Statistically Indifferent Quality Variation: An Approach for Reducing Multimedia Distribution Cost for Adaptive Video Streaming Services

Benjamin Rainer, Stefan Petscharnig, Christian Timmerer, *Senior Member, IEEE*, and Hermann Hellwagner, *Senior Member, IEEE*

***Abstract*—Forecasts predict that Internet traffic will continue to grow in the near future. A huge share of this traffic is caused by multimedia streaming. The quality of experience (QoE) of such streaming services is an important aspect and in most cases the goal is to maximize the bit rate which—in some cases—conflicts with the requirements of both consumers and providers. For example, in mobile environments users may prefer a lower bit rate to come along with their data plan. Likewise, providers aim at minimizing bandwidth usage in order to reduce costs by transmitting less data to users while maintaining a high QoE. Today's adaptive video streaming services try to serve users with the highest bit rates that consequently results in high QoE. In practice, however, some of these high bit rate representations may not differ significantly in terms of perceived video quality compared to lower bit rate representations. In this paper, we present a novel approach to determine the statistically indifferent quality variation of adjacent video representations for adaptive video streaming services by adopting standard objective quality metrics and existing QoE models. In particular, whenever the quality variation between adjacent representations is imperceptible from a statistical point of view, the representation with higher bit rate can be substituted with a lower bit rate representation. As expected, this approach results in savings with respect to bandwidth consumption while still providing a high QoE for users. The approach is evaluated subjectively with a crowdsourcing study. Additionally, we highlight the benefits of our approach, by providing a case study that extrapolates possible savings for providers.**

***Index Terms*—Adaptive video streaming, MPEG–DASH, quality of experience.**

Manuscript received May 24, 2016; revised September 9, 2016 and November 8, 2016; accepted November 14, 2016. Date of publication November 16, 2016; date of current version March 15, 2017. This work was supported in part by the Austrian Science Fund under the CHIST-ERA project CONCERT (A Context-Adaptive Content Ecosystem Under Uncertainty), project number I1402, and in part by the Austrian Research Promotion Agency under the project Advanced Ultra High Definition Dymamic Adaptive Streaming over HTTP. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaoqing Zhu.

B. Rainer, S. Petscharnig, and H. Hellwagner are with the Institute of Information Technology, Alpen-Adria-Universität Klagenfurt, Klagenfurt 9020, Austria (e-mail: benjamin.ainer@itec.aau.at; stefan.petscharnig@itec.aau.at; hermann.hellwagner@itec.aau.at).

C. Timmerer is with the Institute of Information Technology, Alpen-Adria-Universität Klagenfurt, Klagenfurt 9020, Austria, and also with Bitmovin Inc., Palo Alto, CA 94301 USA (e-mail: christian.timmerer@bitmovin.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2629761

I. Introduction

REAL–TIME streaming of audio and video nowadays represents a huge share of the Internet traffic for both fixed and mobile networks. The percentage of traffic caused by audio/video services exceeds already 70% of total traffic during peak (evening) hours [1]. Predictions for the future assume a further increase in multimedia traffic [2].

The vast amounts of the Internet traffic in combination with the aforementioned peaks require tremendous capacities for the delivery of the multimedia content. Content Distribution Networks (CDNs) (e.g., Akamai, Amazon CloudFront, Fastly) provide the network infrastructure for delivering large amounts of multimedia content towards the end users. However, these CDN services are considered as costly and, thus, each company offering such multimedia streaming services aims at minimizing these CDN costs.

Almost all of these multimedia streaming services adopt the principle of adaptive streaming (over HTTP) where a continuous multimedia stream/file is split into constant time slices referred to as segments. Additionally, multimedia content is provided in multiple versions (e.g, different bit rates, resolutions, etc.) and these versions are referred to as representations. The segments are downloaded by clients in a pull–based manner and the complexity is moved towards the clients, reversing the traditional push–based approaches using RTP/RTCP. This allows a simple server architecture and delegates the decision of choosing an appropriate representation to the clients. In this context, various proprietary solutions have been proposed in the past and with the ratification of ISO/IEC 23009-1 also known as MPEG Dynamic Adaptive Streaming over HTTP (DASH) an interoperable solution is available [3]. In this paper, we utilize the MPEG–DASH terminology and its formats but please note that the approach is also applicable for other formats adopting the same principles of adaptive video streaming (over HTTP), e.g., Apple HTTP Live Streaming.

The hypothesis for this work is as follows: for a given set of representations within some time periods (i.e., one or more segments) there exist representations encoded at different bit rates which are (almost) visually indistinguishable. In other words, we presume that the perceived visual quality of higher bit rate representations do not always significantly differ from those of lower bit rate representations considering a statistical point of view. Therefore, we conclude that by selecting lower bit rate representations with negligible differences in perceived visual

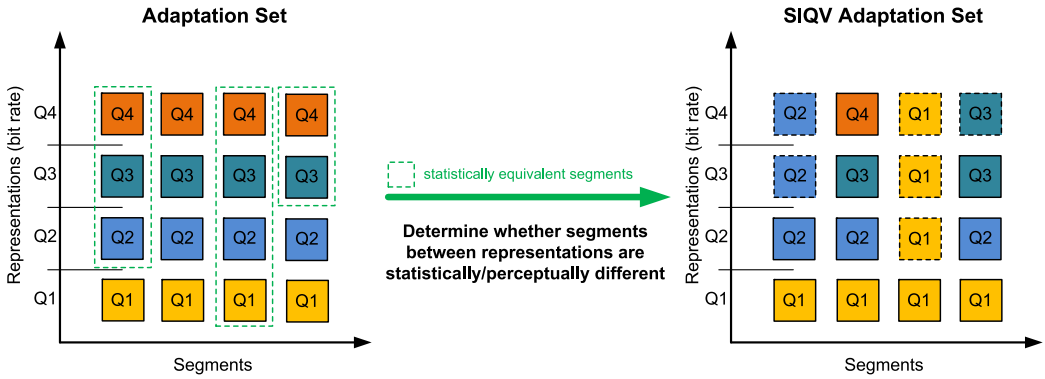

Fig. 1. Idea and ultimate goal of the SIQV approach. Colors indicate the different representations and their qualities.

quality, large amounts of network traffic can be avoided while maintaining a high Quality of Experience (QoE) for the end user. For example, beginnings of movies often consist of many black frames which may look alike in medium and high bit rate encodings. In this case, the QoE for the end user would not decrease by showing the presumably *lower quality* version, which in fact is indistinguishable for the human visual system (HVS). The HVS further motivates this hypothesis, because of its following properties [4]:

1) the HVS does not notice every little change in a video, which is indeed noticed by a signal fidelity metric like PSNR or SSIM;
2) not every image region is equally important (thus, not every region receives the same visual attention);
3) some changes may lead to an enhancement of a video sequence (for example edge sharpening).

With the above considerations, we formulate the research question addressed in this paper as follows: *How can we reduce the network traffic and consequently CDN costs while maintaining the QoE?*

In this paper, we propose a method to address this problem for adaptive video streaming over HTTP utilizing MPEG–DASH formats. First, we calculate objective video quality metrics for each representation and segments. Using an appropriate "objective video quality metric"–to–QoE mapping (i.e., QoE–model), a Mean Opinion Score (MOS) is predicted for every segment. The approach foresees to minimize the size of the representations by substituting segments with their lower–size versions of statistically equivalent *visual quality* (with respect to the employed objective video quality metric).

The main contribution of this work is to define whether a segment of two or more different representations is visually equivalent by determining its *statistically indifferent quality variation (SIQV)*. Fig. 1 illustrates the idea and ultimate goal of our SIQV approach. The basic idea addresses the question whether MOS values of different representations of a segment are statistically different. Therefore, we employ the variation in MOS (which is given by the QoE–model) and perform a Student's t–test for a specific MOS value leading to an enclosing interval. MOS values within this interval are *statistically non-significant different in QoE*. An appropriate "objective video quality metric"–to–QoE mapping (which reflects in some part the human perception) is used to determine the objective value. We do not claim that perceptual indifference is equal to statistical indifference. An objective metric can only cover the human perception of video frames to a certain extent [5]. Thus, the success of this approach relies on the employed objective video quality metric as well as on an appropriate QoE model (which depends on the objective metric used). Please note that we neither propose a new video quality metric nor a QoE model. We instead adopt existing metrics and models which can be superseded in the future. When we refer to *statistically indifferent quality variation*, we actually think of it by means of the employed QoE model (and its parameters) and the employed objective metric. In order to assess the performance and applicability of our approach, we provide a subjective quality assessment using crowdsourcing and an objective evaluation. The subjective quality assessment validates our statistic inference approach and shows that the selected quality metrics and their respective QoE models support the aforementioned assumptions. Finally, a case study is shown to demonstrate how much traffic one may save on average. We extrapolate the savings with estimated numbers from a large Video on Demand platform and CDN costs. We expect that with our approach it is possible to save noticeable amounts of network bandwidth–justifying the effort of implementing it–while maintaining the QoE (compared to using a state-of-the-art approach which aims to maximizing the representation bit rate) independent of the employed client implementation, possible network traffic shaping, or the number and characteristics of the available representations.

The remainder of the paper is structured as follows. Section II gives an overview on related work on quality-aware adaptive video streaming over HTTP approaches and optimal representation selection. Our approach to quantify the SIQV is presented in Section III. The evaluation, case study and a link to a live demonstration is provided in Section IV and the paper is concluded in Section V.

II. Related Work

Recent research on video quality focuses on Quality of Experience (QoE), defined as the user's degree of delight with an application or service [6]. With growing understanding of the QoE while consuming videos over the Internet, insights from QoE research influenced some DASH approaches which are highlighted in the following.

In order to optimize the QoE of a user, Hoßfeld *et al.* [7] introduced an evaluation framework which allows the computation

of QoE-optimal adaptation on a per-user basis. The authors provide various mixed-integer optimization problems that aim at: (i) minimizing the startup delay without stalling, (ii) maximizing the quality without stalling, (iii) minimizing the number of representation switches without stalling at a given quality, (iv) maximizing the quality for a multi-user scenario without stalling, and (v) minimizing the number of quality switches without stalling at a given target quality for a multi-user scenario. Further investigation on the impact buffer starvation and it's impact on the QoE is provided in [8]. The optimization problems provided in [7] influenced our optimization problems provided in Section III-A.

Li *et al*. [9] propose an extension to Probe and Adapt (PANDA) [10], an adaptation mechanism for DASH, that provides *consistent quality*. Dynamic programming is employed to solve the proposed constrained optimization sub-problem at every adaptation step. Based on available bandwidth, buffer size, knowledge of bandwidth requirements, and quality of the encoded video, the downloading choice is calculated following a quality policy (such as maximize minimum quality or maximize average quality) while ensuring that the buffer is full enough for a stall-free video playback.

The work that matches closest to our approach is provided in [11] and [12]. Toni *et al*. provide an Integer Linear Program (ILP) for determining the optimal selection of representations for adaptive video streaming with respect to a user satisfaction metric. Their approach is based on an ILP maximizing average user satisfaction (incorporating a video quality metric) while considering network dynamics, type of video content, and user population characteristics. The results comprise guidelines for the number of representations per video type, allocation of representations across resolutions, allocation of available representations across encoding rate at a given resolution, as well as saving CDN bandwidth while preserving user satisfaction. In contrast, our approach does not aim at providing guidelines or providing an optimized set of representations. Our approach takes as input an already existing set of representations. Thus, the work provided in [11], [12] can be seen as a step before our approach is applied.

For objectively estimating QoE, relevant literature often makes use of PSNR like in [13], [14]. As in the last decade the quality of displays has drastically changed (e.g., resolutions up to 3840×2160, better color space coverage, and size), SSIM has been introduced [15], which exploits the structure of images. It has emerged as a second relevant objective metric for image quality. In order to determine the quality of a video, a *QoE model* $y = f(x; \boldsymbol{\beta})$ (henceforth vectors are indicated by bold symbols) is a mapping from x, an objective video quality, to a MOS value reflecting subjective video quality y. The model in conjunction shall reflect the *human visual perception* by providing the mapping between objective and subjective metrics. The model is estimated based on subjectively evaluated mean opinion scores using parameters $\boldsymbol{\beta}$.

In the literature, there exist a variety of models and for this work, we adopt the following two models [14], [15]:

$$f(x; \boldsymbol{\beta}) = 1 - \frac{1}{1 + exp(\beta_1 \cdot (x - \beta_2))} \quad \boldsymbol{\beta} \in \mathbb{R}^2 \qquad (1)$$

TABLE I
TERMINOLOGY FOR THE OPTIMIZATION PROBLEM

N	number of segments for the video content
$r_{\max}$	maximum number of representations
$c_{i,j} \in \{0, 1\}$	decision variable whether segment i is selected in representation j
q_i^*	best (optimal) quality chosen for segment i
ε_i	statistically indifferent quality variation for segment i, where $\varepsilon_i \in [0, q_i^*]$.
$V(i)$	mapping that provides the cumulative available download capacity until the scheduled PTS of segment i in kilobytes
$size(i, j)$	mapping that maps segment i of representation j to its size in kilobytes
$quality(i, j)$	mapping that maps segment i of representation j to its associated quality

$$g(x; \boldsymbol{\gamma}) = \gamma_1 - \frac{1}{exp(\gamma_2 \cdot (x - \gamma_3))} \quad \boldsymbol{\gamma} \in \mathbb{R}^3. \qquad (2)$$

In particular, (1) is used for PSNR as suggested in [14] and (2) is used for SSIM according to [15].

III. Quantifying the Statistically Indifferent Quality Variation

This section describes our approach for quantifying the statistically indifferent quality variation. Therefore, we introduce simplifications to the adaptive video streaming use case, enabling us to formulate an optimization problem providing a theoretical upper bound on the achieved quality of such streaming systems. The upper bounds on quality allow to precisely formulate the problem behind our research question. In particular, we formulate a general optimization problem modeling minimization of consumed bandwidth while providing videos with statistically indifferent quality variation from the theoretically derived upper quality bound. The main contribution of this work uses existing QoE–models to statistically derive a maximum deviation from a specific objective quality metric (in the QoS space) which is tolerable without a significant impact on the perceived video quality (in the QoE space). We call this deviation the *statistically indifferent quality variation* (SIQV) denoted as ε. Furthermore, we propose an algorithm providing a solution to the aforementioned optimization problem by exploiting the SIQV.

A. Problem Statement

In order to formulate a general optimization problem that provides an upper bound for the quality, we introduce the following assumptions and simplifications. We assume that the playback timestamp of segments coincides with the (latest) deadline for downloading them. It must be noted that we neither aim at modeling channel characteristics nor at estimating the channel's future behavior. Instead, we assume that the behavior of the channel is known in advance (see the definition of $V(\cdot)$ in Table I). We further assume that buffers are large enough such that overflows do not occur.

An overview on the terminology is given in Table I. In our use case of adaptive streaming, a video is split into constant–time segments and encoded in different representations at encoding time. The number of segments for the video content

is denoted by $N \in \mathbb{N}_1$, the maximum number of representations by $r_{\max} \in \mathbb{N}_1$. The decision variable $c_{i,j} \in \{0,1\}$ denotes whether segment i in representation j is selected. Let $quality : \mathbb{N}_1 \times \mathbb{N}_1 \rightarrow [Q_{\min}, Q_{\max}]$ be the mapping that assigns segment i of representation j its associated quality. Please note that we do not explicitly define how quality is measured. We assume the value for quality to be bounded by $Q_{\min}$ and $Q_{\max}$ (given by an already defined video quality measure). $q_i^* \in [Q_{\min}, Q_{\max}]$ denotes the best (optimal) quality chosen for segment i with respect to the available download capacity. The parameter ε_i denotes the statistically indifferent quality variation for segment i, where $\varepsilon_i \in [0, q_i^*]$. ε_i has to be chosen such that the quality loss has no impact on the resulting QoE. Our goal and the main contribution of this paper is to find a suitable approximation of ε_i which is addressed in Section III-B.

Let $V : \mathbb{N}_1 \rightarrow \mathbb{R}_0^+$ be the mapping that provides the cumulative available download capacity until the scheduled playback timestamp (PTS) of segment i in kilobytes. Furthermore, let $size : \mathbb{N}_1 \times \mathbb{N}_1 \rightarrow \mathbb{R}_0^+$ be the mapping that assigns segment i of representation j its size in kilobytes.

Before we investigate how bandwidth consumption can be minimized while maintaining QoE, we have to determine the maximum available quality $q^* \in [Q_{\min}, Q_{\max}]^N$. In order to determine q^*, we introduce the general optimization problem for maximizing the delivered quality for a multimedia streaming session as follows:

$$\underset{(quality(1,\cdot),\ldots,quality(N,\cdot))}{\arg\max} \sum_{i=1}^{N} \sum_{j=1}^{r_{\max}} quality(i,j) \cdot c_{i,j} \tag{3}$$

subject to

$$\forall i \in \{1,\ldots,N\} : \sum_{j=1}^{r_{\max}} c_{i,j} = 1 \tag{3a}$$

$$\forall k \in \{1,\ldots,N\} : \sum_{i=1}^{k} \sum_{j=1}^{r_{\max}} size(i,j) \cdot c_{i,j} \leq V(k). \tag{3b}$$

The objective function in (3) aims at maximizing the quality for a multimedia streaming session. The restrictions denoted in (3a) ensure that exactly one representation must be chosen for each segment. Equation (3b) ensure that we do not exceed the available bandwidth when downloading the segments.

Using this upper bound, we then formulate our research question as the problem of minimizing the utilized bandwidth while maintaining an statistically indifferent quality degradation as follows:

$$\min \sum_{i=1}^{N} \sum_{j=1}^{r_{\max}} [size(i,j) \cdot c_{i,j}] \tag{4}$$

subject to

$$\forall i \in \{1,\ldots,N\} : \sum_{j=1}^{r_{\max}} c_{i,j} = 1 \tag{4a}$$

$$\forall i \in \{1,\ldots,N\} : q_i^* - \sum_{j=1}^{r_{\max}} quality(i,j)\ c_{i,j} \leq \boldsymbol{\varepsilon_i}. \tag{4b}$$

Equation (4) denotes the objective function which shall be minimized for a specific multimedia stream. The first N constraints presented in (4a) ensure that a segment can be selected in a single representation only. The following N constraints in (4b) ensure that the *quality* of the resulting segments only differs at most by ε from the theoretical optimum q^*.

The main focus and contribution of this work is how to determine the SIQV ε_i, whereas how to define $V(\cdot)$ in detail is out of scope. One may obtain $V(\cdot)$ by estimating dynamics of the transmission channel by Markov theory or by estimating its distribution over time.

B. Estimating the SIQV

The general idea is to estimate ε which is referred to as the amount of the maximum quality variation that my not be perceived by the user. Hence, the following question arises: *How much deviation from a specific QoS-metric value q may occur until it may be perceived by the user?*

As starting point for providing an estimate of the maximum statistically indifferent quality variation (the quality variation that may not be perceived by the user) for a specific video region, we consider a QoE model $f(x, \boldsymbol{\beta}) : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ (henceforth denoted as *the model*). This model shall provide MOS values for a specific objective quality measure of a certain video region (in our case a few seconds), which is the case with the two very well-known model introduced in Section II [cf. (1) and (2)]. As discussed above these models have proven useful for estimating the MOS (in terms of QoE) from the objective quality measures PSNR or SSIM, respectively. These models are estimated from data gathered by conducting user studies. MOS is in general defined as $MOS = \mu_O = \frac{1}{n}\sum_{i=1}^{n} o_i$, where o_i are the individual samples for a specific test-case (in our case the video region with a specific encoding and bit-rate having a certain (average) objective quality value).

We assume that the corresponding sample variables O_i are independent and identically distributed (i.i.d.) normal variables $O_i \sim \mathcal{N}(\mu_O, \sigma^2)$ with unknown variance σ^2 and $E(O_i) = \mu_O$. We will later on show that the normality assumption can be fulfilled if the user study is designed accordingly. If we want to know whether two sample means μ_X and μ_Y (i.e., two different test case, the same video region with different bit-rates), with the sample variables $X_i \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y_j \sim \mathcal{N}(\mu_Y, \sigma^2)$, are drawn from different populations and, therefore, are *significantly* different provided an significant threshold α, we apply an student's t-test (if and only if the aforementioned assumptions hold). Thus, we can use the student's t-test and a suitable model (which provides the estimated QoE values/MOSs for specific values of an objective measure) for providing an educated guess whether two (average) objective values for a specific video region result in *signficant*[1] different MOSs or not. In other terms, our goal is to find an $|\varepsilon| \geq 0$ such that $q \pm |\varepsilon|$ (or $\mu_X \pm \varepsilon$) is not significantly different from μ_X. As unbiased estimator for the unknown variance, we use the corrected sample variance $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)^2$.

Assuming a null-hypothesis $H_0 : \mu_X = \mu_Y$, we say that the difference of the two sample means follow a t–distribution with

[1]Significant always refers to a certain significance threshold or error one is willing to accept.

$t_D(1-\frac{\alpha}{2}, n+m-2)$ (where n and m denote the degrees of freedom). H_0 is rejected if the absolute value of the test statistics T is greater than $t_D(1-\frac{\alpha}{2}, n+m-2)$.

The test statistics is given by

$$T = \frac{\mu_X - \mu_Y}{s_{X,Y}\sqrt{\frac{1}{n}+\frac{1}{m}}}. \tag{5}$$

$s_{X,Y}$ denotes the weighted average of the two sample variances of the two populations X and Y [cf. (6)].

$$s_{X,Y}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \tag{6}$$

In order to answer the above stated question, we now make use of our assumptions and set $\mu_X = f(q, \boldsymbol{\beta})$, $\mu_Y = \mu_X - \varepsilon_Q$, $s_X = s_Y = s$, and $n = m$. ε_Q denotes the amount of the maximum statistically indifferent quality variation in the QoE space of the model. Plugging in (5) and (6) results in

$$T = \frac{\varepsilon_Q}{s\sqrt{\frac{2}{n}}}. \tag{7}$$

With respect to H_0, $|T| \leq t_D(1-\frac{\alpha}{2}, 2(n-1))$ states how large $|T|$ should be such that we have no significant difference (with respect to our assumptions) between the *two populations*.

$$|\varepsilon_Q| \leq t_D(1-\frac{\alpha}{2}, 2(n-1)) \cdot s \cdot \sqrt{2} \cdot n^{-\frac{1}{2}} \tag{8}$$

(8) denotes the maximum quality variation in the QoE–space of the model. It depends on the number of observations from which the model parameters were estimated, and the standard deviation the model covers (for specific values, relative to the ground truth/data). Models based on many observations bear little amount of uncertainty compared to models based on few observations. Furthermore, some models may be inappropriate, this can be identified by the R^2 (which is $\frac{\text{explained variation}}{\text{total variation}}$ or also known as the coefficient of determination). The higher the R^2 the better does the model reflect the observed data. This does also hold for the introduced approach. If a model with a low coefficient of determination is used in conjunction with our approach, it will not provide useful information.

The amount of quality loss is constant in the QoE–space for a given model. Especially, it is not affected by the observed quality $f(q, \boldsymbol{\beta})$. Hence, in the QoE–space the statistically indifferent quality variation ε_Q lies within the interval $[-t_D(1-\frac{\alpha}{2}, 2(n-1)) \cdot s \cdot \sqrt{2} \cdot n^{-\frac{1}{2}}, t_D(1-\frac{\alpha}{2}, 2(n-1)) \cdot s \cdot \sqrt{2} \cdot n^{-\frac{1}{2}}]$. Consequently, the value for quality variation ε must be within the interval $f^{-1}([f(q, \boldsymbol{\beta}) - t_D(1-\frac{\alpha}{2}, 2(n-1)) \cdot s \cdot \sqrt{2} \cdot n^{-\frac{1}{2}}, f(q, \boldsymbol{\beta}) + t_D(1-\frac{\alpha}{2}, 2(n-1)) \cdot s \cdot \sqrt{2} \cdot n^{-\frac{1}{2}}])$. The expression on the right hand side of (8) can be easily evaluated given a certain significance threshold α. Typically, one selects $\alpha \in \{0.05, 0.1\}$ (which are the so-to-say standard significance levels in statistics, the interested reader is refered to [16]). Nevertheless, this results in additional restrictions on the model which are continuity and at least local invert-ability within the neighborhood of q. With this transformation from the QoE space back to the QoS space, the non–linearity of the model results in intervals for which the possible quality variations may have different sizes. For example, Fig. 2 depicts quality variation intervals for different values of objective quality using a PSNR–based QoE model. Let's assume that our calculations result in an interval in the MOS space of length ε_Q. For different quality values Q_1 and Q_2, the interval in the MOS space is of constant size. Altering the MOS results in an interval in the PSNR space. As the model is steeper in the middle, Q_1 with a medium–low quality has a smaller resulting interval ε_1 than Q_2, where the model levels out. Hence, for this model, intervals for very high and very low qualities are bigger than intervals in the middle of the QoS range. This matches with the intuition that small changes in the middle of the QoS range have a big influence on the QoE and the variation of these must be handled carefully. Please also note that the intervals in the PSNR space are not symmetrical.

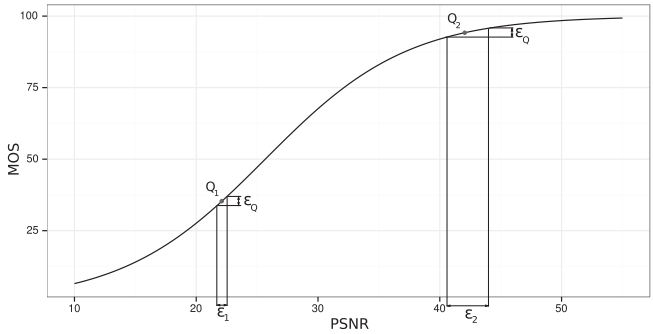

Fig. 2. Intervals depending on PSNR values in the PSNR–based QoE model.

The presented approach relies strongly on the validity of the model, mapping an objective quality measure to the QoE. Inaccurate model assumptions may lead to significant differences in the QoE domain. While the *statistically indifferent quality variation* seems to be linear in the QoE domain, it has to be noted that the interval does rely on the underlying model and is only valid with the used model and its estimated parameters. The model and the objective metric provide the connecting part between subjective measure and the human visual perception (as already discussed in Section II) or at least the part of the human visual perception that can be covered by the objective measure.

C. Example

In this section we show an example for a practical application of the SIQV. Therefore, we use the QoE–model $f(x, \boldsymbol{\beta}) = 100 - \frac{100}{1+e^{\beta_1 * (x-\beta_2)}}$, where x denotes a PSNR value, $\beta_1 = 0.1701$ and $\beta_2 = 25.6675$. The model has been derived from a sample of $n = 15$ Mean Opinion Scores. The actual standard deviation of the samples in the model is $s_X = 16$. Lets assume that the adaptation logic decides for a representation (which in our terminology translates to the representation with optimal quality q^*) with a PSNR value of 50 dB. Translated to the QoE–space ranging from 0 to 100 this is equivalent to a subjective score of $\mu_X = f(q^*; \boldsymbol{\beta}) = 98.431$. With these given values, we can compute our interval using (8). The resulting interval for ε_Q is $[-1.4236, 1.4236]$. Quality variations of video sequences within this interval are (statistically) visually imperceptible. This interval is independent from the considered quality (so it is constant within the used

Algorithm 1: SIQV Approach

Given:
- $r_{\max}$ ▷ Number of different representations
- n ▷ Number of segments per representation
- f, f^{-1} ▷ QoE model and its inverse
- $quality, size$ ▷ Mappings of segments
- n, s^2 ▷ Model parameters
- α ▷ Confidence level

Output:
- S ▷ The resulting substitution set

$S \leftarrow \emptyset$
$\varepsilon_Q \leftarrow t_D(1-\frac{\alpha}{2}, 2(n-1))\ \sqrt{2} \cdot s^2 \cdot n^{-\frac{1}{2}}$
for $i \in \{1 \dots n\}$ **do**
  $S_i \leftarrow \emptyset$
  for $j \in \{1 \dots r_{\max}\}$ **do**
    $tmp \leftarrow j$
    $\varepsilon \leftarrow f^{-1}(f(quality(i,j)) - \varepsilon_Q)$
    for $k \in \{1 \dots r_{\max}\}$ **do**
      if $quality(i,k) > \varepsilon$ **then**
        if $size(i,k) < size(i,tmp)$ **then**
          $tmp \leftarrow k$
    $S_i \leftarrow S_i \cup \{((i,j),(i,k))\}$
  $S \leftarrow S \cup \{S_i\}$

model) and should now be mapped to the actual quantitative quality measure using $f^{-1}([\mu_X - 1.4236, \mu_X + 1.4236]; \beta) = [46.118, 64.069]$. This yields the actual interval for varying the quantitative quality measure (in this example PSNR). If we want to save bandwidth, we obviously want to select the lower end of the interval allowing us to pick a representation with a quantitative measure of 46.118 dB. If such a representation with a quantitative measure in $[46.118, 50)$ exists, it is likely that the segments of the newly selected representation will require less bit rate.

D. QoE–Aware Selection of Representations

Based on the SIQV, we are able to formulate Algorithm 1 enabling a QoE–aware substitution of representations without significant impact on the QoE. Given information about segment quality and size, the algorithm determines whether choosing a higher representation over a lower representation is beneficial with respect to bandwidth savings. In particular, when the quality variation between a representation and its next higher representation is not significant according to our approach and the employed model, then the lower bit rate representation is used instead. The resulting representations thus are equal in perceived quality (compared to the original representation set), but smaller in size. The parameters of the algorithm are presented in the following. Analogous to the terminology of the optimization problem presented in Section III-A, n denotes the number of segments per representation, $r_{\max}$ denotes the number of representations. Let $quality(i,j)$ be the mapping of segment i in representation j to its associated quality and $size(i,j)$ the mapping of segment i in representation j to its associated size. The model function f is used to transform the objectively measured quality into the QoE space. As already mentioned in Section III-B, this transformation bears an uncertainty in the mapping to a QoE value. Rather than considering the QoE value retrieved by the model, one can say that the true QoE value lies in an interval around this value with confidence level $1-\alpha$. This interval depends on the model fit, i.e., the number of observations n and the variance of the model s^2. The inverse model function f^{-1} is used to transform the interval from the QoE space back to an interval in the objective video quality measure space. The resulting interval is then the statistically indifferent quality variation. Whereas parameters f, f^{-1}, n, and s^2 are given by the model, the parameter α may be chosen freely from the interval $[0,1]$. Please note that for bigger values of α, the interval for SIQV increases. Thus, the approach may result in noticeable quality losses. In order to keep the quality deviation from the original representation (statistically) imperceptible, we recommend setting the statistical significance level $\alpha = 0.05$.

Algorithm 1 implements the selection of representations for a given set of representations per resolution. Initially, the algorithm calculates the maximum accepted deviation in the QoE space according to (8). The allowed deviation in subjective quality does not depend on a specific segment, it solely depends on the parameters of the used model and parameter α. We furthermore characterize an individual segment (i,j) by its time slot i and representation j. The algorithm chooses for each segment (i,j) a substitution (i,k), with $quality(i,k) \geq \varepsilon$ and minimal size. Thus, the result is a set S of substitution rules for each segment and representation. The substitution rules are of the form $((i,j),(i,k))$, which means that in the new representation set, representation j may be replaced by representation k for time slot i without significant quality loss.

By executing the algorithm for every resolution (and application of the substitution rules), this algorithm transforms the whole representation set into a new quality–aware set of representations with smaller size compared to the original set. Please note that we do not consider substitution of segments with different resolutions, since the used model does not take the resolution into account. Furthermore, our approach does not affect the client's adaptation logic and, thus, is orthogonal to any adaptation mechanism.

IV. Evaluation

We evaluate our approach subjectively by conducting subjective quality assessments (SQA) using crowdsourcing [17], [18]. The QoE model is derived from an existing dataset containing 1080p video sequences and their MOS values which is used for applying Algorithm 1. Finally, we perform a case study to investigate potential cost savings for a service provider.

A. QoE Model Derivation

In order to apply our approach on video sequences for conducting a subjective quality assessment using crowdsouring, we need an appropiate QoE–model. Therefore, we pick up an existing dataset and try to estimate the model parameters for the models provided by (1) and (2).

The dataset provided in [19] fits our purpose very well because it comprises short video sequences (9 to 12 seconds). The entire dataset consists of 168 video sequences including a wide range of genres encoded at 1080i50 using different bitrates with H.264/AVC. The video sequences were subjectively assessed

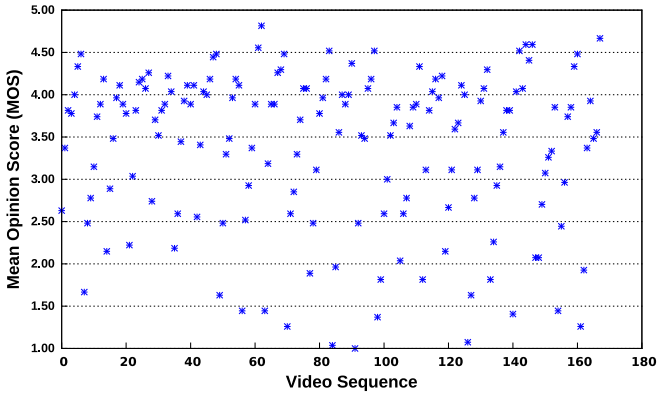

Fig. 3. MOS for all video sequences. The video sequences cover the entire range of the quality scale used in the SQA.

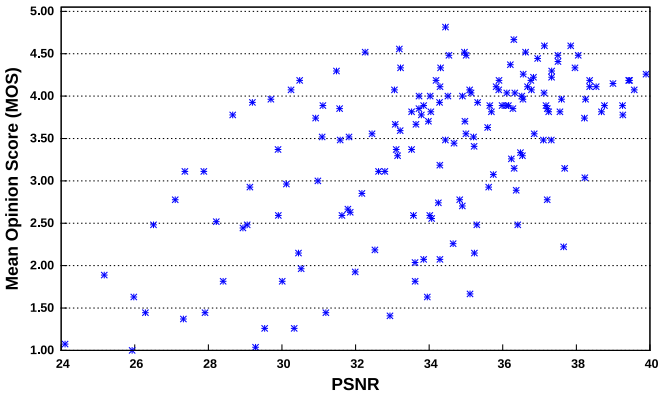

Fig. 4. MOS with respect to PSNR. High PSNR values lead to a low MOS and vice versa.

using the Absolute Category Rating with Hidden Reference (ACR-HR) as recommended by the ITU-T [20]. The ratings for the video sequences were assessed using a five–level scale ranging from 1 (worst) to 5 (best). The obtained scores are normally distributed and, therefore, our assumption in Section III-B hold. In addition to the provided MOS, we calculated the frame–by–frame PSNR and SSIM for each of the video sequences. The duration of the video sequences do fit quite well to usual segment sizes of DASH (especially for bigger segment sizes, e.g., 6s, 8s, and 10s).

In [14] the model given in (1) was estimated from responses from different datasets comprising video sequences with a resolution ranging from QCIF to 4CIF. In [13] both models were estimated from responses gathered by a subjective quality assessment using a Single Stimulus with Continuous Quality Scale with video sequences having a resolution of 768×480. Fig. 3 depicts the QoE range covered by the video sequences from [19]. It indicates that the video sequences do provide a high diversity in quality and that they cover the entire range of the used quality scale.

In order to estimate the parameters of the two models for the MOS–annotated dataset, we minimize the least square estimator using the conjugate gradient method. Fig. 4 depicts the MOS with respect to the PSNR. In the PSNR case, obtaining a statistical significant model was not possible using (1). A thorough

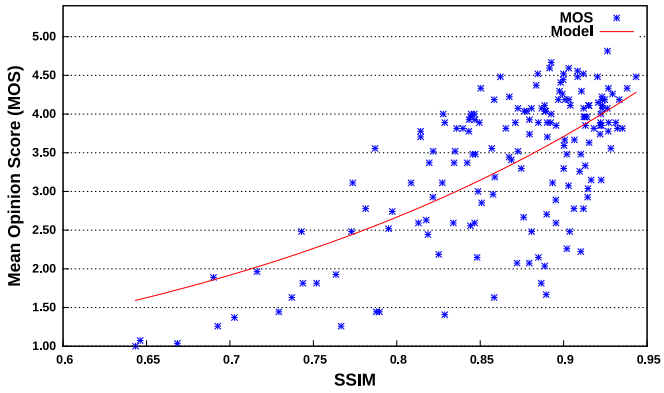

Fig. 5. MOS with respect to SSIM. The red line denotes the SSIM–based QoE model [cf. (2)].

investigation on the quality of the estimate did reveal that the parameters are not significantly different from zero, according to a student's t-test ($\beta_1 : t = 0.049$ $p = 0.96$ and $\beta_2 : t = -0.012$ $p = 0.99$). Thus, it is evident that in this case PSNR is not suited to objectively measure the quality of the video sequences. This contradicts the findings in [13] and [14]. We believe that this is due to the wide range of resolutions (QCIF to 4CIF and 768×480 vs 1920×1080) and the wide range of genres in the dataset from [19]. Fig. 5 depicts the MOS with respect to SSIM and the model with parameters $\boldsymbol{\gamma} \in \mathbb{R}^3$ [cf. (2)]. In the following, we investigate the quality of the estimate. The estimated SSIM model has a R^2 value of 0.44. A student's t-test on the hypothesis that the parameters $(\boldsymbol{\gamma_1, \gamma_2, \gamma_3})$ are zero is rejected ($\gamma_1 : t = 22.376$ $p = 5.4 \cdot 10^{-52}$, $\gamma_2 : t = 9.62$ $p = 1 \cdot 10^{-17}$, and $\gamma_3 : t = 7.16$ $p = 2.38 \cdot 10^{-11}$). These findings for SSIM coincide with the findings in [13] and [15]. Thus, we will use the obtained model for SSIM for applying our approach in the following sections. An in–depth analysis of why PSNR is not able to reflect the quality of the video sequences provided by the used dataset is out of scope of this paper.

B. Subjective Quality Evaluation Using Crowdsourcing

The goal of the SQA using crowdsourcing is to assess whether users notice the quality loss introduced by our approach.

Test content: We select two freely available videos for the SQA, *Tears of Steel* (ToS) [21] and *Sintel* [22]. Both videos are licensed under the Creative Commons Attribution. The movies were down–scaled and padded from the 3840×2140 pixel source versions to a resolution of 1280×720 pixels. In order to apply our approach introduced in Section III we encoded the movies in four different representations: 1000 kbps, 1500 kbps, 2000 kbps, and 2500 kbps. The bitrates of the representations were selected based on the bitrates YouTube provides. For encoding of the video sequences and movies, we used x264 [23] with the encoding parameters provided in Listing 1 which adopts two-pass encoding similar to [24]. The spatial and temporal information of the selected video sequences is provided by Figs. 6 and 7, respectively. The selected video sequence cover a wide range of complexity in the spatial domain as indicated by Fig. 6. Regarding the temporal information, Fig. 7 depicts that Sintel1 provides a high

```
#First Pass
x264 --pass 1 --stats ".stats" --profile baseline --preset
     slow --verbose --fps <FPS> --ref 4 --subme 5 --bitrate
     <BITRATE> --vbv-maxrate <BITRATE> --vbv-bufsize 4*<
     BITRATE> --scenecut 0 --keyint 24 --analyse none --
     output NUL <REFERENCE>
#Second Pass
x264 --pass 2 --stats ".stats" --profile baseline --preset
     slow --subme 5 --ref 4 --verbose --fps <FPS> --bitrate
     <BITRATE> --vbv-maxrate <BITRATE> --vbv-bufsize 4*<
     BITRATE> --scenecut 0 --keyint 24 --output <OUTPUT> <
     REFERENCE>
```

Listing 1. x264 encoding parameters for two pass encoding.

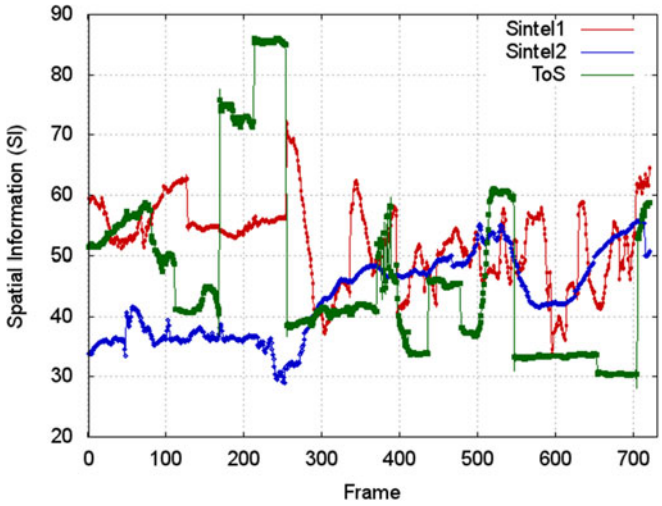

Fig. 6. Spatial information of the used test sequences for each frame.

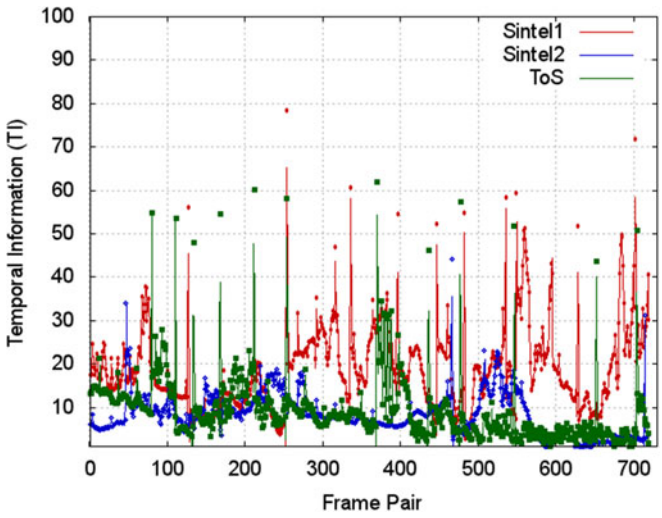

Fig. 7. Temporal information of the used test sequences for each consecutive frame pair.

complexity in the temporal domain (e.g., camera movements, object moving within the scenes), whereas ToS is quite the opposite of Sintel1, and Sintel2 provides a compromise between Sintel1 and ToS. In order to evaluate the SIQV approach, we split each video into segments of one second length. When using one second segments we discovered that the bandwidth savings are larger than with segments of two seconds or more.

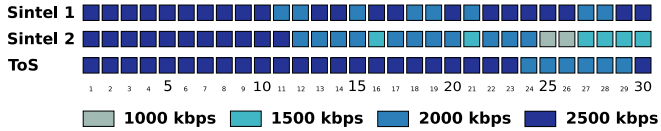

Fig. 8. Temporal series of the SIQV test sequences of the 2500 kbps representation showing which segments have been exchanged for lower quality versions. The horizontal axis denotes the segment number.

TABLE II
OVERVIEW ON EVALUATION TEST CASES

Test case	Stimulus 1	Stimulus 2
1	ToS reference high	ToS SIQV
2	ToS reference high	ToS reference low
3	Sintel 1 reference high	Sintel 1 SIQV
4	Sintel 1 reference high	Sintel 1 reference low
5	Sintel 2 reference high	Sintel 2 SIQV
6	Sintel 2 reference high	Sintel 2 reference low

Providing the possibility to save bandwidth in terms of the actual bit-rate compared to the reference [cf. Section IV-C]. For the evaluation, we selected three test video sequences with a duration of thirty seconds each. Two sequences were extracted from Sintel [22] and one was extracted from ToS [21]. We applied Algorithm 1 for all sequences resulting in a new representation set per *test sequence* assuming that the client has sufficient bandwidth for the 2500 kbps representation. The *reference sequences* comprise the same thirty seconds as the test sequences representing the highest (2500 kbps) and the lowest (1000 kbps) quality versions. We did not include audio because participants shall solely concentrate on the visual information.

Fig. 8 depicts the temporal series of the test sequences generated by our approach (i.e., the chosen representations). These test sequences are henceforth denoted by the suffix SIQV (Statistically Indifferent Quality Variation). For the selected representation of ToS, our algorithm reduces the average bitrate per second by approximately 100 kbps (which is a saving of approximately 3.8% compared to the original sequence), for Sintel 1 SIQV the savings in average bitrate per second are approximately 130 kbps (5%). The selected representations of the Sintel 2 SIQV sequence lead to a video that saves approximately 470 kbps (about 18.7%). The previous mentioned video sequences are the basis for the six different test cases presented in Table II. The first stimulus is always the reference stimulus and the second is the impaired sequence.

SQA design: The SQA follows a simple design separated into four parts. First, an introduction is presented to the participants. This introduction provides a clear and detailed description of the actual experiment. We further require the participants to agree to a disclaimer (e.g., epilepsy warning, visual impairments).

After agreement to this disclaimer, a pre–questionnaire is displayed as the second step of the SQA. Its purpose is to gather demographic information about the participants, i.e., age, gender, nationality, and country of residence. The video sequences are downloaded during the pre–questionnaire stage. This *pre–caching* step avoids bias caused by video playback stalls or long start–up times. The participant is allowed

to continue as soon as all video sequences are downloaded successfully.

The third part is the main evaluation using the Double Stimulus Impairment Scale (DSIS) methodology as recommended by the ITU-T in [20]. The DSIS method is chosen in order to be able to assess whether there exists a significant difference between the highest available representation and the transformed representations by SIQV. Therefore, the participants are shown two video sequence for each test case. In order to validate the results, we also include the case where the participants shall compare the highest and lowest quality representation of each video sequence. The test cases are provided in Table II. The reference sequence is always shown first (the participants do know the order), followed by a short pause (approximately 5 seconds) and the impaired video sequence. For the playback of the video sequences, we deactivate all user controls. This impedes possible attempts to trick within the course of the evaluation. After each pair of video sequences, there is a voting phase. The task of the participants is to categorize the perceptive visual impairment of the video quality on a five–level category scale: 1) very annoying, 2) annoying, 3) slightly annoying, 4) perceptible but not annoying, and 5) imperceptible. The voting is rendered using option boxes allowing only a single selection (initially, no option box is selected).

Within the fourth and last part of the SQA, the participants are asked to fill in a post–questionnaire. This provides the participants the possibility of giving general feedback. We further ask the participants whether they already have participated in a similar SQA.

Participants and screening: The *microworkers* crowdsourcing platform [25] was used in order to hire 321 participants from USA and Europe. The used evaluation framework [26] provides measures to assess the reliability of the participants. In particular, playback time of the video sequences, number of stalls and pauses, time spent during the voting phases, and number of browser focus changes are tracked for each test case in the main evaluation phase. Browser focus changes indicate whether the participants have payed attention during the playback of the video sequences. If the playback time deviates from the nominal playback time of the video sequences it is likely that the participant managed to pause the video playback. Furthermore, we want participants to be intuitive with their rating. Therefore, we reject those ratings for which the rating time is higher than 30 seconds. The users were screened due to invalid rating (2), browser focus change (177), rating time (30) and stalls (2).

Statistical Analysis: After filtering, 72 male and 41 female subjects remained for the statistical analysis. Their age ranges from 15 to 61 years with a median age of 31 years. 32 subjects live in the United States, the rest has residence in Europe. Fig. 9 depicts the relative occurrences of user ratings per test case. In the cases where the participants had to rate the degradation between the reference sequences and the sequences constructed according to SIQV, the majority (at least 51%) did not perceive any differences. Moreover, only about 10 percent of the users stated that they were annoyed from the perceived quality loss within the SIQV test cases. For the cases where the reference sequences and the sequences with the lowest representation were presented, 70% of the participants could identify a difference in the visual quality. Thus, the question arises whether there is a

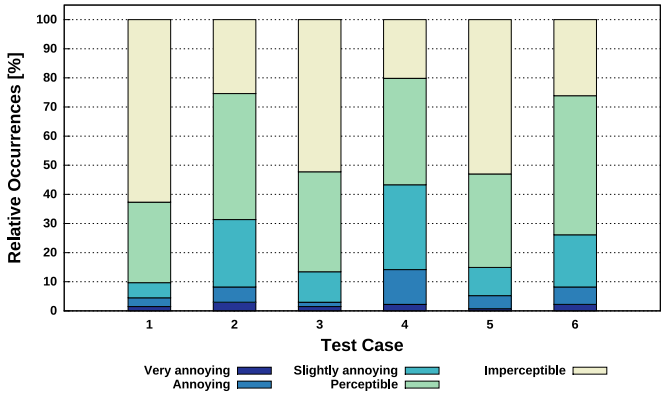

Fig. 9. Relative occurrences of the rated categories and stimuli presentations.

statistical significant difference between the test cases as well as to which category (i.e. very annoying, annoying, slightly annoying, annoying but not annoying, and imperceptible) the expected value of a test case can be assigned. For this purpose, we map these categories to the ordinal scale 1, 2, 3, 4, and 5.

As every user rated each reference–test sequence pair, the samples are not independent. Thus, for the statistical evaluation, one– and two–sample Wilcoxon tests were conducted. The one–sample Wilcoxon test is based on the null hypothesis that a given expected value equals the expected value of the sample. If rejected, there is evidence that the alternative hypotheses *less* or *greater* are appropriate. For the two–sample Wilcoxon test, the alternative hypothesis is *not equal*. The two sample Wilcoxon tests showed that there is a significant difference between ToS SIQV and ToS low (p–value $= 1.166 \cdot 10^{-7}$, V $= 2567.5$), Sintel 1 SIQV and Sintel 1 low (p–value $= 4.839 \cdot 10^{-9}$, V $= 3243.5$), and Sintel 2 SIQV and Sintel 2 low (p–value $= 7.948 \cdot 10^{-5}$, V $= 1887.5$). This confirms the supposition from the exploratory data analysis that the SIQV approach at least produces better results than the low resolution. For the test cases using the low video sequences, the expected user rating is significantly less than 4 (one–sample Wilcoxon test with alternative hypothesis *less*; ToS low: p–value $= 0.028$, V $= 650$; Sintel 1 low: p–value $= 2.2676 \cdot 10^{-6}$, V $= 522.5$, Sintel 2 low: p–value $=$ 0.028 V $= 650$). We expect the content to be rated *slightly annoying* or worse. Hence, the users recognize the quality degradation from the highest to the lowest quality. The SIQV sequences are tested against the alternative hypothesis *greater*. The tests reveal that the expected rating is significantly greater than 4 (ToS SIQV: p–value $= 2.646 \cdot 10^{-8}$, V $= 2698$; Sintel 1 SIQV: p–value $= 3.515 \cdot 10^{-5}$, V $= 1995$, Sintel 2 IQV: p–value $= 4.818 \cdot 10^{-4}$, V $= 2190$). This means that we can expect the SIQV representations to be rated category 5, *Imperceptible*. However, Fig. 9 shows that in the SIQV cases (test cases 1, 3, 5) the quality degradation was imperceptible for the majority, but for an approximately 30% of the participants the quality degradation was perceptible. Nevertheless, this perceptible quality degradation was *not annoying* which is in contrast to the other test cases not using our SIQV approach. We thus conclude that the proposed SIQV approach provides an at least non-annoying difference in video quality compared to the reference.

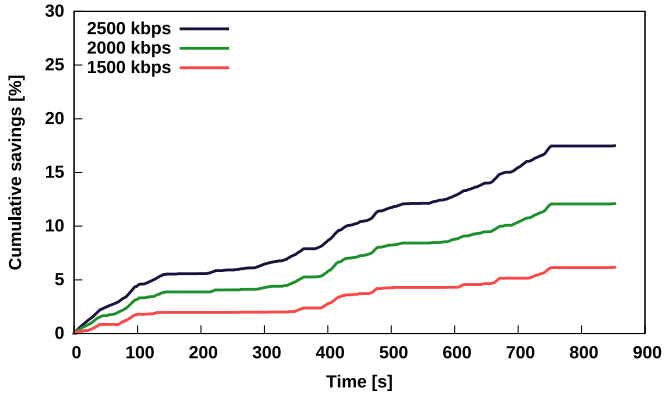

Fig. 10. Savings over time in download size relative to the total size of the highest available bit rate (1500, 2000, and 2500 kbps) for the movie Sintel in resolution 720p.

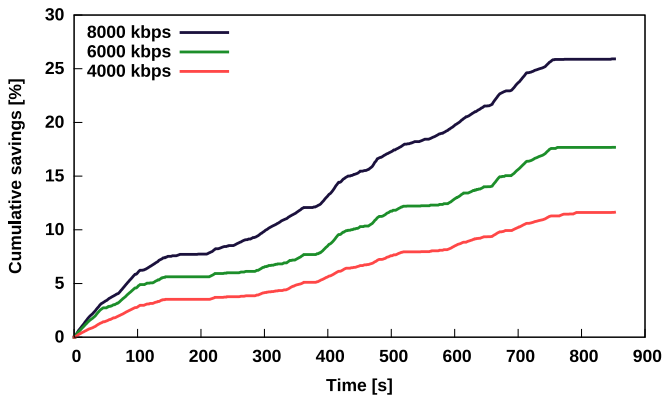

Fig. 11. Savings over time in download size relative to the total size of the highest available bit rate (4000, 6000, and 8000 kbps) for the movie Sintel in resolution 1080p.

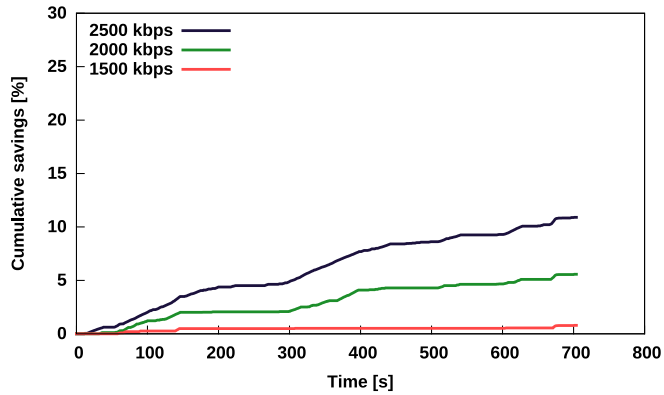

Fig. 12. Savings over time in download size relative to the total size of the highest available bit rate (1500, 2000, and 2500 kbps) for the movie ToS in resolution 720p.

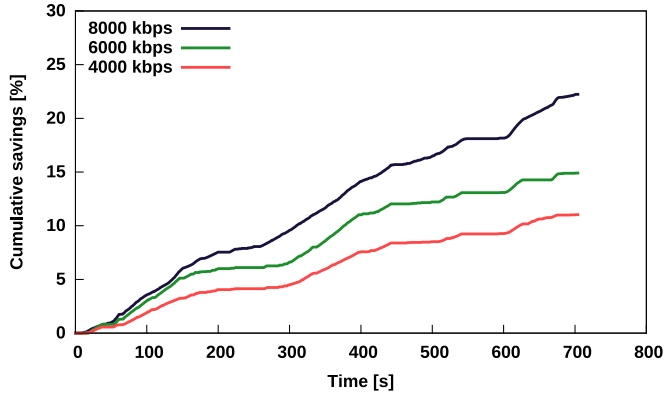

Fig. 13. Savings over time in download size relative to the total size of the highest available bit rate (4000, 6000, and 8000 kbps) for the movie ToS in resolution 1080p.

C. Case Study

According to the SQA, our approach leads to *perceptible, albeit not annoying* quality losses which in turn provide certain bandwidth savings. We now show the benefit of our approach by presenting a case study.

For the case study, we use the movie Sintel [22] and ToS [21]. We apply our approach on different segment sizes (1s, 2s, 4s, 6s, and 8s) as well as different resolutions 360p (640×360), 480p (720×480), 720p (1280×720), 1080p (1920×1080), and 1440p (2560×1440). The model obtained in Section IV-A is used with $\alpha = 0.05$. Each resolution is encoded at four bit rate levels, e.g., 1080p is encoded at 3000 kbps, 4000 kbps, 6000 kbps, and 8000 kbps. The bit rates of the representations are based on the bit rates adopted by YouTube. In the following, we focus on 720p and 1080p which are nowadays considered as the *de facto* standard. The movies have been encoded according to Section IV-B. Figs. 10 and 11 depict the cumulative savings provided by the proposed approach over the whole duration of Sintel relative to the size of the highest–quality representation. For the representation with the highest bit rate, the savings are greater than 15% for 720p and 25% for 1080p compared to the unmodified representation, respectively. Figs. 12 and 13 depict the cumulative relative savings for ToS for the resolutions 720p and 1080p. Our approach generates savings of more than 10% and 20%, respectively. Fig. 14 provides an overview of the savings for an average segment (file size) based on the representations of each resolution and for different segment sizes. As every resolution has four associated representations and we omit the SIQV calculation for the lowest resolution, we refer to the representation with the highest bit rate as *high*, the representation with the second–highest bit rate as *mid–high*, and the representation with the

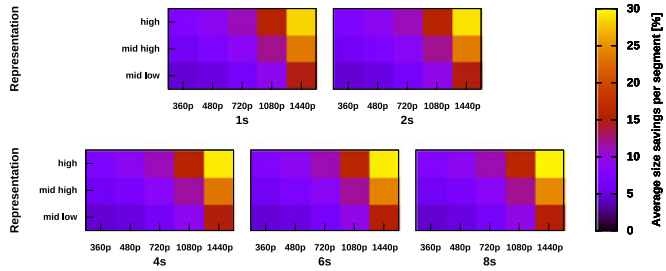

Fig. 14. Average bandwidth saving over representation for video ToS and different segment lengths.

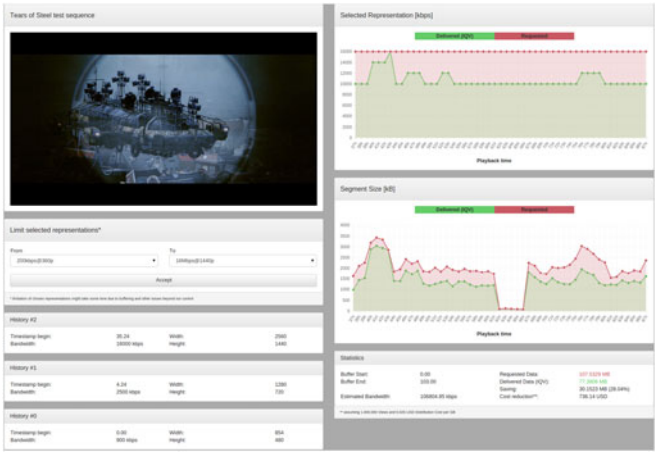

Fig. 15. Online live demo of our SIQV approach available at https://demo-itec.aau.at/livelab/siqv/demo/

third–highest bit rate as *mid–low*. The highest savings on average are obtained when the highest available resolution and bit rate are selected which is roughly 30%. Please note that these savings are given for the whole videos *Sintel* and *ToS*, in contrast to the aforementioned savings which were given for the 30-second SQA test sequences.

The last part of the case study is an estimation of potential CDN cost savings. We model the CDN cost per year C according to $C = 12 \cdot u \cdot b \cdot c$, where u denotes the number of subscribers, b denotes the average bandwidth consumption per subscriber and month, and c denotes the data delivery cost per GB in USD. As a case study we use Netflix [27]. According to [28], Netflix has 69.17 million subscribers worldwide in the third quarter of 2015. A survey on time use by the Bureau of Labor Statistics [29] reports that a person in the USA spends on average 2.8 hours on watching TV every day. If we extrapolate this to a month we get 78.4 hours. According to Netflix [27] one hour of streaming 1080p in the best quality takes three GB per hour. In [30] it is reported that the average bandwidth of an Internet access in the USA is approximately 9 Mbps (second quarter 2015). Thus, we assume that the Netflix users (at least in the USA) are capable of streaming the best quality of HD (three GB per hour). We also assume that the client adaptation logic remains unchanged, ignore multi-user scenarios competing for bandwidth of a network bottleneck and severe bandwidth fluctuations. With an estimated cost of 0.025 USD per GB [31], and an approximate yet conservative saving of 10% gained by our approach [cf. Figs. 10–13], the yearly CDN savings are approximately 488 million USD considering our previously stated assumptions.

D. Live Demonstration of the SIQV Approach

In order to show a working example of our SIQV approach, we provide an online live demonstration. The demonstration employs a Web-based DASH player (i.e., the Shaka Player provided by Google as open source) and the freely available video sequence Tears of Steel encoded in different representations with varying bit rates and resolutions, ranging from 200 kbps@360p to 14 Mbps@1440p. We applied our SIQV approach on the representation set according to Algorithm 1. Fig. 15 depicts a screenshot of the demonstration. The Web interface provides information about the selected representations and their size. It further shows which representation is currently requested by the DASH player and which is actually provided by using our SIQV approach. The demonstration can be found at: https://demo-itec.aau.at/livelab/siqv/demo/

V. Conclusion

In this paper we introduced an approach that enables a significant reduction of bandwidth consumption for adaptive video streaming services by exploiting the so–called *statistically indifferent quality variation (SIQV)* adopting existing quality metrics and models. The SIQV approach suggests that certain video representations can be substituted with a lower bit rate representation without significantly impacting the Quality of Experience. We conducted a crowdsourced subjective quality assessment to confirm our findings. The results are promising and we found that the IQV approach provides bit rate savings of up to 30% compared to original representations using two freely available video sequences (ToS and Sintel). We further presented a case study providing (under reasonable assumptions) a conservative estimation of CDN cost reductions (i.e., only 10%) and found that, for example, Netflix could save approximately 488 million USD per year.

Future work includes large-scale testing on actual deployments of adaptive video streaming services and fine-tuning of the SIQV approach. In particular, we would like to investigate whether it is worth to apply this approach at a frame level instead on a per segment basis. Another topic of interest is the experimental proof of scalability of our approach when using more sophisticated QoE models for adaptive video streaming services.

Acknowledgment

This work was performed in part in the Lakeside Laboratories research cluster at Alpen-Adria-Universität.

References

[1] Sandvine, "Global internet phenomena: Africa, Middle East & North America," *Sandvine Intelligent Broadband Networks*, 2015. [Online]. Available: https://www.sandvine.com/trends/global-internet-phenomena/

[2] Cisco, "Cisco visual networking index: Forecast and methodology, 2014–2019," Cisco Systems Inc., San Jose, CA, USA, May 27, 2015.

[3] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the internet," *IEEE Multimedia*, vol. 18, no. 4, pp. 62–67, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1109/MMUL.2011.71

[4] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320311000204

[5] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb. 2013.

[6] K. Brunnström *et al.*, "Qualinet white paper on definitions of quality of experience," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds., Lausanne, Switzerland, Ver. 1.2, March 2013.

[7] T. Hoßfeld, M. Seufert, C. Sieber, T. Zinner, and P. Tran-Gia, "Identifying QoE optimal adaptation of HTTP adaptive streaming based on subjective studies," *Comput. Netw.*, vol. 81, pp. 320–332, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128615000626

[8] Y. Xu *et al*., "Analysis of buffer starvation with application to objective QoE optimization of streaming services," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 813–827, Apr. 2014.
[9] Z. Li *et al*., "Streaming video over HTTP with consistent quality," in *Proc. 5th ACM Multimedia Syst. Conf.*, 2014, pp. 248–258.
[10] Z. Li *et al*., "Probe and adapt: Rate adaptation for HTTP video streaming at scale," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 719–733, Apr. 2014.
[11] L. Toni *et al*., "Optimal selection of adaptive streaming representations," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 2s, pp. 43:1–43:26, 2015.
[12] L. Toni, R. Aparicio-Pardo, G. Simon, A. Blanc, and P. Frossard, "Optimal set of video representations in adaptive streaming," in *Proc. 5th ACM Multimedia Syst. Conf.*, 2014, pp. 271–282.
[13] A. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 587–599, Apr. 2010.
[14] M. Pinson and S. Wolf, "Video quality measurement techniques," Nat. Telecommun. Inform. Admin., Washington, D.C., USA, Tech. Rep. TR-02-392, 2002.
[15] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[16] F. Poletiek, *Hypothesis-Testing Behaviour.* New York, NY, USA: Taylor & Francis, 2013. [Online]. Available: https://books.google.at/books?id=Z-YPbeLUCSUC
[17] T. Hoßfeld *et al*., "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.
[18] C. C. Wu, K. T. Chen, Y. C. Chang, and C. L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1121–1137, Aug. 2013.
[19] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm," in *Proc. Int. Workshop Image Media Quality Appl.*, 2008. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00300182
[20] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Int. Telecommun. Union., Geneva, Switzerland, Tech. Rep., 2008.
[21] "Tears of Steel, Mango open movie project," 2012. [Online]. Available: https://mango.blender.org/
[22] "Sintel, the Durian open movie project," 2010. [Online]. Available: https://durian.blender.org/
[23] "VideoLAN – x264," accessed on Nov. 6, 2015. [Online]. Avaialble: http://goo.gl/rH8w4J
[24] Y. C. Lin, H. Denman, and A. Kokaram, "Multipass encoding for reducing pulsing artifacts in cloud based video transcoding," in *Proc. Int. Conf. Image Process.*, Sep. 2015, pp. 907–911.
[25] "Microworkers," accessed on Jun. 2016. [Online]. Available: http://www.microworkers.com
[26] B. Rainer, M. Waltl, and C. Timmerer, "A web based subjective evaluation platform," in *Proc. 5th IEEE Int. Workshop Quality Multimedia Experience*, Jul. 2013, pp. 24–25.
[27] "Netflix," accessed on Nov. 5, 2015. [Online]. Available: http://www.netflix.com
[28] "Statista," accessed on Nov. 5, 2015. [Online]. Available: http://goo.gl/eiQ6wS
[29] "American time use survey," accessed on Nov. 5, 2015. [Online]. Available: http://goo.gl/eP8qzR
[30] "State of the internet," accessed on Nov. 5, 2015. [Online]. Available: https://goo.gl/0jtW9rf
[31] "Bizety," accessed on Nov. 5, 2015. [Online]. Available: https://goo.gl/zsmRcs

Benjamin Rainer received the B.Sc., M.Sc. (Dipl.-Ing.), and Ph.D. (Dr. techn.) degrees with distinction from the Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria, in 2010, 2012, and 2015, respectively.

He is a PostDoc Researcher with the Department of Information Technology in the Multimedia Communication (MMC) research group working in the CONCERT project, Alpen-Adria-Universität Klagenfurt. His research Interests include audio/video encoding, parallel computing, quality of experience, and ICN/NDN.

Stefan Petscharnig received the B.S. and M.S. degrees from Alpen-Adria-Universität Klagenfurt (AAU), Klagenfurt, Austria, in 2014 and 2015, respectively.

Since 2016, he has been working on the Ph.D. (Dr. techn.) degree with the KISMET research project with a focus on the analysis of endoscopic video data using machine learning at the Department of Information Technology, Distributed Multimedia Systems research group, AAU Klagenfurt. He was a Research Assistant for the AdvUHD-DASH project in 2015.

Christian Timmerer (M'08–SM'16) is an Associate Professor with Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria. He is a Cofounder of Bitmovin Inc., Palo Alto, CA, USA, as well as the CIO and the Head of Research and Standardization. He has authored or coauthored more than 150 publications. He participated in several EC-funded projects, notably, DANAE, ENTHRONE, P2P-Next, ALICANTE, SocialSensor, and the COST Action IC1003 QUALINET. He also participated in ISO/MPEG work for several years, notably, in the areas of MPEG-21, MPEG-M, MPEG-V, and MPEG-DASH. His research interests include immersive multimedia communication, streaming, adaptation, and quality of experience.

Prof. Timmerer was the General Chair of WIAMIS 2008, QoMEX 2013, and ACM MMSys 2016.

Hermann Hellwagner (S'85–A'88–M'95–SM'11) is a Full Professor of computer science with the Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria, leading the Multimedia Communications group. He has received many research grants from national (Austria, Germany) and European funding agencies as well as from industry, is the editor of several books, and has authored or coauthored more than 200 scientific papers on parallel computer architecture, parallel programming, and multimedia communications and adaptation. His current research areas include distributed multimedia systems, multimedia communications, ICN/NDN and quality of service.

Prof. Hellwagner is a Member of the ACM, GI (German Informatics Society) and OCG (Austrian Computer Society), and a Vice President of the Austrian Science Fund.