaches for the modeling and the prediction of satisfied user ratio for compressed video," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 4213–4217.

[22] S. Nami, F. Pakdaman, and M. R. Hashemi, "BL-JUNIPER: A CNN-assisted framework for perceptual video coding leveraging block-level JND," *IEEE Transactions on Multimedia*, 2022, Early Access.

[23] J. Zhu and P. L. Callet, "Just noticeable difference (JND) and satisfied user ratio (SUR) prediction for compressed video: research proposal," in *Proceedings of the 13th ACM Multimedia Systems Conference*, 2022, pp. 393–397.

[24] H. Wang, I. Katsavounidis, Q. Huang, X. Zhou, and C.-C. J. Kuo, "Prediction of satisfied user ratio for compressed video," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6747–6751.

[25] D. Yuan, T. Zhao, Y. Xu, H. Xue, and L. Lin, "Visual JND: A perceptual measurement in video coding," *IEEE Access*, vol. 7, pp. 29 014–29 022, 2019.

[26] H. Wang, X. Zhang, C. Yang, and C.-C. J. Kuo, "Analysis and prediction of JND-based video quality model," in *Picture Coding Symposium (PCS)*, 2018, pp. 278–282.

[27] Y. Zhang, L. Zhu, G. Jiang, S. Kwong, and C.-C. J. Kuo, "A survey on perceptually optimized video coding," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–37, 2023.

[28] *Advanced Video Coding for Generic Audio-Visual Services*, ISO/IEC 14496-10, ISO/IEC JTC 1, May 2003.

[29] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[30] *Versatile Video Coding*, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, July 2020.

[31] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[32] L. Jin, J. Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," *Electronic Imaging*, no. 13, pp. 1–9, 2016.

[33] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "MCL-JCV: a JND-based H.264/AVC video quality assessment dataset," in *International Conference on Image Processing (ICIP)*, 2016, pp. 1509–1513.

[34] X. Liu, Z. Chen, X. Wang, J. Jiang, and S. Kowng, "JND-Pano: Database for just noticeable difference of JPEG compressed panoramic images," in *Pacific Rim Conference on Multimedia (PCM)*. Springer, 2018, pp. 458–468.

[35] C. Fan, Y. Zhang, H. Zhang, R. Hamzaoui, and Q. Jiang, "Picture-level just noticeable difference for symmetrically and asymmetrically compressed stereoscopic images: Subjective quality assessment study and datasets," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 140–151, 2019.

[36] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, and S. Kwong, "A JND dataset based on VVC compressed images," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[37] H. Lin, M. Jenadeleh, G. Chen, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Subjective assessment of global picture-wise just noticeable difference," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[38] G. A. Gescheider, *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, Inc, 1997.

[39] F. A. Wichmann and F. Jäkel, "Methods in psychophysics," *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, vol. 5, pp. 1–42, 2018.

[40] ISO/IEC 29170-2, "Information technology – Advanced image coding and evaluation – Part 2: Evaluation procedure for visually lossless coding," 2015.

[41] H. Men, H. Lin, M. Jenadeleh, and D. Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, vol. 9, pp. 138 939–138 975, 2021.

[42] S. Le Moan and M. Pedersen, "Subjective image fidelity assessment: Effect of the spatial distance between stimuli," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 445–449.

[43] B. R. Wooten, L. M. Renzi, R. Moore, and B. R. Hammond, "A practical method of measuring the human temporal contrast sensitivity function," *Biomedical Optics Express*, vol. 1, no. 1, pp. 47–58, 2010.

[44] D. F. Stolitzka, P. Schelkens, and T. Bruylants, "New procedures to evaluate visually lossless compression for display systems," in *Applications of Digital Image Processing XL*, vol. 10396. SPIE, 2017, pp. 98–108.

[45] H. Lin, V. Hosu, C. Fan, Y. Zhang, Y. Mu, R. Hamzaoui, and D. Saupe, "SUR-FeatNet: Predicting the satisfied user ratio curve for image compression with deep feature learning," *Quality and User Experience*, vol. 5, no. 1, pp. 1–23, 2020.

[46] M. Taylor and C. D. Creelman, "PEST: Efficient estimates on probability functions," *The Journal of the Acoustical Society of America*, vol. 41, no. 4A, pp. 782–787, 1967.

[47] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, "VVenC: An open and optimized VVC encoder implementation," in *Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2021, pp. 1–2.

[48] N. Barman, N. Khan, and M. G. Martini, "Analysis of spatial and temporal information variation for 10-bit and 8-bit video sequences," in *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2019, pp. 1–6.

[49] B. Münzer, K. Schoeffmann, L. Böszörményi, J. Smulders, and J. J. Jakimowicz, "Investigation of the impact of compression on the perceptional quality of laparoscopic videos," in *IEEE 27th International Symposium on Computer-Based Medical Systems*, 2014, pp. 153–158.

[50] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," *IEEE Access*, vol. 9, pp. 72 139–72 160, 2021.

[51] D. McNally, T. Bruylants, A. Willème, T. Ebrahimi, P. Schelkens, and B. Macq, "JPEG XS call for proposals subjective evaluations," in *Applications of Digital Image Processing XL*, vol. 10396. International Society for Optics and Photonics, 2017, p. 103960P.

[52] M. Jenadeleh, J. Zagermann, H. Reiterer, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Relaxed forced choice improves performance of visual quality assessment methods," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 37–42.

[53] A. T. Woods, C. Velasco, C. A. Levitan, X. Wan, and C. Spence, "Conducting perception research over the internet: a tutorial review," *PeerJ*, vol. 3, p. e1058, 2015.

[54] Q. Li, S. J. Joo, J. D. Yeatman, and K. Reinecke, "Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[55] ISO 9241-303, "Ergonomics of human-system interaction – part 303: Requirements for electronic visual displays," 2011.