



TECHNISCHE UNIVERSITÄT
ILMENAU

Faculty of Electrical Engineering and Information Technology
Institute for Media Technology
Audio Visual Technology

DISSERTATION

zur Erlangung des Akademischen Grades Doktoringenieur (Dr.-Ing.)

**Bitstream-based Video Quality Modeling and
Analysis of HTTP-based Adaptive Streaming**

vorgelegt von: Rakesh Rao Ramachandra Rao
Betreuender Gutachter: Prof. Dr.-Ing. Alexander Raake
Gutachter: Prof. Dr. Chulhee Lee
Gutachter: Prof. Dr. Maria Martini
Datum der Einreichung: 24.11.2022
Datum der Verteidigung: 24.05.2023

URN: [urn:nbn:de:gbv:ilm1-2023000123](https://nbn-resolving.org/urn:nbn:de:gbv:ilm1-2023000123)

DOI: [10.22032/dbt.57583](https://doi.org/10.22032/dbt.57583)

To my family

"I was born not knowing and have had only a little time to change that here and there." - Richard P. Feynman, The Life and Science of Richard Feynman (Letter to Armando Garcia J, December 11, 1985)

Acknowledgments

I have always believed that the process is as important as the result if not more. The process that has now brought me to this stage of writing up a thesis has been quite a ride. I wouldn't say that it was always a fun ride to be on but in the end, the ups seemed to have trumped the downs. It has certainly been a journey from going to being perceived as someone with limited scientific ambition to reaching a stage where I can certainly be proud of writing a thesis. Along the way have come people (and a few things!) to whom I owe my thanks and gratitude and so, now is the time to express that.

At the outset, I would like to thank Prof. Alexander Raake for firstly giving me an opportunity to work on the topic and for the many insightful discussions we have had especially during the P.NATS Phase 2 competition. I would also like to thank Prof. Chulhee Lee and Prof. Maria Martini for being reviewers of my thesis.

I have always liked a "pressure cooker" kind of a situation to deliver something (maybe because I am just too lazy to do things without any added additional pressure). One person who did a commendable job at maintaining that pressure all through this thesis, be it in terms of annoying me with paper titles or by being a devil's advocate at every step of the way, or by himself progressing at a faster pace thereby creating an atmosphere of friendly competition (one thing that I thrive on) among us was Steve Göring. I certainly owe a big thank you to him for all that and the innumerable late-night sessions that eventually led to many a publication.

This journey wouldn't have been as wonderful and enjoyable as it has turned out to be without the exchange with and company of Stephan Fremerey, Dominik Keller (extra thanks for our squash sessions, was good to take out the frustration on the court sometimes), Frank Hofmeyer, and Janto Skowronek. I would also like to thank Werner Robitza especially for being helpful at the beginning of this Ph.D. to get into the topic and being a constant collaborator throughout. I would also thank Tatiana Surdu for undertaking the painful task of proofreading this thesis.

Along this journey, owing to different projects, I have had the fortune of meeting some wonderful people. Firstly, this work would not have been possible without important inputs and contributions from Peter List, Bernhard Feiten, and Ulf Wüstenhagen. Also, I would like to extend my thanks to Saman Zadtootaghaj and Nabajeet Barman for our fruitful discussions. In particular, I would like to thank David Lindero for our many interesting and amazing discussions and also for his whiskey suggestions. I would also thank all the colleagues who were part of the P.NATS Phase 2 competition.

I would most definitely have to thank Monique Rodegast for making my life simpler in all administrative-related matters despite the lack of a stable common language. I would also extend my thanks to all the institute staff members who have been always helpful and been there to help whenever there was a need.

Now that I have thanked many people who have made this journey a memorable one, I would like to at least extend a minor thank you to a few very important things. First among them would be our office whiteboard (oh what would have I done without that, where would I track all the deadlines that I missed). I should certainly thank the magic of command-line tools without which this journey would have been a bit more painful than it actually turned out to be. Finally, how can I forget my thesis topic itself? After all these years, it gave me a legitimate reason to justify the disproportionate amount of time I spend watching movies and series and my parents can't complain anymore as I am working, you see!

I will end this by extending my biggest thanks to the three most important people: my father Ramachandra Rao (the original one) for his unwavering belief in me, my mother Meera for always being worried about me for no reason and my brother Roshan for holding the fort back home all these years.

Abstract

The pervasion of both affordable capture technology and increasing average bandwidth of internet connections has resulted in videos of high-quality (resolutions $\geq 1080p$ and framerates $\geq 60fps$) being streamed on the internet. The most preferred method to stream these videos is HTTP-based adaptive streaming, where usually an adaptation of video quality according to the available bandwidth is implemented using different media representations. Although adaptive streaming reduces the occurrences of video playout being stopped (called “stalling”) due to narrow network bandwidth, the automatic adaptation has an impact on the quality perceived by the user, which results in the need to systematically assess the perceived quality. Such an evaluation is usually done on a short-term (few seconds) and overall session basis (up to several minutes). In this thesis, both these aspects are assessed using subjective and instrumental methods. The subjective assessment of short-term video quality consists of a series of lab-based video quality tests which have resulted in publicly available datasets. The overall integral quality was subjectively assessed in lab tests with human viewers mimicking a real-life viewing scenario. In addition to the lab-based tests, the out-of-the-lab test method was investigated for both short-term video quality and overall session quality assessment to explore the possibility of alternative approaches for subjective quality assessment. This resulted in proposing an approach based on a pre-defined crop of the video cut out from the centre of the video for the out-of-the-lab settings. The instrumental method of quality evaluation was addressed in terms of bitstream- and hybrid pixel-based video quality models developed as part of this thesis. For this, a family of models, namely *AVQBits* has been conceived using the results of the lab tests as ground truth. Based on the available input information, four different instances of *AVQBits* are presented, that is, a Mode 3 model with full access to the bitstream, a Mode 0 variant using only metadata such as bitrate, resolution, framerate and codec as input, a Mode 1 model using metadata and frame-type and framesize information, and a Hybrid Mode 0 model that is based on Mode 0 and the decoded video pixel information. The Mode 3 model

that forms the core of *AVQBits* was developed in the context of the “P.NATS Phase 2 competition” conducted by ITU-T Study Group 12, Question 14, and has been standardized as ITU-T Rec. P.1204.3. Based on the winning model determination criteria outlined in the “P.NATS Phase 2 competition”, this model has been adjudged the winning bitstream model and, indirectly, also the best model among the 35 models in different categories consisting of bitstream-based, pixel-based, and hybrid models following an extensive validation. The *AVQBits* model instances have been evaluated under a large variety of different conditions and show either better or on-par performance in comparison with other state-of-the-art models considering their specific use case. The *AVQBits* models have further been evaluated for other application scopes such as 360° video, high framerate content, gaming videos, and images. Also, to assess the overall integral quality, a long-term integration model based on the standardized ITU-T P.1203.3 model is presented that can be applied for 30 s up to 5 min long audiovisual sequences. The different instances of *AVQBits* with the per-1-sec scores output are employed as the video quality component of the proposed long-term integration model. All *AVQBits* variants as well as the long-term integration module and the subjective test data have been made publicly available following an open-science approach for use by the community for further research.

Zusammenfassung

Die Verbreitung von erschwinglichen Aufnahmetechnologien und die zunehmende durchschnittliche Bandbreite von Internetverbindungen hat dazu geführt, dass Videos in hoher Qualität (Auflösungen $\geq 1080p$ und Frameraten $\geq 60fps$) über das Internet gestreamt werden. Die bevorzugte Methode zum Streamen dieser Videos ist das HTTP-basierte adaptive Streaming, bei dem in der Regel eine Anpassung der Videoqualität an die verfügbare Bandbreite unter Verwendung verschiedener Mediendarstellungen vorgenommen wird. Obwohl adaptives Streaming die Probleme des Anhaltens ("stalling") von Videos aufgrund geringer Netzwerkbandbreite reduziert, hat die automatische Anpassung Auswirkungen auf die vom Benutzer wahrgenommene Qualität, was die Notwendigkeit einer systematischen Bewertung dieser zur Folge hat. Eine solche Bewertung erfolgt in der Regel für einen kurzen Zeitraum (mehrere Sekunden) und für längere Sequenzen (bis zu mehreren Minuten). In dieser Arbeit werden diese beiden Aspekte mit subjektiven und instrumentellen Methoden bewertet. Die im Rahmen dieser Arbeit vorgenommene subjektive Bewertung der kurzfristigen Videoqualität besteht aus einer Reihe von laborgestützten Videoqualitätstests, die zu öffentlich verfügbaren Datensätzen geführt haben. Die gesamte integrale Qualität wurde in subjektiven Labortests mit menschlichen Betrachtern bewertet, die ein reales Betrachtungsszenario nachahmen. Zusätzlich zu den laborbasierten Tests wurde die Methode des Out-of-the-Lab-Tests sowohl für die kurzfristige Videoqualität als auch für die Bewertung der integralen Gesamtqualität untersucht, um die Möglichkeit alternativer Ansätze für die subjektive Qualitätsbewertung zu erkunden. Als Ergebnis wurde ein Ansatz vorgeschlagen, der auf einem vordefinierten Ausschnitt des Videos basiert, der aus der Mitte des Videos für die Out-of-the-Lab-Einstellungen herausgeschnitten wurde. Die instrumentelle Methode der Qualitätsbewertung wurde in Form von Bitstrom- und hybriden Pixel-basierten Videoqualitätsmodellen behandelt, die im Rahmen dieser Arbeit entwickelt wurden. Zu diesem Zweck wurde eine Familie von Modellen, im Folgenden *AVQBits*, konzipiert, wobei die Ergebnisse der Labortests als Basiswahrheit verwendet wurden.

Basierend auf den verfügbaren Eingabeinformationen werden vier verschiedene Instanzen von *AVQBits* vorgestellt, d. h. ein Mode-3-Modell mit vollem Zugriff auf den Bitstrom, eine Mode-0-Variante, die nur Metadaten (Bitrate, Auflösung, Framrate, Videocodec) als Eingabe verwendet, ein Mode-1-Modell, das Metadaten und Frame-Informationen verwendet, und ein hybrides Mode 0 Modell, das auf Mode 0 und den dekodierten Videopixelinformationen basiert. Das Mode-3-Modell, das den Kern von *AVQBits* bildet, wurde im Rahmen des von der ITU-T Study Group 12, Question 14, durchgeführten Wettbewerbs "P.NATS Phase 2" entwickelt und als ITU-T Rec. P.1204.3 standardisiert. Basierend auf den im "P.NATS Phase 2 Wettbewerb" beschriebenen Kriterien zur Bestimmung des Gewinners wurde dieses Modell nach einer umfangreichen Validierung als bestes Modell unter den eingereichten 35 Modellen in verschiedenen Kategorien, bestehend aus bitstrombasierten, pixelbasierten und hybriden Modellen, eingestuft. Die *AVQBits*-Modellinstanzen wurden unter einer Vielzahl unterschiedlicher Bedingungen bewertet und zeigen entweder eine bessere oder gleichwertige Leistung im Vergleich zu anderen State-of-the-Art-Modellen unter Berücksichtigung ihres spezifischen Anwendungsfalls. Die Modelle wurden auch für andere Anwendungsbereiche wie 360°-Videos, Inhalte mit hoher Framerate, Spielevideos und Bilder bewertet. Zur Bewertung der integralen Gesamtqualität wird außerdem ein Langzeit-Integrationsmodell auf der Grundlage des standardisierten Modells ITU-T P.1203.3 vorgestellt, das für 30 Sekunden bis zu 5 Minuten lange audiovisuelle Sequenzen angewendet werden kann. Die verschiedenen Instanzen von *AVQBits* mit einer Ausgabe von einem Qualitätsschätzwert pro Sekunde werden als Videoqualitätskomponente des vorgeschlagenen Langzeitintegrationsmodells verwendet. Alle *AVQBits*-Varianten sowie das Langzeitintegrationsmodul und die subjektiven Testdaten wurden im Rahmen eines Open-Science-Ansatzes öffentlich zugänglich gemacht, damit sie von der Forschungsgemeinschaft für weitere Forschungsarbeiten genutzt werden können.

Contents

1	Introduction	1
1.1	Quality of Experience	4
1.2	HTTP-based Adaptive Streaming	6
1.3	Short-term Video Quality Models	9
1.4	Overall Session Quality	12
1.5	Research Questions	13
1.6	Contributions by the Author	15
1.6.1	Publications	16
1.6.2	Open Source Software and Data	20
1.6.3	Patent Applications	20
1.7	Thesis Structure	21
2	State of the Art	23
2.1	Commonly Used Acronyms	24
2.2	Short-term Video Quality Assessment	24
2.2.1	Subjective Studies	24
2.2.2	Video Quality Models	30
2.3	Overall Integral Quality of a HAS session	35
2.3.1	Subjective Studies	35
2.3.2	Quality Models	36
2.4	Summary and Conclusion	37
3	Subjective Quality Assessment of 4K/UHD-1 Videos	39
3.1	Lab-based Subjective Quality Assessment	41
3.1.1	Short-term Video Quality Assessment	41
3.1.2	Overall Quality Assessment of a HAS Session	60
3.2	Out-of-the-lab Subjective Quality Assessment	63
3.2.1	Short-term Video Quality Assessment	64
3.2.2	Overall Quality Assessment of a HAS Session	72
3.3	Summary	79

Contents

4	<i>AVQBits</i>: Adaptive Bitstream-based Video Quality Model	83
4.1	P.NATS Phase 2 Competition	85
4.1.1	General Details	85
4.1.2	P.NATS Phase 2 Statistical Evaluation	86
4.2	Short-term Video Quality Models: Model Description	90
4.2.1	<i>AVQBits</i> M3 / P.1204.3	91
4.2.2	<i>AVQBits</i> M0	98
4.2.3	<i>AVQBits</i> M1	100
4.2.4	<i>AVQBits</i> H0	102
4.3	Short-term Video Quality: Model Training	104
4.3.1	<i>AVQBits</i> M3 / P.1204.3	105
4.3.2	<i>AVQBits</i> M0	106
4.3.3	<i>AVQBits</i> M1	107
4.3.4	<i>AVQBits</i> H0	108
4.4	Short-term Video Quality: Model Evaluation	108
4.4.1	Validation of <i>AVQBits</i> M3 / P.1204.3	109
4.4.2	Evaluation of <i>AVQBits</i> Model Instances	110
4.5	Other Prototype Models	113
4.5.1	ITU-T P.1203.1 Mode 0 Extension	114
4.5.2	Hybrid-VMAF	117
4.6	Summary	121
5	Overall Integral Quality	125
5.1	Model Description	126
5.2	Evaluation of the Overall Integral Quality Model	127
5.3	Summary	130
6	Extended Application Scopes of <i>AVQBits</i>	131
6.1	Gaming	132
6.1.1	Related Work for Gaming Video Quality Assessment	132
6.1.2	Datasets	134
6.1.3	Evaluation	136
6.2	360° Video	142
6.2.1	State of the Art for 360° Video Quality Assessment	143
6.2.2	360 Streaming Video Quality Dataset	144
6.2.3	Evaluation	145

6.3	High Framerate Video	149
6.3.1	Related Work for Quality Assessment of HFR Videos	149
6.3.2	LIVE-YT-HFR Dataset	150
6.3.3	Evaluation	151
6.4	Live Streaming Sports	152
6.4.1	Related work	154
6.4.2	LIVE-APV Dataset	155
6.4.3	Evaluation	156
6.5	Quality Evaluation of Videos with Pre-Existing Distortions	157
6.5.1	Related work	158
6.5.2	LIVE Wild Compressed Video Quality Database	160
6.5.3	Evaluation	161
6.6	Image Quality Evaluation	162
6.6.1	Related work	163
6.6.2	Dataset	164
6.6.3	Evaluation	164
6.7	Summary	168
7	Conclusion and Future Work	171
A	Subjective Test	179
B	P.NATS Phase 2 Test Plan	187
C	AVQBits Helper Functions	215
C.1	RfromMOS (5-Point MOS Scale to 100-Point Scale)	215
C.2	MOSfromR (100-Point Scale to 5-Point MOS Scale)	216
C.3	Scaletto5 (4.5-point scale of MOSfromR to 5-point scale)	216
	Bibliography	219
	List of Figures	245
	List of Tables	247
	List of Acronyms	251

Introduction

Video today has become the most dominant type of all the data uploaded, shared, and streamed on the internet, due to the availability of affordable capture technology and an increase in the average bandwidth of internet connections available to the users. Also, these factors have resulted in an increased proportion of these videos being of high-resolution. It was predicted that video traffic would account for 82% of all consumer traffic in 2022, up from 75% in 2017 with 4K/UHD-1 (3840×2160 pixels) video traffic accounting for 22% of this global video traffic [Cis22]. This increase in video traffic is also very well reflected in different statistics of popular streaming service providers, e.g., an average of 500 hours of video being uploaded per minute on YouTube [Woj20] and more than one billion videos watched per day on the same platform [Woj21]. Similar statistics have been reported for other streaming service providers as well. This trend is also substantiated by the increase in the paid subscriber count of these services, e.g., the number of paid subscribers on Netflix has increased to approximately 223 million in the third quarter of 2022 as compared to approximately 221 million in the second quarter of 2022 [Sto22]. In addition, these services have also been increasingly spending on producing their own content. One example of this is Netflix, which has increased its spending on content creation from approximately \$12 billion in 2020 to \$17 billion in 2021 [Iqb21]. This increase is not limited to traditional 2D video with a significant upsurge being witnessed in streaming of other video formats such as gaming video streaming [Gno21; Cha21] and 360° videos.

The main focus of any video streaming service provider is to enable an increase in the overall Quality of Experience (QoE) of the viewing session for the users/customers for a given bandwidth connection. Two key components that have attracted consid-

erable attention in this regard are the choice of the appropriate video compression method and the associated streaming technology.

A typical uncompressed 24-bit 1080p video at 60 fps requires a data rate of 2.98 Gbit/s - which makes transmission over networks an impractical task. Hence, videos are typically compressed before transmission. For this, efficient compression methodologies also called “video codecs” have to be developed which are capable of efficiently compressing videos while still maintaining a good visual quality for a specific target bitrate. The focus is on maintaining a good visual quality as that forms an important constituent in the overall QoE of the end user. In this regard, there has been continuous development and improvement over the past decades. Most notably, the Moving Pictures Expert Group (MPEG) has been involved in developing multiple standardized video codecs including H.264/ AVC (Advanced Video Coding) [ITU21a; Wie+03], H.265 (High Efficiency Video Codec (HEVC)) [ITU21b; Sul+12] and the more recently developed H.266 (Versatile Video Codec (VVC)) [ITU22; Bro+21]. In addition, other bodies such as the Alliance for Open Media (AOM) have also been involved in codec development, resulting, among others in the AV1 codec [Che+18]. Several studies have demonstrated the higher compression efficiency of newer codecs in comparison with older codecs. For example, studies have shown that VVC outperforms HEVC enabling the same visual perceived quality with a bitrate reduction of approximately 40% [Sid+19] for 4K/UHD-1 videos. Similarly, it has been reported that AV1 outperforms HEVC with average bitrate savings of approximately 8% for 4K/UHD-1 content [Zha+20].

The other component that plays a vital role in guaranteeing an overall good QoE is the streaming mechanism. The available bandwidth is expected to fluctuate during a streaming session. Hence, any kind of streaming approach should be able to adapt to these fluctuations and be able to stream the video maintaining a desired level of QoE for the user. This has led to the widespread adoption of HTTP-based adaptive streaming (HAS) [Sod11] as the preferred mechanism for both Video on Demand (VoD) and live-streaming use cases. Prominent streaming service providers such as YouTube [Ben13], Netflix, Amazon Prime Video, Vimeo, etc. use HAS as the underlying mechanism to stream videos. To enable HAS to adapt to the available bandwidth and also the device type, different representations of the video in terms of the considered resolution, framerate, and video codec are stored on the server and

the most fitting one is streamed based on the available bandwidth. Optimal encoding settings have to be determined to create such representations to enable a smooth and good streaming experience as the available bandwidth fluctuates. Different strategies have been proposed to choose optimal encoding settings, e.g. fixed-bitrate ladder [Inc14] by Apple, per-title encoding [Blo15] and per-shot encoding [Kat18] by Netflix and context-aware encoding [Rez+20] by Brightcove. To determine which encoding settings are appropriate for a given available bandwidth, it is required to quantify the perceived quality of the videos that are encoded with different settings.

All these developments demonstrate the need for an accurate assessment of the perceived video quality. This can usually be done using two different approaches. One approach is conducting subjective studies for perceptual assessment of short-term video quality, especially for videos of higher resolutions and framerate, and overall integral quality assessment of a HAS session. Subjective studies are considered a gold standard in multimedia quality assessment. Another approach is using highly accurate quality prediction models. These models are developed using the data obtained from subjective studies as ground truth.

This doctoral work focuses on perceptual video quality assessment, mainly in the context of HTTP-based adaptive streaming and consists of two main parts. The first part includes subjective assessment of short-term video quality with focus on videos of up to a resolution of 4K/UHD-1 and framerate up to 60fps being displayed on screens with a resolution $\geq 1080p$ and also the overall integral quality of a HAS session. The second part mainly focuses on the development of bitstream-based and hybrid quality models for videos up to a resolution of 4K/UHD-1 in the context of HAS.

To enable the reader to better understand this work and its potential applications, the subsequent sections in this chapter will introduce the concepts of QoE, HAS, short-term video quality models, and overall session quality assessment in the context of HAS.

1.1 Quality of Experience

Traditionally, Quality of Service (QoS) has been used for assessing different technologies. ITU-T Rec. E.800 [ITU08] defines QoS as “Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.”

However, over the years, technology-related assessment has evolved from system-centric QoS-oriented approaches to user-centric QoE-oriented approaches. In this work, the focus has been on developing models capable of estimating the QoE or a constituent of QoE, i.e., video quality in the context of multimedia adaptive streaming.

The Qualinet White Paper [LMP+12] defines QoE as follows:

Definition 1 *Quality of Experience (QoE): “is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user’s personality and current state.”*

This definition is restrictive as it only addresses experiencing a particular service or application. Instead, defining QoE from a global view that comprises not only experiencing a particular service or application, but also evaluating the contribution of a given application, system, or service implementation to the overall quality of experiencing is desired [RE14]. As a result, Raake and Egger [RE14] proposed the following definition of QoE.

Definition 2 *Quality of Experience (new) (QoE) is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person’s evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person’s context, personality and current state.*

QoE of a particular application or service may be influenced by several factors. The factors influencing QoE in the context of multimedia services, also called “Influence Factors” is defined in the Qualinet White Paper [LMP+12] as follows:

Definition 3 *Influence Factor (IF): Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user.*

The IFs can be classified into three different categories, namely, Human IFs (HIFs), System IFs (SIFs), and Context IFs (CIFs) [Bru+13]. These IFs are frequently interrelated, as illustrated in Figure 1.1.

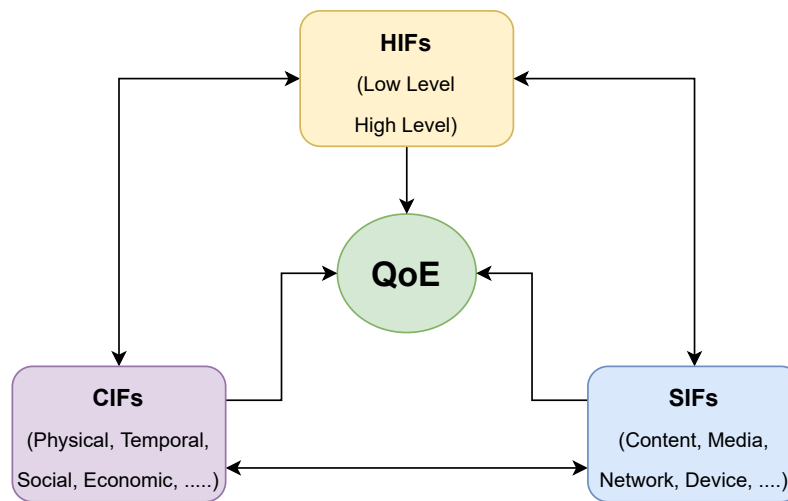


Figure 1.1: Factors influencing Quality of Experience.

A HIF is any characteristic of a human that impacts QoE. This is further classified into factors resulting from low- and high-level processing that may influence the perceptual and quality formation process. Low-level processing-related HIFs include demographic aspects such as visual acuity, age, and gender. In addition, it also comprises aspects related to the user's mood, motivation, attention, etc. Elements such as understanding and interpreting the stimuli that are presented, and the knowledge a particular user brings into a situation constitute the high-level processing-related HIFs.

Any aspect that affects the technically produced quality of an application or service falls into the category of SIFs. In the particular context of HAS, it may relate to any component in the end-to-end video processing chain. This can be associated with the content characteristics of the video and media-related aspects such as encoding, resolution, framerate, etc. Furthermore, network-related factors in the end-to-end

video processing chain such as bandwidth, packet loss, jitter, etc., and device-related factors associated with the end device also form part of SIFs.

The final category of IFs affecting QoE i.e., CIFs relates to aspects related to the user's environment. This comprises physical (location and space), temporal (time of day, content duration, etc.), social, economic, task, technical and information contexts.

For a typical media viewing session, QoE comprises different "constituents" such as audio quality, video quality, overall audiovisual quality, immersion, presence and more. Video quality forms one of the major "constituents" of QoE that is of concern to any streaming service provider, both for VoD and live-streaming scenarios. Hence, the main focus of this work is to model the effects of technical parameters such as bitrate, resolution, framerate etc. on the perception of video quality by users, in particular by using bitstream information. Following this, as one of the applications scopes, the developed models are used to estimate the overall QoE of a HAS viewing session by incorporating HAS-related aspects such as initial loading delay, stalling and quality switches, the details of which are explained in Section 1.4.

1.2 HTTP-based Adaptive Streaming

In recent years, HAS has replaced progressive download as the most widely used streaming technology for delivering videos on the internet. Different implementations of HAS have been used with Dynamic Adaptive Streaming over HTTP (DASH) being one of the most commonly used implementations [Bit21]. Additionally, other proprietary implementations are also available, e.g. Apple Inc.'s HTTP Live Streaming (HLS) [Inc14].

To illustrate the concept of HAS, MPEG-DASH is considered in the following as an example. The entire MPEG-DASH streaming process consists of three main steps, namely, encoding and segmentation, delivery, and decoding and playback.

- ▷ Encoding and segmentation: On the server side, a video is encoded and divided into smaller temporal chunks referred to as *segments*. The duration of the segments depends on the particular implementation of the streaming protocol and the considered application (e.g. VoD, Live streaming, etc.). Usually, in MPEG-

DASH, the typical segment duration is between 2-10 s [Pan11; ISO19]. The encoding settings for video consist of varying parameters such as resolution, bitrate, framerate, codec, etc. Over the years, several encoding strategies have been used to efficiently encode videos, with the aim of delivering optimal quality in case of constrained bandwidth. Starting with a fixed bitrate ladder [Inc14] which is content agnostic, several encoding strategies have been developed. Notable examples are per-title [Blo15] and per-shot [Kat18] encoding, which take content into consideration.

In addition to the representations, manifest files also called Media Presentation Description (MPD) files are created and stored on the server, which assign a segment to a particular representation and additional metadata that is required to enable playback.

- ▷ Delivery: Content delivery networks (CDNs) are used to deliver the requested representations to the client devices, along with the manifest files.
- ▷ Decoding and playback: On the client side, the DASH client first requests the MPD file to play the content. MPD describes a manifest of the available content, its representations, their URL addresses, and other characteristics. The client then parses the MPD and extracts information about the program timing, available media content, media types, resolutions, minimum and maximum bandwidth, and other features. Then, based on the factors such as available bandwidth, preferred settings, etc. the client selects the appropriate streaming alternative and plays out the content [Sod11]. The MPD hierarchical model based on the work from Sodagar [Sod11] is illustrated in Figure 1.2.

The behaviour of the client w.r.t requesting segments can further be used to create better bitrate ladders, for optimizing the overall streaming quality by considering network and device playback statistics, as it is done for example in the Context-Aware Encoding developed by Brightcove [Rez+20].

HAS can be implemented in various forms, for example, MPEG DASH, Apple's HLS or Microsoft Smooth Streaming (MSS). Furthermore, to simplify the delivery of HTTP-based adaptive streaming media, MPEG created the Common Media Application Format (CMAF) [Cis]. CMAF is a container format with tools that enables single-

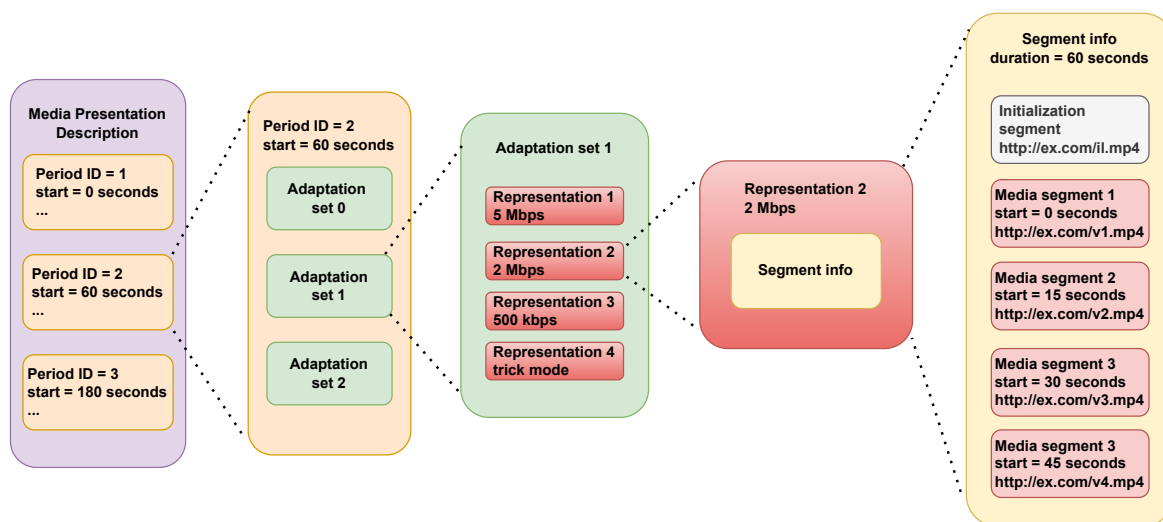


Figure 1.2: The MPD hierarchical model. This example shows how the client requests the appropriate representation and plays out the segment (adapted from [Sod11]).

approach video streaming that works with different protocols like MPEG-DASH, HLS, etc.

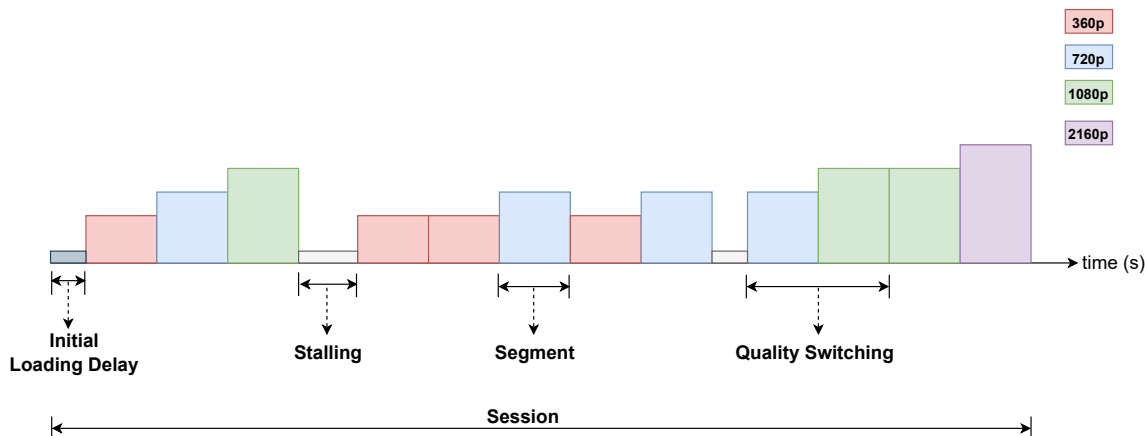


Figure 1.3: An example HAS session

Figure 1.3 illustrates a typical HAS session. It can be characterized by three types of streaming-related events that occur during such a session. These are *initial loading delay*, *stalling* and *quality switching*. Initial loading delay (ILD) is defined as the time elapsed between the request for a video and the first played-out frame [ITU16b]. Also, during an HAS session, the playout buffer may run empty resulting in the freezing of the video. This event is called *stalling*. [ITU16b; ITU16a]. Here, the

number and the duration of stalling play an important role in the overall quality perception of a user [Hoß+11]. In addition to ILD and stalling, a typical HAS session involves the client adapting the playout depending on the target device and its screen or playout window resolution, as well as the available network bandwidth. Due to this, the video quality varies based on the representation that is chosen for playout. This variation of video quality at different instances during a HAS session is termed as *quality switching*.

As mentioned above, over the years, significant advancements have been made in terms of encoding optimization so as to create an efficient representation of segments. This is done with the aim of both reducing the number and duration of stalling events and also quality switches and reducing the overall used bandwidth. Based on a study by Wassermann, Wehner, and Casas [WWC19], more than 90% of the video sessions played smoothly without stalling events on smartphones as of 2018. This is significantly better compared to 2016 during which only 60% of the video sessions played smoothly without stalling events [WWC19].

All these special characteristics of HAS including the encoding- and playout-related aspects require models for assessing the quality of the segments and the overall QoE of an HAS session. Hence, there is a need for the development of both short-term video quality models and overall QoE estimation models. These models are also needed for usage in various other applications, starting from encoding optimization, to overall quality monitoring. In the following, the fundamentals of both these model types are described.

1.3 Short-term Video Quality Models

Short-term video quality models refer to objective quality models used to assess the quality of the encoded video representation, in the special case of HAS, encoded segments. They measure the distortion due to compression on the perceived video quality. Such short-term video quality models can then be used together with the measurement of other degradations such as stalling or perceivable quality switches to predict the overall QoE of a HAS session. They can additionally be used to optimize

the encoding settings to generate optimal encodes, thus leading to a high QoE of the end user.

In general, based on the input information used for quality assessment, video quality models can be classified into several categories [Chi+11; Sha+14b; Raa+11], for example, metadata-based, pixel-based, bitstream-based, or hybrid models.

For the first type of models, namely, metadata-based models, the input information available is limited to video resolution, video bitrate, video framerate, and video codec and usually does not require parsing of the given bitstream. Since no information related to the underlying video content is available to the model, these models are content-agnostic.

The second type of models is called pixel-based and has access to pixel information to estimate video quality. Based on the availability of the pixel information of a reference (uncompressed source) video, pixel-based models can be further divided into three categories:

- ▷ Full-Reference (FR) models: FR models have complete access to the pixel information of the reference video, in addition to the distorted video; e.g.: VMAF [Net18]
- ▷ Reduced-Reference (RR) models: These models have “reduced”/“partial” access to the reference video, along with the distorted video; e.g.: ITU-T P.1204.4 [ITU19c]
- ▷ No-Reference (NR) models: NR models only have access to the distorted video for quality estimation; e.g: Devisq [GSR18]

The third type of models is called bitstream-based, which are usually NR models that just use the encoded bitstream without decoding to estimate the visual quality. Based on the degree of availability of the bitstream information, bitstream models can be categorized into the following modes of operation:

- ▷ Mode 0: This is a metadata-based model and has access to bitrate, resolution, framerate and codec as input information for quality estimation. A notable example of such a model is the ITU P.1203.1 Mode 0 [ITU19e; Raa+17].

- ▷ Mode 1: This category of models have access to frame size and frame type information (I and Non-I) in addition to metadata; for example ITU-T P.1203.1 Mode 1 [ITU19e; Raa+17]
- ▷ Mode 3: A Mode 3 model has complete access to the bitstream to estimate the video quality; for example, ITU-T P.1203.1 Mode 3 [ITU19e; Raa+17], ITU-T P.1204.3 [ITU19b; Raa+20a; Rao+20a]

It is noted that also Mode 2 was proposed as another model category, cf. [ITU16b; ITU19e; Raa+17], with access to the full bitstream like Mode 3, yet with a maximum of 2% of the bitstream being parsed. In ITU-T Rec. P.1203 [ITU16b], this model type is still comprised. The idea was to enable in-network measurements with a massive number of streams parsed at the same time. With today's encrypted traffic, this model variant has become mostly obsolete.

The last type of models to estimate video quality are hybrid models. Here, usually, a combination of bitstream and pixel information is used to estimate video quality. One example of a hybrid model is ITU-T P.1204.5 [ITU19d; Raa+20a].

Each model category has different application scopes. For example, a bitstream-based Mode 3 model can be used for bitrate ladder derivation by a service provider as the provider has complete access to the encoded bitstream. The advantage of such bitstream models is they are usually less computationally complex.

Pixel-based models can be used for different use cases depending on the availability of the reference video. FR models can be efficiently used for bitrate ladder derivation at the server side where there is potential access to the reference video. The advantage of a pixel-based model over a bitstream-based model for such a use case is that the pixel-based model is codec-independent. Another advantage of the pixel models is the quality monitoring at the client side despite having encrypted streams as the decoded pixels are available. The realistic model type that can be used is the NR type of models as the reference video is typically unavailable at the client side. However, studies have shown that pixel-based NR models perform worse than a Mode 0 model [RGR22]. Furthermore, as compared to the bitstream-based models, pixel models are usually computationally more complex.

Similarly, hybrid models can be used for client-side monitoring, depending on the type of the hybrid model and the information that is available from the bitstream.

In general, different approaches have been used to develop such models. These approaches range from a simple curve-fitting-based approach [ITU12b; YG13; ITU19e; ITU19c; ITU19d] to more complex machine-learning-based (ML) approaches [MMB12; MSB13; GRR19] and approaches based on deep neural network (DNN) [GSR18; Utk+20; Zad+20a].

As part of this work, a combination of approaches consisting of traditional curve fitting and machine-learning has been used to develop the proposed models. DNN models are not considered due to the following two reasons:

1. The ML-alternatives employed in the models proposed as part of this thesis are more light-weight compared to DNN models.
2. The lack of availability of large amount training data needed to develop a well-performing DNN model, e.g. ImageNet uses more than 1 million images for training [Den+09]. Obtaining such large amounts of video data with quality annotations obtained from subjective tests as ground truth is infeasible.

1.4 Overall Session Quality

There is a general tendency to regard QoE as a static event and as a result, the QoE measured for a stimulus of delimited length is assumed to be stable along with its duration. However, in an audiovisual session extending over several minutes, this is rarely the case [Wei+14]. This is more evident in a typical HAS session lasting several minutes, as such a session would include different quality-related events, which are, for example, audio and video quality switching, initial loading delay, and stalling as illustrated in Figure 1.3. Hence, to assess the overall session quality it is important to include the time-dependent impact of these different quality-related events, and thus also the recency and primacy effects [GPL00] on the quality perception of users. One notable example of such a model is ITU-T Rec. P.1203.3 [ITU20; RGR17]. The current amendment of this model is based on the author's contribution to ITU-T SG12/Q14. In the present work, an adaptation of the ITU-T Rec. P.1203.3 model is proposed by taking into account the advancements in encoding strategies and the updated user preferences and is included as an appendix in ITU-T Rec. P.1204.3 [ITU19b].

1.5 Research Questions

Based on the considerations above, as well as the need to develop both short-term and long-term quality models for newer codecs and high-resolution contents ($> 1080p$), the following research questions have been derived to address the existing challenges in the quality assessment and modeling of high-resolution videos considering HTTP-based adaptive streaming applications.

Research Question 1 *Can video quality be accurately predicted for higher resolutions using only bitstream information?*

The primary focus of this research question is to investigate the possibility of developing quality models using only bitstream information in the context of 4K/UHD-1. These models should be able to estimate the quality of videos with higher resolutions such as 4K/UHD-1, and framerates up to 60 fps, and which are encoded with widely used video codecs, e.g. H.264, H.265, and VP9. Moreover, it is important to ensure to evaluate the performance of such models on unknown data, to investigate the robustness of the models on varying encoding conditions. The main intention of considering only bitstream-based models is that they are usually both computationally less complex and faster in comparison with traditional FR or other pixel-based models. It is also beneficial to develop models with a common architecture that can be adapted to different scenarios, based on the available input information and specific use cases.

Research Question 2 *How can bitstream models be used in cases where only a limited amount of input data is available for precise video quality estimation?*

There exist scenarios where the entire bitstream information may not be available to assess the video quality and the overall QoE in an HAS context. In such cases, it is desirable to have a model that can easily be adapted to such scenarios. Hence, this research question 2 addresses this need and involves adapting a model which has access to the complete bitstream information to different scenarios. For example, in such scenarios, a model may only have access to metadata, a reduced set of bitstream data, or the pixel information of the distorted video.

Research Question 3 *Can bitstream and hybrid models be applied for video quality assessment of application scopes other than traditional 2D videos?*

In addition to different scenarios based on the availability of input information, there are use cases that differ based on video formats such as 360° video, High Framerate (HFR) video, etc. Here, it would be advantageous that a traditional 2D video quality model could be applied to the new use cases either out-of-the-box or with minimal modifications. This would reduce the development time of models for newer applications. Therefore, as part of this research question, models developed in the process of addressing research questions 1 and 2 will be assessed for their applicability and adaptability to other applications such as 360°, HFR and gaming videos and also images encoded with video codecs [GR19].

Research Question 4 *Can bitstream and hybrid models be used to predict the overall QoE of a longer ($\geq 1min$) HAS session?*

A typical HAS session lasts more than a few minutes and is characterized by various factors outlined in Section 1.4. One of the main factors affecting the overall perceived QoE of a HAS session is the perceivable video quality switches. Therefore, in this research question, the focus is on assessing the applicability of the developed models for the prediction of the overall quality of a HAS session involving audiovisual sequences of duration ranging from 1 *min* to 5 *min* by employing them as the video quality component in a long-term integration model.

Research Question 5 *How can quality assessment of high resolution videos be conducted in an out-of-the-lab setting?*

Traditionally, video quality assessment tests have mostly been conducted in a controlled lab setting. However, conducting lab tests is both time-consuming and expensive. Moreover, other aspects, e.g., the Covid-19 pandemic, have shown that there may be instances where such lab tests cannot be conducted, due to reasons that are not just limited to technical aspects. Hence, alternative testing paradigms for

out-of-the-lab settings have to be conceived. There is a vast literature on using crowd and remote testing for multimedia quality assessment [Hos+17; SB19; Hos+20]. However, most of the conducted studies cannot be applied to videos of higher resolution such as 4K/UHD-1, for example, since users do not necessarily have 4K/UHD-1 capable screens and hardware. As a result, with this research question, the focus is on developing a methodology for the assessment of video quality in the case of higher-resolution videos in an out-of-the-lab setting.

1.6 Contributions by the Author

In the course of addressing the research questions outlined in the previous section, the author has contributed to the state-of-the-art in various ways. The contributions can be classified into three categories, namely, publications, open source software and data, and patent applications. The publications are organized into different categories based on the research focus. Firstly, publications that are relevant to this thesis are listed. These include publications related to high resolution video quality datasets, bitstream-based, pixel-based, and hybrid video quality models and contributions to standardization. Following this, other publications of the author are listed. As part of this thesis, data and software have been made publicly available for reproducibility and further development. These include high resolution video quality datasets consisting of source videos, distorted videos, quality ratings in the form of mean opinion scores (MOS) and other metadata, and also the reference implementation of the proposed models. Finally, the patent applications that resulted as part of this thesis are listed.

In addition, the “Mode 3” type model developed by the author as part of this work has been standardized as ITU-T Rec. P.1204.3 [ITU19b]. Furthermore, the models developed by the author in the categories of “Mode 0” and “Mode 1” and submitted to the “P.NATS Phase 2” competition conducted by ITU-T Study Group 12 / Question 14 (SG12/Q14) were part of the winning groups in both the categories.

1.6.1 Publications

The publication categories are ordered according to their applicability to this thesis.

Dataset Publications

- [Rao+19a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Werner Robitza, Bernhard Feiten, and Alexander Raake. “AVT-VQDB-UHD-1: A Large Scale Video Quality Database for UHD-1”. In: *21st IEEE International Symposium on Multimedia (IEEE ISM)*. Dec. 2019
- [RGR21b] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Towards High Resolution Video Quality Assessment in the Crowd”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 2021
- [GRR23] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Quality Assessment of Higher Resolution Images and Videos with Remote Testing”. In: *Quality and User Experience (QUEx) 8* (2023)
- [Rao+23] **Rakesh Rao Ramachandra Rao**, Silvio Borer, David Lindero, Steve Göring, and Alexander Raake. “PNATS-UHD-1-Long: An Open Video Quality Dataset for Long Sequences for HTTP-based Adaptive Streaming QoE Assessment”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023

Publications related to Bitstream Models

- [Rao+19b] **Rakesh Rao Ramachandra Rao**, Steve Göring, Patrick Vogel, Nicolas Pachatz, Juan Jose Villamar Villarreal, Werner Robitza, Peter List, Bernhard Feiten, and Alexander Raake. “Adaptive video streaming with current codecs and formats: Extensions to parametric video quality model ITU-T P.1203”. In: *Electronic Imaging* (2019)
- [Rao+20a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Peter List, Werner Robitza, Bernhard Feiten, Ulf Wüstenhagen, and Alexander Raake. “Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analysis and Open Source Implementation”. In: *Twelfth IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020

- [Rao+20b] **Rakesh Rao Ramachandra Rao**, Steve Göring, Robert Steger, Saman Zad-tootaghaj, Nabajeet Barman, Stephan Fremerey, Sebastian Möller, and Alexander Raake. “A Large-scale Evaluation of the bitstream-based video-quality model ITU-T P.1204.3 on Gaming Content”. In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020
- [Raa+20a] Alexander Raake, Silvio Borer, Shahid Satti, Jörgen Gustafsson, **Rakesh Rao Ramachandra Rao**, Stefano Medagli, Peter List, Steve Göring, David Lindero, Werner Robitza, Gunnar Heikkilä, Simon Broom, Christian Schmidmer, Bernhard Feiten, Ulf Wüstenhagen, Thomas Wittmann, Matthias Obermann, and Roland Bitto. “Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204”. In: *IEEE Access* 8 (2020)
- [GRR20] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Prenc – Predict Number Of Video Encoding Passes With Machine Learning”. In: *Twelfth IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020
- [Rob+21] Werner Robitza, **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Impact of Spatial and Temporal Information on Video Quality and Compressibility”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. June 2021
- [RGR22] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications”. In: *IEEE Access* 10 (2022)
- [Rob+22] Werner Robitza, **Rakesh Rao Ramachandra-Rao**, Steve Göring, Alexander Dethof, and Alexander Raake. “Deploying the ITU-T P.1203 QoE Model in the Wild and Retraining for New Codecs”. In: *Proceedings of the 1st Conference on Mile-High Video*. MHV '22. Denver, Colorado: Association for Computing Machinery, 2022
- ▷ Contributions to standardization: ITU-T Rec. P.1203.3 [ITU20], P.1204 [ITU19a], P.1204.3 [ITU19b]

Publications related to Pixel Models

- [GRR19] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “nofu - A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content”. In: *Eleventh IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany, June 2019
- [Gör+20] Steve Göring, Robert Steger, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Automated Genre Classification for Gaming Videos”. In: *22nd IEEE International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020
- [Zad+20a] Saman Zadtootaghaj, Nabajeet Barman, **Rakesh Rao Ramachandra Rao**, Steve Göring, Maria G. Martini, Alexander Raake, and Sebastian Möller. “DEMI: Deep Video Quality Estimation Model using Perceptual Video Quality Dimensions”. In: *22nd IEEE International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020
- [RGR21a] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Enhancement of Pixel-based Video Quality Models using Meta-data”. In: *Electronic Imaging, Human Vision Electronic Imaging*. 2021
- [Gör+21a] Steve Göring, **Rakesh Rao Ramachandra Rao**, Bernhard Feiten, and Alexander Raake. “Modular Framework and Instances of Pixel-Based Video Quality Models for UHD-1/4K”. in: *IEEE Access* 9 (2021)

Publications on 360° Video, VR and other Video Aspects/Formats

- [Sin+19] Ashutosh Singla, **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Assessing Media QoE, Simulator Sickness and Presence for Omnidirectional Videos with Different Test Protocols”. In: *26th IEEE Conference on Virtual Reality and 3D User Interfaces*. Osaka, Japan, Mar. 2019
- [Raa+20b] Alexander Raake, Ashutosh Singla, **Rakesh Rao Ramachandra Rao**, Werner Robitza, and Frank Hofmeyer. “SiSiMo: Towards Simulator Sickness Modeling for 360° Videos Viewed with an HMD”. in: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2020
- [Fre+20] Stephan Fremerey, Steve Göring, **Rao Rakesh Ramachandra Rao**, Rachel Huang, and Alexander Raake. “Subjective Test Dataset and Meta-data-based

- Models for 360° Streaming Video Quality”. In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020
- [Sin+21] Ashutosh Singla, Steve Göring, Dominik Keller, **Rakesh Rao Ramachandra Rao**, Stephan Fremerey, and Alexander Raake. “Assessment of the Simulator Sickness Questionnaire for Omnidirectional Videos”. In: *28th IEEE Conference on Virtual Reality and 3D User Interfaces*. 2021
- [Kel+21] Dominik Keller, Markus Vaalgamaa, Erkki Pajanen, **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Groovability: Using Groove as a Novel Measure for Audio QoE with the Example of Smartphones”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 2021
- [Gör+21b] Steve Göring, **Rakesh Rao Ramachandra Rao**, Stephan Fremerey, and Alexander Raake. “AVRate Voyager: An open source online testing platform”. In: *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2021
- [Kel+23] Dominik Keller, Felix von Hagen, Julius Prenzel, Kay Strama, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Influence of Viewing Distances on 8K HDR Video Quality Perception”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023
- [Bra+23] Florian Braun, **Rakesh Rao Ramachandra Rao**, Werner Robitza, and Alexander Raake. “Automatic Audiovisual Asynchrony Measurement for Quality Assessment of Videoconferencing”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023
- [Gör+23] Steve Göring, **Rakesh Rao Ramachandra Rao**, Rasmus Merten, and Alexander Raake. “Appeal and quality assessment for AI-generated images”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023
- [Dia+23] Chenyao Diao, Luljeta Sinani, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Revisiting Videoconferencing QoE: Impact of Network Delay and Resolution as Factors for Social Cue Perceptibility”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023

1.6.2 Open Source Software and Data

In addition to the aforementioned publications, some software tools and datasets that have been used as part of the publications have been made publicly available, to promote reproducible research and aid further development.

- ▷ *AVT-VQDB-UHD-1*¹: A database consisting of data of the conducted 4K/UHD-1 video quality tests, presented in [Rao+19a; RGR21b].
- ▷ *ITU-P.1204.3 reference implementation*²: P.1204.3 reference implementation, presented in [Rao+20a].
- ▷ *ITU-P.1204.3 video bitstream parser*³: Contributions to the reference video parser for the P.1204.3 prediction model; see [Rao+20a].
- ▷ *ITU-T P.1204.3 extensions*⁴: Reference implementation of different bitstream and hybrid-based extensions of ITU-T P.1204.3, presented in [RGR22].
- ▷ *PNATS-UHD-1-Long*⁵: A database consisting of data of the conducted 4K/UHD-1 long duration (1-5 min) audiovisual quality tests, presented in [Rao+23].

1.6.3 Patent Applications

- ▷ R. Ramachandra, S. Göring, A. Raake, P. List, W. Robitza, B. Feiten, U. Wüstenhagen. WO002021064136: Information-adaptive mixed deterministic/machine-learning-based bit stream video quality model.
- ▷ P. List, R. Ramachandra, W. Robitza, A. Raake, S. Göring, U. Wüstenhagen, B. Feiten. WO002021013946: System and method to estimate blockiness in transform-based video encoding.

¹<https://github.com/Telecommunication-Telemedia-Assessment/AVT-VQDB-UHD-1>

²https://github.com/Telecommunication-Telemedia-Assessment/bitstream_mode3_p1204_3

³https://github.com/Telecommunication-Telemedia-Assessment/bitstream_mode3_videoparser

⁴https://github.com/Telecommunication-Telemedia-Assessment/p1204_3_extensions

⁵<https://github.com/Telecommunication-Telemedia-Assessment/PNATS-UHD-1-Long>

1.7 Thesis Structure

To address the research questions outlined in Section 1.5, this thesis is organized into different chapters. Firstly, a detailed overview of the state-of-the-art is provided in Chapter 2 “State of the Art” which includes the subjective assessment of video quality of high-resolution videos, quality models for estimating short-term video quality, and overall QoE of a HAS session to get better insights to answer the defined research questions. Following this, a series of datasets is described, related to short-term video quality assessment of 4K/UHD-1 videos and overall QoE assessment of an HAS session. These datasets have been created as part of this thesis and are described in Chapter 3 “Subjective Quality Assessment of 4K/UHD-1 Videos”. These comprise tests conducted in traditional lab settings and also out-of-the-lab. Afterwards, in Chapter 4 “*AVQBits*: Adaptive Bitstream-based Video Quality Model”, the different models developed during this thesis are presented. These include three bitstream-based models and two versions of a hybrid model. The performance of the developed models is then extensively evaluated using the datasets described in Chapter 3, and also other publicly available video quality datasets to assess the robustness of the models. In addition to the proposed models, the extension of the ITU-T P.1203.1 mode 0 model [ITU19e] for newer codecs and higher resolutions and framerates, and a hybrid variant of VMAF [Net18] are described in this chapter. Following this, an overall QoE assessment model based on ITU-T Rec. P.1203.3 is described in Chapter 5 “Overall Integral Quality”. This chapter also includes an evaluation of the proposed models using the long-video dataset described in Chapter 3.

Furthermore, in Chapter 6 “Extended Application Scopes of *AVQBits*”, the applicability of the developed bitstream models for different application scopes such as 360°, gaming, HFR videos, live-streamed sports content and images is evaluated. For this purpose, publicly available datasets are used. Finally, in Chapter 7 “Conclusion and Future Work”, a brief conclusion of this thesis is presented, along with an outlook for future work concerning the subjective assessment of high-resolution videos and video quality model development for newer use cases.

State of the Art

Multimedia quality assessment can be conducted using two approaches, namely, subjective and instrumental methods. Subjective methods consist of conducting quality assessment studies and gathering opinion scores. Instrumental methods involve developing models for quality prediction. As already mentioned in Chapter 1, both these approaches are explored in this thesis for quality assessment of short-term video quality and overall quality of a HAS session. Hence, this chapter will first provide an in-depth overview of subjective studies using both lab- and crowd-based approaches, and bitstream-based and hybrid models related to short-term video quality assessment of videos up to a resolution of 4K/UHD-1 videos. Following this, subjective studies and models related to the overall quality assessment of an HAS session reported in literature are outlined.

This chapter is based on the following publications:

- [Rao+19a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Werner Robitza, Bernhard Feiten, and Alexander Raake. “AVT-VQDB-UHD-1: A Large Scale Video Quality Database for UHD-1”. In: *21st IEEE International Symposium on Multimedia (IEEE ISM)*. Dec. 2019
- [RGR21b] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Towards High Resolution Video Quality Assessment in the Crowd”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 2021
- [Rao+20a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Peter List, Werner Robitza, Bernhard Feiten, Ulf Wüstenhagen, and Alexander Raake. “Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analy-

sis and Open Source Implementation”. In: *Twelfth IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020

[RGR22] Rakesh Rao Ramachandra Rao, Steve Göring, and Alexander Raake. “AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications”. In: *IEEE Access* 10 (2022)

2.1 Commonly Used Acronyms

Before proceeding further with describing the state-of-the-art (SoA) studies related to subjective and instrumental quality assessment of videos with a resolution of up to 4K/UHD-1, a list of commonly used acronyms related to the quality assessment of videos, in general, is presented in this section.

- ▷ SRC (Source): The original undistorted source material that is subjected to different encodings is referred to as SRC. This is also called as reference video.
- ▷ HRC (Hypothetical Reference Circuit): This refers to the various encoding conditions that are applied to a SRC.
- ▷ PVS (Processed Video Sequence): This is the result of the application of a HRC to a SRC which is then shown to the subjects for rating the video quality.

2.2 Short-term Video Quality Assessment

In this section, different subjective studies for quality assessment of videos with a resolution of up to 4K/UHD-1 reported in literature are presented. In addition, different SoA bitstream and hybrid video quality models for the prediction of video quality are also presented in this section.

2.2.1 Subjective Studies

In general, two approaches have been widely used for subjective video quality assessment. One approach is the traditional lab-based subjective tests under controlled

conditions. The other approach is based on crowdsourcing, where usually the test environment is not well controlled. Hence, this section provides an overview of both lab-based and crowdsourcing-based subjective quality assessments of high-resolution videos. It should be noted that the primary approach for video quality assessment of 4K/UHD-1 resolution used in this thesis has been lab-based subjective tests.

2.2.1.1 Lab-based Approach

Following the standardization of UHDTV [ITU12a] by the International Telecommunication Union (ITU-R) in 2012, several studies on perceptual assessment of 4K/UHD-1 content have been presented [Bae+13; Van+16; XJ13]. In addition, some studies have been published on benchmarking of SoA objective models [CL14; HKE13], and some on making 4K/UHD-1 datasets publicly available [CL18; Son+13].

Song et al. [Son+13] present a study describing a set of 15 4K/UHD-1 video contents that are made publicly available for further research. It was focused on the qualitative analysis of the video contents in the form of spatiotemporal complexity of the source contents and hence no quality-related analysis regarding encoding-related effects was made. On the other hand, Bae et al. [Bae+13] present the results of a subjective quality test with 4K/UHD-1 content which focused on investigating the impact of encoding on the perceived quality of 4K/UHD-1. For this purpose, videos encoded with H.265 were considered. Unlike many experiments using the Absolute Category Rating (ACR) method for quality assessment of 4K/UHD-1 content, the Double-stimulus Impairment Scale (DSIS) [ITU14b] method was used to assess perceptual quality differences between 4K/UHD-1 contents encoded at different bitrates, color formats, and viewing distances in this test. Note that, studies have shown that there is no significant difference between DSIS and ACR test methodologies [Tom+10; CWH16]. Another subjective evaluation, conducted by Bae et al. [Bae+13], focused on comparisons between contents in 4K/UHD-1 resolution with different encoding parameters, but not between different resolutions.

In Xu and Jiang [XJ13], a subjective quality assessment of 4K/UHD-1 videos using the Double-Stimulus Continuous Quality Scale (DSCQS) [ITU14b] method was presented. The goal of the test was to analyze the effect of bitrate on 4K/UHD-1

content. H.264 was the only codec used for encoding. One limitation of this study is that there is no comparison between different resolutions in the subjective test. This is vital as HAS consists of adapting the resolution of the video depending on the available bandwidth and hence it is important to assess the perceived quality differences across different resolutions.

Furthermore, to analyze the differences between full-HD and UHD, and if users can perceive a difference, Berger et al. [Ber+15] present a study comparing the perceived quality of transmitting 4K/UHD-1 content compared to Full-HD content at the same bitrate, encoded with HEVC. In total, 15 different contents and 4 different bitrate settings were chosen to span a wide range of the employed quality scale. The ACR-HR (ACR with Hidden Reference removal) test method was used. From the results it can be concluded that there is not always a significant quality difference between videos transmitted in Full-HD and 4K/UHD-1 resolution, however, the results strongly depend on the content type and the capture quality. No analysis based on SoA models is included in that study. Van Wallendael et al. [Van+16] performed a similar test, where 4K/UHD-1 and Full-HD resolutions were compared. They also come to a similar conclusion as [Ber+15], namely that the perceptibility of a 4K/UHD-1 advantage over Full-HD is highly content-dependent. In [Gör+19], Göring et al. developed an automated system to predict whether there is a benefit of using 4K/UHD-1 over Full-HD. They conclude that nearly 50% of their analyzed source videos will not have any perceivable benefit in 4K/UHD-1.

Besides the aforementioned research, there are also studies available that evaluate the applicability of SoA models for the quality evaluation of 4K/UHD-1 videos. For example, the authors in Hanhart, Korshunov, and Ebrahimi [HKE13] compare several common quality models such as PSNR, VSNR, SSIM, MS-SSIM, VIF, and VQM for 4K/UHD-1 videos based on a subjective quality test conducted by the authors. The authors conclude that VIF may be used as a general-purpose metric for 4K/UHD-1 videos. However, only 4 contents in total (1 for training + 3 for testing) have been considered and no modern quality features and models such as VMAF were used, mainly because those models were not available when the research was conducted. Further, the study only analyses the effect of two codecs, H.264 and HEVC.

2.2 Short-term Video Quality Assessment

Similarly, Lee et al. [Lee+17], investigated the applicability of traditional models such as PSNR, SSIM, etc., to monitor the perceptual quality of adaptive streaming services. For evaluation, subjective test data for high-quality 4K/UHD-1 video sequences along with the down-scaled versions were used. However, no coding degradations were considered.

In Rassool [Ras17], VMAF has been analyzed to determine whether it can be used for quality prediction in the case of 4K/UHD-1 content. The study uses 10 video sequences from a Xiph dataset [VQEG4K] which are encoded with bitrates ranging between 3mb/s and 10mb/s and show a high correlation between MOS and VMAF scores. However, one limitation of this study is that only one proprietary video codec (rmXD) was used.

Cheon and Lee [CL18] present a larger comparison of subjective and instrumental quality assessments of compressed 4K/UHD-1 videos. Three codecs, H.264, HEVC, and VP9 are addressed. In addition to analyzing the effect of different encoding parameters and benchmarking the existing SoA quality models, they also make the video and subjective data publicly available. The study only uses source contents with a framerate of 30 fps whereas it is recommended to use a framerate of at least 50fps in the UHD-1 specification. Moreover, only a comparison test of 4K/UHD-1 and Full-HD was conducted. However, in a classical HAS or DASH scenario, several low-resolution representations should also be included. Further, no results for more recent models such as VMAF are reported. All the analysis was based on just one subjective test with 25 participants who rated 250 video sequences (240 compressed + 10 reference sequences). This dataset has been made publicly available. In addition, other 4K/UHD-1 datasets have also been made publicly available [Li+19b; MVV20].

More recently, video quality assessment studies for higher resolution have also focused on other aspects which are, for example, higher framerates, and newer application scopes such as assessment of live-streamed content and user-generated content (UGC). Madhusudana et al. [Mad+21] conducted a large-scale study on the subjective and instrumental quality of high framerate video with framerates up to 120 fps. For this purpose, a large dataset called the LIVE-YouTube-HFR (LIVE-YT-HFR) containing 480 PVSs was created which are subjectively evaluated by a total of 85 participants. The LIVE-YT-HFR dataset was made publicly available. An evaluation of existing FR and NR models has been performed and it has been

reported that the GSTI [Mad+20a] model outperforms all the considered SoA models including VMAF. GSTI uses a statistical entropic differencing method based on a Generalized Gaussian Distribution model expressed in both the spatial and temporal band-pass domains to measure the difference in quality between reference and distorted videos. Furthermore, Lee et al. [Lee+21] conducted a subjective and instrumental assessment of the video quality of space-time subsampled videos. The ETRI-LIVE Space-Time Subsampled Video Quality (ETRI-LIVE STSVQ) database was created for the purpose and contains a total of 437 PVSs with framerates varying between 30 fps and 120 fps. The evaluation shows that the VSTR model proposed by Lee et al. [Lee+20], which is specifically developed to take into account the joint perceptual effects of spatio-temporal subsampling and compression, outperforms all the considered SoA models including VMAF.

For quality assessment of live streaming videos, Shang et al. [Sha+22] present a study with a particular focus on high motion live streaming videos. This included an assessment of a total of 315 sequences derived from 45 SRCs. The source sequences were subjected to six different distortions namely, compression (H.264 encoding), aliasing, judder, flicker, frame drops, and interlacing. The evaluation also includes a comparison of different NR and FR models for the particular use case. Furthermore, the developed dataset is made publicly available.

2.2.1.2 Crowdsourcing Approach

Crowdsourcing as a viable alternative to lab-based tests for perceptual quality assessment of both images and audiovisual content has garnered considerable attention in recent years [Hos+17; SB19; Hos+20]. Consequently, several studies have investigated the applicability and reliability of crowdsourcing for perceptual quality assessment of audiovisual content. In this section, a brief overview of best practices in crowdsourcing, frameworks for conducting such studies, and using them for quality assessment will be presented.

Two important aspects of conducting crowdsourcing tests are selecting the appropriate crowdsourcing framework and ensuring the reliability of the obtained results. A number of crowdsourcing frameworks have been proposed in the literature [Kei+12; Rib+11], and Hoßfeld et al. [Hoß+14a] provided a survey of different web-based

crowdsourcing frameworks for subjective quality assessment. With the goal of ensuring validity and reliability of the results, different studies have analyzed the results of crowd tests, or recommend a set of best practices for crowdsourcing QoE testing [Hoß+14b]. To evaluate the reliability of the crowdsourcing paradigm, in the best case, a comparison of crowd results and lab tests is performed, e.g. as it has been done in [FAK13; Sha+14a; Kei+12].

For the assessment of video quality, Hoßfeld et al. [Hoß+11] propose a generic subjective QoE assessment methodology for multimedia applications based on crowdsourcing. They conclude that crowdsourcing is a highly effective method not only for QoE assessment of online videos but also for other current and future internet applications.

A study on the usage of crowdsourcing for subjective quality assessment in the HAS context was conducted by Shahid et al. [Sha+14a]. Here, the results of the crowdsourcing test showed a strong correlation with the corresponding lab test. Similarly, Rainer and Timmerer [RT14] conducted a crowdsourcing study in the HAS context with the objective of comparing QoE performance of different HAS-based web clients namely, YouTube, DASH-JS and dash.js. Rainer and Timmerer conclude that the delivered representation bitrate and the number of stalls are the main influencing factors of QoE, as can also be confirmed by lab-based studies [Rob+18b].

In addition, crowdsourcing has been used to create large datasets annotated with human ratings. A few examples are the Konvid-1K database by Hosu et al. [Hos+17] which consists of 1200 public-domain video sequences sampled from YFCC100m, containing a very small number of high-quality videos, and the LIVE-VQC dataset by Sinno and Bovik [SB19], consisting of 585 videos with 240 recorded human ratings per video.

Notably, Seufert and Hossfeld [SH16] conducted a crowdsourcing study to test the limits of crowdsourced subjective video quality testing. They investigated the extreme case of presenting only a single test condition with a stimulus duration of 10 s to each subject (i.e. fully corresponding to a between-subjects test design) and the possibility of using such a simple “one-shot” design with a large number of subjects instead of using sophisticated test designs in crowdsourcing. The results suggest that when training effects are negligible, the extreme case of the “one-shot”

design seems to be applicable. In this study, source videos of 1080p were downscaled to 576p to meet the possibly low internet connections of the crowd users.

Moreover, crowdsourcing has also been widely used in the perceptual assessment of image quality, and in creating large image datasets annotated with human ratings. Ghadiyaram and Bovik [GB16a] designed and created the “LIVE in the Wild” image quality challenge database consisting of 1162 images rated by over 8100 unique observers. In addition, Hosu et al. [Hos+20] created an image database consisting of 10073 images scored in terms of quality by 1459 crowd users. On the other hand, Bosse et al. [Bos+16] investigated the feasibility of patch-based image quality assessment and found that humans can evaluate perceived quality on patch size of 128×128 pixels from a source image of 512×512 pixels.

In addition to image and video quality assessment, crowdsourcing has been used in other multimedia applications such as image annotation [NR10; Ras+10], video summarization [SDF12; TB12], speech quality assessment [Nad+20] and visual attention [Leb+15].

Although crowdsourcing was widely used for subjective image and video quality assessment, most research was focused on low-quality/-resolution content, due to issues such as lack of control on the display device, low bandwidth connections of crowd users, etc. Hence, there is a clear lack of crowdsourcing methods and also studies for quality assessment of high-quality/-resolution videos.

2.2.2 Video Quality Models

As the focus of the thesis is the development of bitstream-based and hybrid quality models for the prediction of video quality, the SoA survey will focus only on these two model types. For an overview of other pixel-based models, the reader is referred to, for example [Gör+21a].

2.2.2.1 Bitstream Models

As mentioned in Chapter 1, bitstream models consists of three different types, namely, Mode 0, 1 and 3. It should be noted that a detailed survey of Mode 2 models is not

considered because with today's encrypted traffic, this model variant has become mostly obsolete. This section presents a detailed overview of the SoA models of these three types.

Mode 0 This section briefly summarizes the SoA of Mode 0 models. A Mode 0 model has access to metadata such as bitrate, resolution, framerate and codec for video quality estimation. The most notable Mode 0 model for quality monitoring of video streaming is the ITU-T P.1203.1 Mode 0 model [ITU19e]. As mentioned before, this model is applicable for H.264 encoded videos for resolutions of up to 1080p and framerates up to 30 fps. A first extension of this model for newer codecs such as H.265 and VP9 was provided by a proprietary implementation from TU Ilmenau which has been made publicly available¹ [Raa+17; Rob+18a]. This extension used VMAF scores as ground truth to derive the mapping coefficients for the newer codecs.

Furthermore, Rao et al. [Rao+19b] propose an extension of this model to newer codecs such as H.265, VP9, AV1, and also for videos up to a resolution of 4K/UHD-1 and framerate up to 60 fps. However, this extension was based on only two subjective tests with limited encoding settings unlike the original standardized Mode 0 model in ITU-T Rec. P.1203, which was developed based on a large-scale dataset containing 17 training and 13 validation databases. In addition to this, Lebreton and Yamagishi [LY19] have also extended the application scope of the ITU-T P.1203.1 Mode 0 model for H.265 encoded videos for resolution up to 4K/UHD-1.

To shorten the development time and the associated subjective quality assessment tests needed for such newer extensions, Yamagishi et al. [Yam+21] proposed a generic method to derive coefficients for metadata-based models for adaptive bitrate streaming services. The proposed method uses full-reference model scores as ground truth to estimate the new coefficients.

Mode 1 A Mode 1 model has access to frame type and frame size information for quality estimation, in addition to metadata such as bitrate, resolution and fram-

¹<https://github.com/Telecommunication-Telemedia-Assessment/itu-p1203-codecextension>

erate, as for Mode 0 models. This additional access to the frame type and frame size information allows the quality estimation process to be content-dependent to a certain extent. As with the Mode 0 model, the ITU-T Rec. P.1203.1 Mode 1 model [ITU19e; Raa+17; Rob+18a] is the first standardized model of this type for the HAS scenario and has been trained on the same 17 databases and validated on the same 13 databases as the Mode 0 model.

Another example of a Mode 1 model is the Bitstream-based Quality Prediction of Gaming Video (BQGV) [Zad+20b]. It has been developed along the lines of P.1203.1 Mode 1. It takes a multi-dimensional approach to quality modeling, where the model consists of quality dimensions such as video discontinuity, video fragmentation, and video unclearness. Video discontinuity is related to the degradation caused due to the variation in framerate, while video fragmentation is mainly due to the chosen bitrate. Furthermore, video unclearness is the impairment that is a result of the scaling of the encoded video to the display resolution. The model was developed with gaming video quality estimation as the main focus and its efficacy is to be tested for traditional 2D video quality estimation. As it is the case for the P.1203.1 Mode 1 model, too, this model is applicable to videos of resolutions up to FHD (1920×1080 pixels).

Mode 3 A Mode3-type model has complete access to the bitstream for estimating video quality. Many early Mode 3-type models that have been proposed were mainly focused on non-reliable transport and lower resolutions ($< 1080p$) [Raa+08; ITU07; GSR10; GR11; Lin+12; Gar+13; SRL13; RL08; SRL11; Moc+15; DG17]. Hence, these models include degradations due to packet loss, besides coding- and resolution-related effects.

One of the first Mode 3-type models that focused on reliable transport is the extension of P.1201.2 for progressive download for H.264 encoded videos. As with H.264 encoded videos, Izumi et al. [Izu+14], developed a Mode 3-based model using QP and spatial features based on coding units to estimate the quality of H.265 encoded videos [Sul+12] based on bitstream information. Also, Huang, Sogaard, and Forchhammer [HSF15] proposed an approach to estimate the quality of H.265 encoded videos in terms of PSNR that can be used either as a bitstream-based or a pixel-based method. The model includes QP and transform coefficients as

features and has been trained on the LIVE dataset [Ses+10] and validated on the SJTU dataset [Son+13].

ITU-T Rec. P.1203 [ITU16b] is the first standardized model for a holistic evaluation of HAS-type video streaming. This recommendation consists of three different modules corresponding to video quality [ITU19e], audio quality [ITU17], and the overall integral quality [ITU20]. The video quality models in ITU-T Rec. P.1203.1 [ITU19e] are further divided into four different modes of operation, depending on the input information available for quality estimation, namely, Mode 0, 1, 2, and 3 [Raa+17]. These models have been specifically developed for the HAS scenario and are applicable for videos encoded with H.264 for resolutions up to 1080p and framerates up to 30 fps. The reference implementation of this model is publicly available² [Rob+18a]. The Mode 3 model corresponding to the standard ITU-T Rec. P.1203.1 has been extended to be applicable to H.265 encoded videos of resolution up to 4K/UHD-1 by Lebreton and Yamagishi [LY19].

Furthermore, He et al. [He+18] present a model for quality assessment of H.264 and H.265 encoded bitstreams. This model uses QP, skip ratio, motion information, bitrate, and framerate as features and shows performance comparable to the ITU-T Rec. P.1203.1 Mode 3 model. In addition, early models for reliable transport and HAS have been proposed by [SHR12; Tra+16a]. The different approaches and models related to holistic QoE evaluation where the cumulative effects of HAS-specific distortions such as momentary audio and video quality and quality switches, and stalling on quality perception are included, will be discussed in Section 2.3.

2.2.2.2 Hybrid Models

A hybrid model has access to both pixel and bitstream information for estimating video quality. Similar to pixel-based models, hybrid models can be classified into different categories depending on the access to the reference video for quality estimation. These include hybrid-FR, hybrid-RR, and hybrid-NR models which have complete, partial, and no access to the reference video, respectively. Furthermore, each of the categories can be divided into Mode 0, 1, and 3 based models, depending on the amount of bitstream information available as input.

²<https://github.com/itu-p1203/itu-p1203>

Yamagishi, Kawano, and Hayashi [YKH09] present a hybrid-NR model for the IPTV scenario using information from packet headers and pixel-based spatial and temporal information for quality estimation [ITU99]. The model is applicable for H.264 encoded videos of resolutions up to 1440×1080 and framerates up to 30 fps.

Another example of a hybrid model for non-reliable transport is the model proposed by Farias et al. [Far+11]. This model estimates blockiness and blurriness with the pixel information, which are then combined with the packet loss rate information to predict video quality. Like the model presented in [YKH09], this model, too, is applicable only for H.264 encoded videos, in light of the video technology primarily used at the time. Similarly, the ITU-T J.343 series of recommendations also propose standardized hybrid models of all types, for the case of non-reliable transport.

Moreover, Osamu et al. [Osa+09] propose a mode 3 hybrid-NR model where the QP is used as the bitstream feature, along with the pixel-based spatial and temporal information to calculate video quality [ITU99]. This model is again restricted to videos encoded with H.264 only.

More recently, hybrid models have been developed also for the HAS scenario. One example is the recently standardized ITU-T Rec. P.1204.5 which is a Mode 0 hybrid-NR model. It was developed as part of the same modeling competition as the bitstream-instance of *AVQBits*, ITU-T Rec. P.1204.3 which was developed as part of this thesis. Like all P.1204 models, the P.1204.5 model is applicable to videos encoded with H.264, H.265 and VP9 with resolutions up to 4K/UHD-1 and framerates up to 60 fps.

Another Mode 0 hybrid-NR model called “hyfu” has been developed by Göring et al. [Gör+21a] as part of a larger framework for pixel-based video quality models using machine learning. Accordingly, at its core, “hyfu” is a random forest (RF) based model. The model has been trained on four databases and validated independently on the four tests of the AVT-VQDB-UHD-1 database [Rao+19a]. The application scope of “hyfu” is the same as ITU-T Rec. P.1204.5.

2.3 Overall Integral Quality of a HAS session

This section details the different subjective studies and quality models presented for the estimation of the overall QoE of a HAS session reported in literature.

2.3.1 Subjective Studies

In addition to video compression-related distortions, a typical HAS session is also affected by rebuffering-related events in the form of initial loading delay (ILD) and stalling events. Hence, it is important to assess how these factors are perceived by humans in terms of overall quality perception. As a result, a number of studies have been reported in the literature conducting subjective tests for the overall quality assessment of an HAS session.

A notable study is based on the ITU-T P.1203 standardization project where 30 different databases were created for quality assessment of audiovisual content in a HAS scenario [Raa+17; Rob+18a]. These databases were created with the goal of developing quality models capable of estimating the short-term video quality, audio quality, and overall integral quality of a HAS session. This resulted in the P.1203 series of ITU recommendations [ITU19e; ITU17; ITU20]. The factors that were varied in these databases consisted of the sequence duration (1min to 5min), quality switches, stalling events, and the used display device (PC/TV and mobile). Of the 30 databases, four have been made publicly available [Rob+18a].

Furthermore, Bampis et al. [Bam+18] conducted a study to assess the effects of aspects such as bitrate adaptation algorithms, network conditions, and video content on the overall QoE with the aim of perceptually optimized end-to-end adaptive video streaming. The resulting dataset has been made publicly available for further research.

In addition to assessing the cumulative effect of quality switches and stalling events on perceived quality, various studies have analyzed the impact of these factors separately. A study on the impact of ILD and stalling on the overall perceived quality was presented by Duanmu et al. [Dua+17]. Extending this work, the same authors evaluated the effect of quality switches (direction of switch, intensity of

switch, duration between switches etc.) on the overall quality without any stalling events [DMW17]. Furthermore, the authors investigated the cumulative effect of both stalling events and quality switches on the overall perceived quality in their subsequent study [DRW18]. Although these studies perform a systematic investigation of the effects of different factors related to adaptive streaming on the overall perceived quality, the duration of the sequences used is short (10s) and hence the question of the impact of video duration on the perceived quality is not addressed. The datasets resulting from all three studies have been made publicly available.

Although the aforementioned evaluations investigate the HAS-related factors on perceived quality, the highest resolution of the videos considered is restricted to 1080p and also the range of the factors addressed represents the user preference at the time of the conducted subjective evaluation. Further studies with videos of higher resolution such as 4K/UHD-1 and updated user preferences are needed to understand the effects of quality switching and stalling on overall QoE. As part of this thesis, videos of 4K/UHD-1 are considered for quality assessment of long-duration videos in a HAS session.

2.3.2 Quality Models

In general, a typical HAS session is characterized by various factors such as initial loading delay, momentary audio, and video quality and quality switches, and stalling as illustrated in Figure 1.3. A holistic QoE evaluation model has to consider all these factors, while also taking into account the time at which these changes occur in a video viewing session (see also [Wei+14]). ITU-T Rec. P.1203.3 [ITU20; RGR17] is the first standardized model that incorporates all these factors. Here, ITU-T Rec. P.1203.1 and P.1203.2 are used to compute the video and audio quality, respectively, of each segment at a per-second level. In the integration module P.1203.3, the per-second audio and video quality values are further aggregated with regard to their time of occurrence, the longest quality change, and the total number of quality changes, to obtain the final audiovisual quality of the video. A second component called “stalling quality” that handles the impact of initial loading delay and stalling is computed using the number of stalls, average stalling duration, and average interval between stalls as features. Then, the overall audiovisual quality and the stalling

quality are integrated to obtain the initial overall quality. Besides a parametric, curve-fitting-based model component, an additive RF-based component is used to compute the overall quality using features such as per-second video and audio quality scores, stalling ratio, stalling frequency, duration before the last stalling event etc. The final overall integral quality is the convex linear combination of the initial overall quality and RF-based overall quality. This model is applicable to videos of durations between 1 and 5 minutes and the implementation is publicly available³.

As the ITU-T Rec. P.1203 model only covers H.264 encoded videos, Lebreton and Yamagishi [LY19] have further extended ITU-T Rec. P.1203 for H.265 encoded videos of resolutions up to 4K/UHD-1. For this purpose, six subjective tests with varying encoding conditions involving up to 192 participants in total were used.

In addition to this, other models for holistic QoE evaluation have been proposed [Tra+16a; Tra+16b], but unlike the ITU-T Rec. P.1203.3, these models have not been trained and validated on large-scale databases.

2.4 Summary and Conclusion

As described in this chapter, various studies have been conducted for perceptual quality assessment of videos of 4K/UHD-1 resolution and also for overall QoE assessment of a HAS session. Some of the conducted research has also resulted in publicly available databases that can be used to train and benchmark existing video quality models and also develop newer quality models for estimating the quality of high resolution. However, these datasets cannot be considered large-scale subjective studies in terms of the number of different source contents and also the number of encoded sequences used. Also, several of these datasets do not provide access to the bitstreams and hence not suitable for this thesis work. Also, the number of encoding parameters tested as part of the reported evaluations is limited. In addition to this, a number of models both bitstream and hybrid, have been proposed in the literature for predicting video quality. However, these models have a limited scope in terms of the applicability considering video resolution and the different codecs. Furthermore, there is a lack of studies focusing on assessing the applicability of out-of-the-lab

³<https://github.com/itu-p1203/itu-p1203>

Chapter 2 State of the Art

testing method for quality assessment of high-resolution videos ($> 1080p$). The identified limitations with datasets, subjective test method and models are part of the research questions of this thesis and will be handled in the subsequent chapters.

Subjective Quality Assessment of 4K/UHD-1 Videos

Subjective testing is considered a gold standard in multimedia quality assessment. Hence, subjective studies play an important role in investigating the impact of the different degradations that are encountered in a typical HAS session on the perception of the end-user. These tests can either be conducted in a well-controlled lab setting or in an out-of-the-lab setting with less control over the testing environment but closer to real-life settings. The impact of the different influence factors on the final perceived quality can either be investigated by considering each degradation individually or by considering the cumulative effect of the different degradations. In this thesis, subjective tests have been conducted both in a well-controlled lab setting and in an out-of-the-lab setting. The out-of-the-lab tests are conducted using two different methods. The first method was an online study approach in which participants were recruited from among the university student and staff body and were not compensated. The second method was a crowdsourcing-based approach using a crowdsourcing framework to recruit participants and also compensate them. The first part of the chapter focuses on tests conducted in a lab setting following standard recommendations. Firstly, the influence of the video encoding settings on short-term video quality ($\approx 7-10$ s) is assessed in a series of subjective tests. Following this, the effect of quality changes, initial loading delay and stalling events on the overall perceived quality is evaluated with various audiovisual sequences ranging between 1 min to 5 min duration.

The second part of the chapter presents tests conducted in an out-of-the-lab setting. For this, an approach to assess both short-term video quality and the overall quality

of a HAS session in an out-of-the-lab setting considering videos of high resolution ($> 1080p$) is proposed. This is done with the aim of answering research question 4 as defined in Chapter 1. Using the proposed approach, two separate tests, one each focusing on short-term video quality and the overall integral quality of a HAS session conducted in an out-of-the-lab setting are described in detail. In addition to this, an analysis of the agreement of the proposed approach in comparison with traditional lab tests is performed. These studies are conducted in both well-controlled lab settings and out-of-the-lab settings using online testing and crowdsourcing.

It should be noted that the analyses of the results of the subjective test presented in this chapter are limited to the overall MOS distributions. No further benchmarking of the codecs or subjective test results analysis in terms of user-specific rating behaviour is conducted because the primary objective of designing and conducting tests was to gather ground truth for model development and use it further for comparison of performances of the proposed and the SoA models.

This chapter is based on the following publications:

[Rao+19a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Werner Robitza, Bernhard Feiten, and Alexander Raake. “AVT-VQDB-UHD-1: A Large Scale Video Quality Database for UHD-1”. In: *21st IEEE International Symposium on Multimedia (IEEE ISM)*. Dec. 2019

[Raa+20a] Alexander Raake, Silvio Borer, Shahid Satti, Jörgen Gustafsson, **Rakesh Rao Ramachandra Rao**, Stefano Medagli, Peter List, Steve Göring, David Lindero, Werner Robitza, Gunnar Heikkilä, Simon Broom, Christian Schmidmer, Bernhard Feiten, Ulf Wüstenhagen, Thomas Wittmann, Matthias Obermann, and Roland Bitto. “Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204”. In: *IEEE Access* 8 (2020)

[RGR21b] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Towards High Resolution Video Quality Assessment in the Crowd”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 2021

[RGR22] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications”. In: *IEEE Access* 10 (2022)

[GRR23] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Quality Assessment of Higher Resolution Images and Videos with Remote Testing”. In: *Quality and User Experience (QUEX) 8* (2023)

3.1 Lab-based Subjective Quality Assessment

In this section, the design and results of the short-term video quality and overall integral quality of a HAS session tests conducted during the course of this work are described in detail.

The details of the protocol followed and instructions provided during these tests are described in Appendix A.

3.1.1 Short-term Video Quality Assessment

Two sets of lab-based subjective tests consisting of a total of eight different subjective tests have been performed. These resulted in two datasets, namely, AVT-PNATS-UHD-1 and AVT-VQDB-UHD-1. Hence, this section is further organized in terms of these two sets.

For all the lab-based tests, the test environment in terms of lighting, curtains and viewing distance follows ITU-R Rec. BT.500-13 [ITU14b] in order to ensure a controlled test environment and also to guarantee repeatability of the tests. Two different screens were used during the course of these subjective assessment studies, namely, a 65" 4K Panasonic VIERA TX-65CXW804 display and a 55" 4K LG OLED55C7D screen. Furthermore, to ensure a seamless playback, an interface to a DeckLink 4K Extreme 12G card was used.

All ratings were collected using the AVRateNG¹ tool. As the test method, 5-point absolute category rating (ACR) [ITU14b] was used in all eight tests. Prior to the ACR-based subjective test, every test participant underwent a visual acuity test using Snellen charts. In total, each test lasted approximately 60 minutes with two optional 5 min breaks in between. Each test included a short training phase in which

¹<https://github.com/Telecommunication-Telemedia-Assessment/avrateNG>

5 different videos spanning the entire quality range were shown to the participants to get them acquainted with the test procedure and the expected quality range. For all the tests, the outlier detection criterion was based on a threshold of 0.75 Pearson Correlation Coefficient (PCC) [ITU16b; Raa+17].

In addition to AVT-PNATS-UHD-1 and AVT-VQDB-UHD-1, the author of this thesis was actively involved in the design of a third dataset, namely, PNATS-UHD-1 which encompasses the 26 different subjective tests designed and conducted as part of the “P.NATS Phase 2” modeling competition in ITU-T SG12/Q14. As outlined above, the “P.NATS Phase 2” competition resulted in the ITU-T P.1204 [ITU19a] series of Recommendations. The respective dataset is also described in this section, as it is used to train and validate the ITU-T P.1204.3 [ITU19b] model which forms a major component of this thesis (cf. Chapter 4). It should be noted that the AVT-PNATS-UHD-1 dataset is a subset of the PNATS-UHD-1 dataset.

Furthermore, the “AV1 dataset” created to extend the ITU-T Rec. P.1203.1 Mode 0 model for newer codecs and higher resolutions and framerate is described in this section.

3.1.1.1 PNATS-UHD-1

This dataset consists of 26 different tests that were designed and conducted by nine proponents. Of the 26 different tests, 13 were used for training the models submitted to the competition and the remaining 13 were used for model validation. It should be noted that the 13 validation databases were created after model submission. Out of the 13 training tests, nine were created with a PC/TV as the viewing device and four with mobile devices for viewing. For validation, nine tests used a PC/TV as viewing devices, three a mobile and one a tablet. For the PC/TV tests, the considered display resolution was 4K/UHD-1 (3840×2160) with a viewing distance of 1.5H [ITU14b]. For Mobile/Tablet (MO/TA) tests, a display resolution of 2560×1440 was used with a viewing distance of 5-7H [ITU14b]. The viewing distances for both device classes follow the recommendation in ITU Rec. BT.500-13 [ITU14b]. All tests included PVSs with a duration of 7–9 s. As a result of all tests, 2464 PVSs were used for training, and 2483 for validation, resulting in a total number of 4947 PVSs.

3.1 Lab-based Subjective Quality Assessment

A detailed overview of the used source contents is provided in Section 3.1.1.1 and about the encoding parameter ranges and the test design are given in Section 3.1.1.1. It should be noted that these details are sourced from Raake et al. [Raa+20a], which is co-authored by the author of this thesis as well as by the other proponents involved in the competition.

Source Contents: In the process of gathering source contents, as a first step, 4K/UHD-1 *source footages* that are publicly available and provided by some of the proponents (TU Ilmenau, Yonsei University, and Ericsson AB) were collected. In this step, in addition to the 4K/UHD-1 footage, 1440p *source footage* was also considered for databases that were planned to be run on Mobile or Tablet. From these different *source footages*, cuts of 7-9 s duration were defined and created. These cuts were then manually reviewed and some which have a scene cut either in the first 2 or last 2 seconds were rejected. The selected video cuts after this review process formed the sources (SRCs) for the training and validation phases. Three SRCs were chosen to be used in both training and validation databases to create the “common set PVSs”. The overall number of unique *source footages* and SRCs used to generate the training and validation databases are summarized in Table 3.1.

Table 3.1: Number of unique footages and SRC files used in the training (TR) and validation (VL) databases in the P.NATS Phase 2 competition [Raa+20a].

		TR	VL	TOTAL
Footages	50/60 fps	27	20	43 (4 common TR/VL)
	24/25/30 fps	32	97	129
	Total	59	117	172 (4 common TR/VL)
SRC files	50/60 fps	203	79	278 (4 common TR/VL)
	24/25/30 fps	138	294	432
	Total	341	373	710 (4 common TR/VL)

The SRCs that were chosen had a wide range of spatial and temporal complexity. This spatial and temporal complexity is characterized in terms of the SI and TI measures respectively as specified in ITU-T Rec. P.910 [ITU99] and is illustrated in Figure 3.1. The reason to select such a wide range of content is, to ensure that the videos used for the datasets are realistic for common TV/streaming content. Moreover, the focus

was on using pristine quality professional content and excluded the user-generated content.

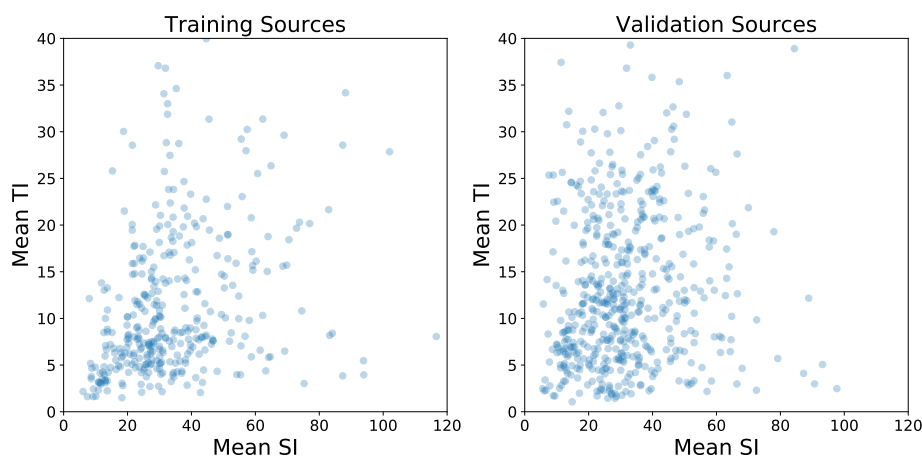


Figure 3.1: SI-TI of all the sources used in training and validation in the P.NATS Phase 2 competition.

Test Design: The hypothetical reference circuits (HRCs) that have been used for the tests were created by varying different encoding parameters. The parameters that were considered were as follows: video codec, resolution, bitrate, framerate, encoding presets/speed, GOP size, encoder implementations, chroma subsampling, bit-depth, encoding types, and bitstream container. The ranges of these different parameters that were used are reported in Table 3.2. In addition, the bitrate ranges used for different encoders for each resolution are depicted in Figure 3.2. A detailed per-database test plan is provided in Appendix B.

As shown in Table 3.2, in addition to the FFmpeg-based encoder implementations, three different online services, namely, Youtube, Bitmovin, and Vimeo were considered to produce the encoded bitstream. The HRCs corresponding to these services was termed as “online conditions”. For these cases, SRCs were uploaded to these services and the corresponding bitstreams were then downloaded. In the case of YouTube and Vimeo, no encoding parameters were allowed to be specified whereas, for Bitmovin, specific input parameters could be specified. However, for all three services, the actual encoding process was unknown. The reason to select such online cases was to simulate real-world encodings to enable the developed models to be capable of handling such conditions.

3.1 Lab-based Subjective Quality Assessment

Table 3.2: Parameter ranges considered in the P.NATS Phase 2 competition [Raa+20a].

Parameter	Range
Video Codec	H.264, H.265, VP9
Encoded Resolution	TV/Monitor: 640×360 – 3840×2160 , Mobile/Tablet: 426×240 – 2560×1440
Framerate	15, 24, 25, 30, 50, 60 frames per seconds
Presets	H.264/H.265: online, i.e. Youtube, Bitmovin or Vimeo; medium, ultrafast, fast, veryfast, slower, slow, veryslow. VP9: speed presets 0, 1, 2, 3, 4
GOP Size	Auto, 2, 5 seconds
Encoder Implementation	H.264: libx264 (FFmpeg), H.265: libx265 (FFmpeg), VP9: libvpx-vp9 (FFmpeg), YouTube, Bitmovin, Vimeo
Chroma Subsampling	YUV420, YUV422
Bit-depth	8,10 bits
Encoding Types	1-pass, 2-pass (with and without min max bitrate constraints), Constant rate factor (CRF) encoding. Unknown encoding recipes employed by YouTube, Vimeo, Bitmovin
Bitstream Container	mp4, webm, mkv

These HRCs were then coupled with different SRCs and assigned to individual databases by random sampling of the bitrate ranges. In this process, it was ensured that roughly equal representations of different codecs, resolutions, and framerates were mapped to individual databases. After the HRCs were defined for each database, these were processed with different SRCs and manually checked for quality distribution to enable having PVSs covering the entire range of the 5-point ACR scale. In addition to ensuring that HRCs are assigned based on the aforementioned criteria, SRCs were also assigned to individual databases in a manner so as to balance the content complexity of the SRCs used in each database. For this purpose, a content complexity measure based on CRF encoding using H.264 codec was defined (more details are described in [Raa+20a]). Using this measure, the SRCs were classified into four different complexity categories ranging from 0 to 4. Based on the complexity of

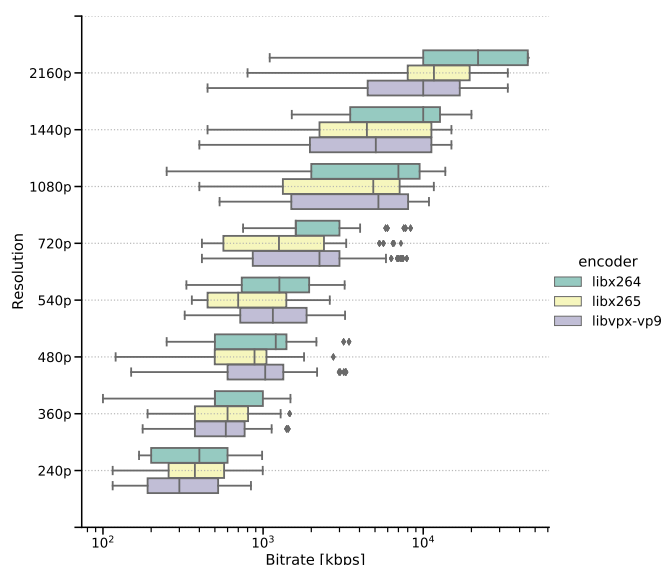


Figure 3.2: Bitrate ranges for each encoder–resolution pair used in the P.NATS Phase 2 competition [Raa+20a].

the SRC, the encoding bitrate was adjusted. One other aspect that was considered while creating HRCs was that none of the PVSs generated would either have resolution and framerate higher than that of the used SRC. The videos were encoded using a dedicated processing chain².

Furthermore, to have a way to normalize the subjective results across different subjective tests, anchor conditions were defined. For this purpose, 5 common HRCs in combination with 3 common SRCs were used in each database. The highest and lowest anchors were adjusted considering the display devices as the subjective expectations on these different device types differ. Hence, for the case of PC/TV, the highest anchor had a resolution of 4K/UHD-1 whereas for the Mobile/Tablet case, it was 2560×1440 . Similarly, the lowest anchor for the PC/TV scenario was 640×360 at a framerate of 24/25/30 fps (the fps was source dependent) and for Mobile/Tablet, it was 426×240 at a framerate of 15 fps. The MOS range for the common HRCs across different databases is reported in Table 3.3.

The training and validation databases in terms of the average confidence interval (Avg. CI), average correlation (Avg. Correlation), target display (Display), number

²<https://github.com/pnats2avhd/processing-chain>

3.1 Lab-based Subjective Quality Assessment

Table 3.3: Common HRCs used in the P.NATS Phase 2 competition. The video codec is H.264 for all common conditions [Raa+20a].

HRC-ID	Resolution	Bitrate (kbps)	FPS	MOS Range	
				PC/TV	MO/TA
HRC0001	240p	100/200	15	-	1.167 - 2.476
HRC0115	360p	300/500	24/25/30	1.160 - 2.917	1.792 - 3.571
HRC0388	720p	800/1600	50/60	1.500 - 3.917	2.833 - 4.542
HRC0436	1080p	3500/7000	50/60	2.958 - 4.833	3.833 - 4.810
HRC0484	1440p	6000/10000	50/60	3.333 - 4.875	4.083 - 4.762
HRC0571	2160p	30000/45000	50/60	3.667 - 5.000	-

of test participants (N), and the number of PVSs (PVSs) rated by each participant are summarized in Tables 3.4 and 3.5.

Table 3.4: Training database details [Raa+20a].

DB-ID	Display	N	Avg. Correlation	Avg. CI	PVSs
P2STR01	Mobile	26	0.82	0.29	203
P2STR02	Mobile	24	0.87	0.27	199
P2STR03	Mobile	30	0.87	0.23	200
P2STR04	PC	26	0.91	0.24	199
P2STR05	PC	26	0.84	0.27	187
P2STR06	Mobile	24	0.82	0.25	187
P2STR08	TV	24	0.89	0.26	179
P2STR09	PC	25	0.86	0.25	187
P2STR10	PC	34	0.86	0.21	187
P2STR11	TV	24	0.89	0.25	187
P2STR12	PC	24	0.85	0.28	183
P2STR13	TV	25	0.87	0.25	187
P2STR14	TV	24	0.84	0.24	179

A dedicated screening process was used to remove some training PVSs due to bad content or wrong encoding settings. The total number of training and validation PVSs after the screening process was 2464 and 2483 respectively.

The details pertaining to the competition structure and the statistical evaluation of the models will be presented in Chapter 4.

Table 3.5: Validation database details [Raa+20a].

DB-ID	Display	N	Avg. Correlation	Avg. CI	PVSs
P2SVL01	TV	30	0.82	0.25	185
P2SVL02	Mobile	24	0.82	0.26	186
P2SVL03	Mobile	21*	0.82	0.30	186
P2SVL04	Mobile	24	0.88	0.28	195
P2SVL05	TV	25	0.87	0.28	194
P2SVL06	TV	24	0.89	0.26	191
P2SVL07	TV	25	0.86	0.26	188
P2SVL08	PC	27	0.82	0.29	195
P2SVL09	TV	28	0.81	0.28	191
P2SVL10	TV	26	0.86	0.21	195
P2SVL11	TV	24	0.87	0.27	195
P2SVL12	Tablet	24	0.84	0.20	195
P2SVL13	TV	26	0.84	0.25	187

* Extra subjects were removed from this database due to file copying bugs. Database was kept since correlation and CI was deemed ok after extensive analysis.

3.1.1.2 AVT-PNATS-UHD-1

Four out of the 26 training and validation databases from the “P.NATS Phase 2” competition that were assigned to TU Ilmenau for conducting subjective tests form the AVT-PNATS-UHD-1 dataset. In the following, these four tests are described in detail.

The tests were targeted to cover a wide range of source contents and hence more than 50 source contents were used in each of the four tests. Due to a large number of source contents, the tests are not full-factorial in their design as that would result in an infeasible number of PVSs that would have to be assessed by participants. An SRC was repeated between 3 and 5 times within a test. Three sources were used across all tests and are referred to as “common sources”. A 55" 4K LG OLED55C7D screen was used to present the videos in all four tests.

In the first test, 52 different SRCs were included and encoded with different HRCs which resulted in a total of 187 PVSs. These 187 PVSs were rated by 27 participants. Following the outlier detection criterion based on PCC described earlier, two outliers were detected and removed from further analysis. The second test covered a total of 53 different SRCs with 187 PVSs created from these, which were rated by a total

3.1 Lab-based Subjective Quality Assessment

of 36 participants. Further analysis based on the aforementioned outlier criterion detected two outliers in this test. 52 different sources were used in the third test, and the 185 PVSs resulting from the HRC processing were rated by 30 participants, with five outliers detected. The fourth test had 53 SRCs processed according to different HRCs, resulting in 191 PVSs that were rated by 28 participants. Here, 3 outliers were detected following the PCC-based criterion.

The distribution of the mean opinion scores (MOS) is illustrated in Figure 3.3. It can be observed that there is a tendency towards higher quality. This test design was motivated to yield better distinction for higher quality levels by test participants and also models.

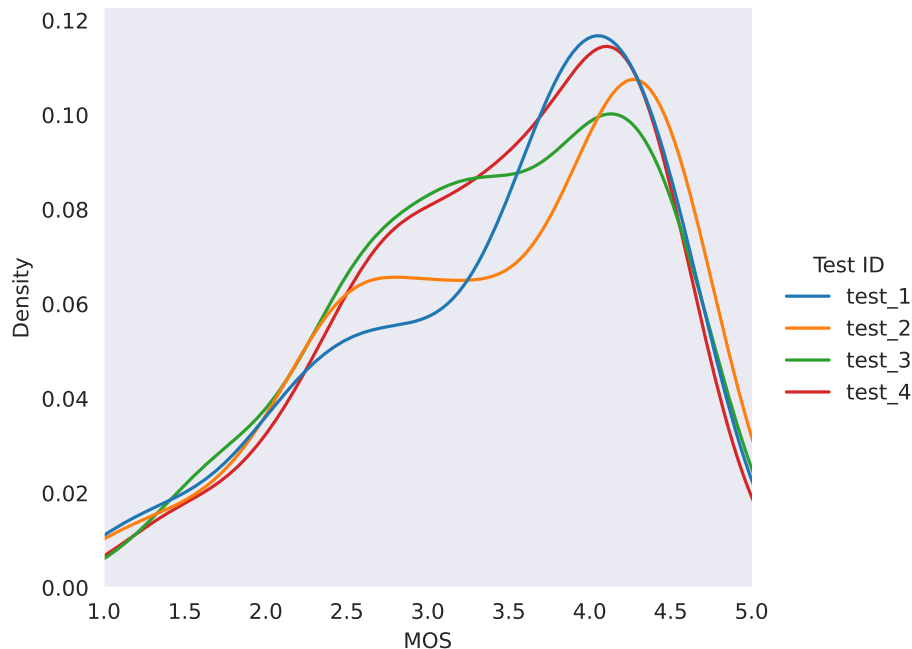


Figure 3.3: MOS distribution of AVT-PNATS-UHD-1 dataset.

The SOS analysis as described in [HSE11] was conducted for each of the four tests of the AVT-PNATS-UHD-1 dataset and the results of this are illustrated in Figure 3.4. The values of the SOS parameter “a” for all the tests are in the typical range reported in the SoA for video quality assessment studies.

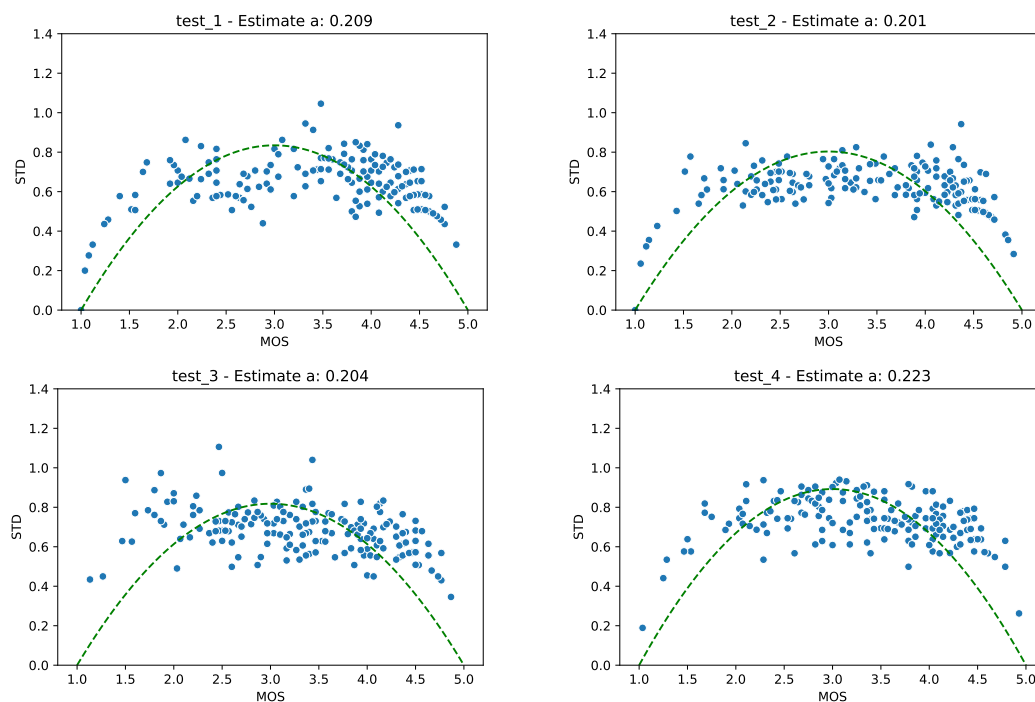


Figure 3.4: SOS analysis of the AVT-PNATS-UHD-1 dataset.

3.1.1.3 AVT-VQDB-UHD-1

Like the AVT-PNATS-UHD-1 dataset, this dataset also consists of four different subjective tests. All four tests had a full-factorial design. In total, 17 different SRCs with a duration of 7-10 s were used across all four tests. All the sources had a resolution of 3840×2160 pixels and a framerate of 60 fps. For HRC design, bitrate was selected in fixed (i.e. non-adaptive) values per PVS between 200 *kbps* and 40000 *kbps*, resolution between 360*p* and 2160*p* and framerate between 15 *fps* and 60 *fps*. In all the tests, a 2-pass encoding approach was used to encode the videos, with *medium* preset for H.264 and H.265, and the *speed* parameter for VP9 set to the default value “0”. As with the tests in AVT-PNATS-UHD-1, the same PCC-based criterion is used for outlier detection. Unlike the AVT-PNATS-UHD-1 dataset, this dataset is publicly available³.

³<https://github.com/Telecommunication-Telemedia-Assessment/AVT-VQDB-UHD-1>

3.1 Lab-based Subjective Quality Assessment

Source Contents: The thumbnails of the used source contents are depicted in Figure 3.5. The source contents for each of the four tests were selected based on their spatial and temporal complexities using spatial information (SI) and temporal information (TI) as described to ITU-T P.910 [ITU99]. These SI and TI scores were calculated using the publicly available implementation of SITI⁴. The distribution of SI and TI of all the source contents as shown in Figure 3.6 indicate that they cover a wide range of values. The duration of the contents used in the tests was between 8-10 seconds, and no quality-switching was used. This way, scores can be considered as “per-segment” scores for different DASH implementations, where segment sizes typically range from 1-15 s [Seu+15]. All the source contents have a framerate of 60 fps. More details about the sources used in the tests are summarized in Table 3.6. The "Shareable" column in the table indicates whether the video (source + encoded versions) is publicly available. In case "No" is stated, all the other features such as MOS and predicted scores by the considered quality models are publicly available.

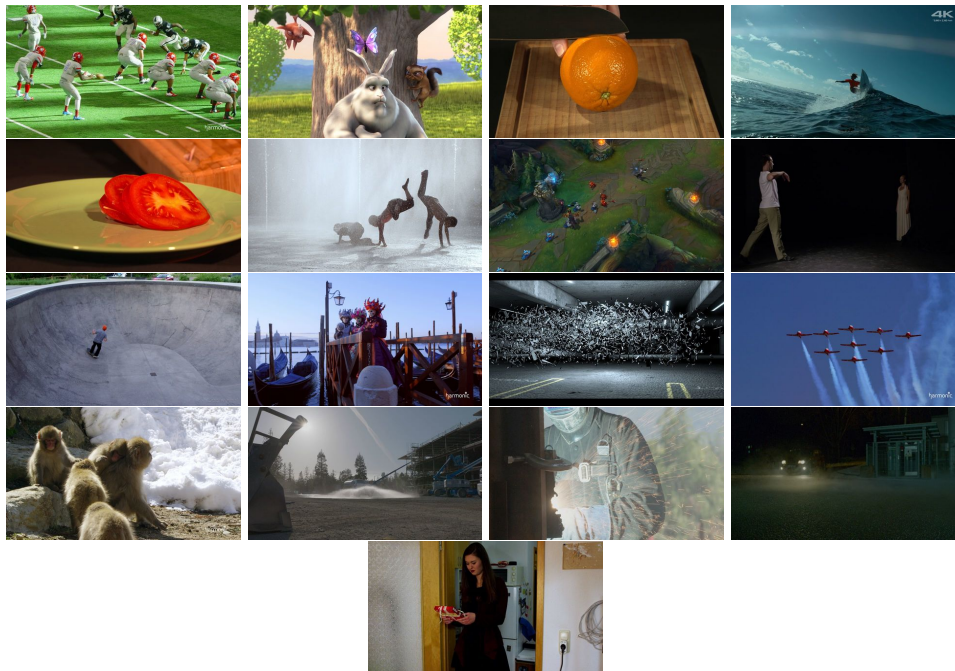


Figure 3.5: Thumbnails of source videos in the AVT-VQDB-UHD-1 dataset.

⁴<https://github.com/Telecommunication-Telemedia-Assessment/SITI>

Table 3.6: Source details for the AVT-VQDB-UHD-1 dataset.

Name	Duration	Details	Origin	Shareable
American Football (AF)	10 s	3840 × 2160, 60 fps	Undisclosed	No
Big Buck Bunny (BBB)	10 s	3840 × 2160, 60 fps	Blender Foundation [Ble]	Yes
Cutting Orange (CO)	10 s	3840 × 2160, 60 fps	TU Ilmenau	Yes
Surfing (SU)	10 s	3840 × 2160, 60 fps	Undisclosed	No
Vegetables (VE)	10 s	3840 × 2160, 60 fps	TU Ilmenau	Yes
Water (WA)	10 s	3840 × 2160, 60 fps	Netflix Inc.	Yes
League of Legends (LoL)	8 s	3840 × 2160, 60 fps	Private	No
Dancers (DA)	8 s	3840 × 2160, 60 fps	Netflix Inc.	Yes
Moment of Intensity (MoI)	8 s	3840 × 2160, 60 fps	Cable Labs	No
Venice (VN)	8 s	3840 × 2160, 60 fps	Undisclosed	No
fr-041-debris (FR)	8 s	3840 × 2160, 60 fps	NASA	Yes
Air acrobatics (AA)	8 s	3840 × 2160, 60 fps	Undisclosed	No
Monkeys (MO)	8 s	3840 × 2160, 60 fps	Undisclosed	No
Sparks (SP)	8 s	3840 × 2160, 60 fps	Netflix Inc.	Yes
Daydreamer (DD)	8 s	3840 × 2160, 60 fps	TU Ilmenau	Yes
Giftmord (GI)	8 s	3840 × 2160, 60 fps	TU Ilmenau	Yes

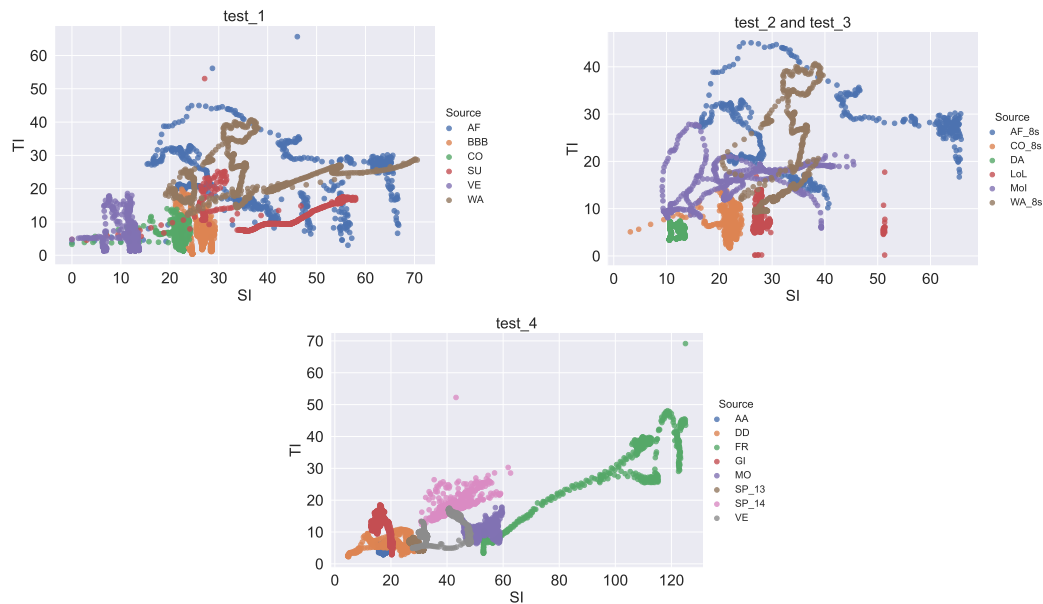


Figure 3.6: Spatial and temporal complexities SI, TI of all the video contents used in the AVT-VQDB-UHD-1 dataset.

3.1 Lab-based Subjective Quality Assessment

test_1 The HRC design of this test was based on choosing from different bitrates for each of the different resolutions. For this purpose, four different resolutions were considered, namely 360p, 720p, 1080p and 2160p. Two bitrates each were selected for 360p and 720p and three bitrates each for 1080p and 2160p. The detailed test design is presented in Table 3.7. Three different codecs, namely, H.264, H.265, and VP9 were used to encode the videos. These HRCs were applied to six different SRCs of 9-10 s duration. This resulted in a total of 180 PVSs. The framerate of all the PVSs was kept at the source framerate of 60 fps. A 65" Panasonic VIERA TX-65CXW804 display was used to present the videos in the test. The 180 PVSs were rated by 29 participants. Following the PCC-based outlier criterion, no outliers were detected.

Table 3.7: Test Design – test_1.

Resolution	Bitrate [kbit/s]			
360p	200	750		
720p		750	2000	
1080p			2000	7500 15000
2160p				7500 15000 40000

test_2 For this test, the HRC design was based on using different bits-per-pixel (*bpp*) settings for different resolutions. Four different *bpp* values were considered, per each of the same four resolutions used also in test_1. As the number of *bpp* – resolution combinations considered was higher than the bitrate – resolution combinations for test_1, only H.264 and H.265 were used to encode the videos. Details of the test design are described in Table 3.8. Six SRCs including the three common set SRCs from test_1 were used. The SRCs had a duration of 7-9 s. As in test_1, the framerate of the PVSs was kept at the source framerate of 60 fps. Overall, 192 PVSs were created using the six SRCs and the defined HRCs. The test videos were presented to the participants on a 55" LG OLED55C7D screen. A total of 24 participants rated these PVSs, and no outliers were detected.

test_3 This test followed the same philosophy for HRC design as test_2, and hence the same *bpp* values and resolutions were used. Also, the same SRCs were employed. Mainly H.265 and VP9 were selected to encode the videos. In test_2, it was observed that some of the PVSs associated with one of the sources (*Dancers_8s*) had

Table 3.8: Test Design – test_2 and test_3 (Bit-per-pixel based test).

Resolution	Bits-per-pixel (bitrate in kbps)			
360p	0.007 (97)	0.0447 (617)	0.0823 (1138)	0.12 (1659)
720p	0.007 (387)	0.0447 (2470)	0.0823 (4553)	0.12 (6636)
1080p	0.007 (871)	0.0447 (5557)	0.0823 (10244)	0.12 (14930)
2160p	0.007 (3484)	0.0447 (22229)	0.0823 (40974)	0.12 (59720)

uncharacteristically low scores due to encoding errors. The HRCs associated with these PVSs corresponding to H.264 were repeated in this test. The corresponding HRCs associated with H.265 were dropped, to keep the total number of PVSs at 192 as in test_2. 26 participants rated the 192 PVSs presented on a 55" LG OLED55C7D screen. No outliers were detected in this test. As test_2 and test_3 are based on the same design philosophy, these two tests can be combined for further analysis.

test_4 The objective of this test was to assess the effect of different framerates on perceived video quality. Hence, the HRC design was based on selecting from different framerates for each of the chosen different resolutions. Four different framerates, namely, 15 *fps*, 24 *fps*, 30 *fps*, and 60 *fps* were used across six different resolutions between 360*p* and 2160*p*. Only H.264 was selected to encode videos for this test. Table 3.9 provides the details of the test design. Eight SRCs with a duration of 7-9 s each was used, with no overlap between sources from the other tests. The selected HRCs in combination with these eight SRCs resulted in a total of 192 PVSs. These PVSs were presented on a 55" LG OLED55C7D screen. 25 participants took part in the test, with two outliers being detected.

Table 3.9: Test Design – test_4 (Framerate variation test).

Resolution	Bitrate [kbit/s] (framerate)			
360p	200 (15)	500 (15)	500 (24)	1000 (24)
480p	500 (15)	1000 (15)	1000 (24)	2000 (24)
720p	1000 (24)	2000 (24)	2000 (30)	4000 (30)
1080p	2000 (24)	4000 (24)	4000 (30)	6000 (30)
1440p	4000 (30)	6000 (30)	6000 (60)	8000 (60)
2160p	6000 (30)	8000 (30)	8000 (60)	15000 (60)

The distribution of MOS for each of the four tests is illustrated in Figure 3.7.

3.1 Lab-based Subjective Quality Assessment

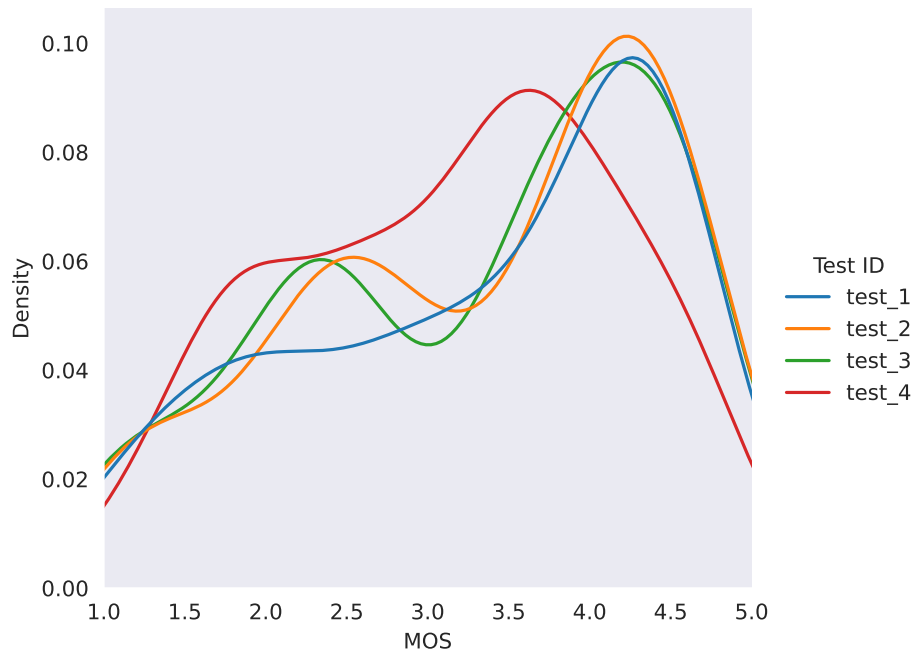


Figure 3.7: MOS distribution of AVT-VQDB-UHD-1 dataset.

In addition to the analysis of the MOS distribution for each of the four tests, SOS analysis was also performed for these tests. The result of this analysis is presented in Figure 3.8 and it can be seen that the values of the SOS parameter “a” for all the tests are in the typical range reported in the SoA for video quality assessment studies.

During analysis of the results for test_2, it was observed that the scores were uncharacteristically low for the `Dancers_8s` sequence, for specific cases of 11 Mbps and 40 Mbps for H.264. On further analysis, it was found that they were some errors during processing due to a bug in the encoding pipeline. To confirm that this indeed was a processing issue, these specific HRCs were re-processed and repeated in test_3, although it was otherwise focused on HEVC and VP9. In test_3, the results for these “problematic” videos from test_2 were as expected for the specific bitrate and resolution conditions, confirming the assumed processing issue for test_2. The respective cases are not included in the presentation of the test_2 results and are also removed from the publicly shared dataset.

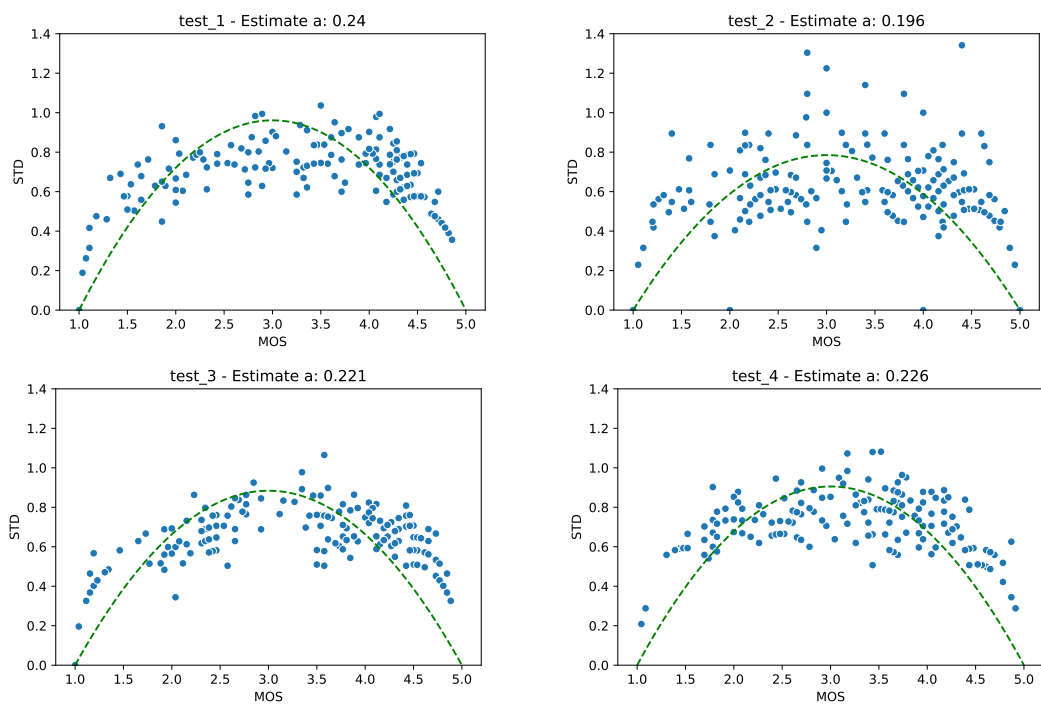


Figure 3.8: SOS analysis of the AVT-VQDB-UHD-1 dataset.

Further analysis was performed on the 60 repeated conditions between test_2 and _3 to analyze the inter-test correlation. It can be concluded from Figure 3.9 that the tests are well correlated and the conditions are rated similarly in both of the tests.

3.1.1.4 AV1 Dataset

In addition to the subjective tests described in the aforementioned sections, further tests were conducted during this work to extend existing SoA models such as ITU-T P.1203.1 Mode 0 [ITU19e] for newer codecs such as AV1 and higher resolutions and framerates. This section describes a dataset consisting of H.265- and AV1-encoded videos.

Seven different source contents of 10 s duration were used in the development of the AV1 dataset. The used SRCs are illustrated in Figure 3.10. The source contents were selected based on different spatial and temporal complexities. Similar to the other datasets described in this chapter, the spatial and temporal complexities were characterized using the SITI metric. Figure 3.11 shows the SI and TI of all the SRCs

3.1 Lab-based Subjective Quality Assessment

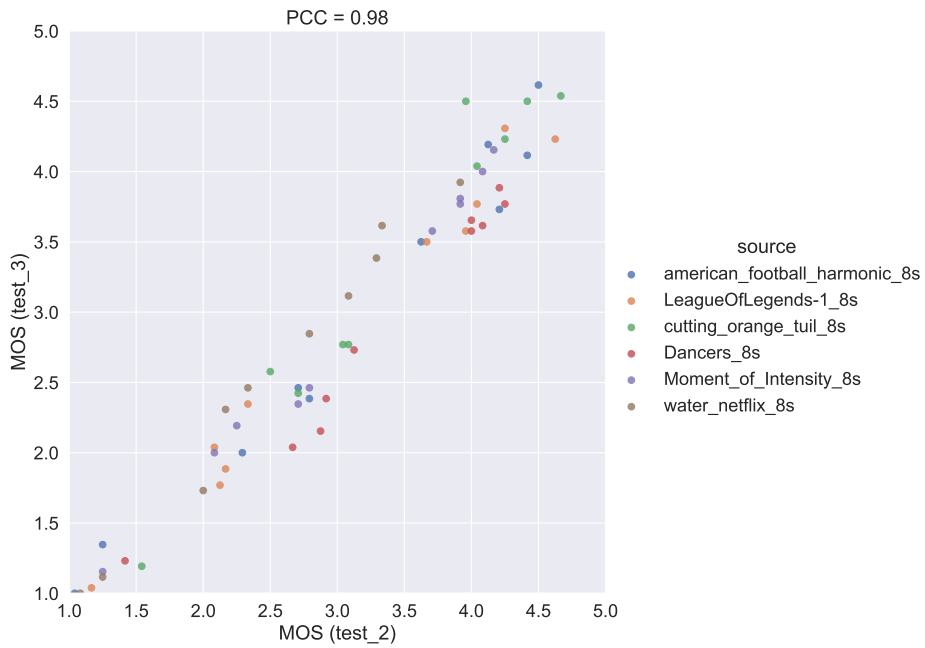


Figure 3.9: Inter-test correlation (test_2 and test_3).

of the AV1-dataset. It can be observed that the SRCs span a wide range of SI and TI. Further details of the source contents are mentioned in Table 3.10. As it can be seen in Table 3.10, the "Space" video was originally in 30 fps. This source was sped up to 60 fps with no negative impact on the content.

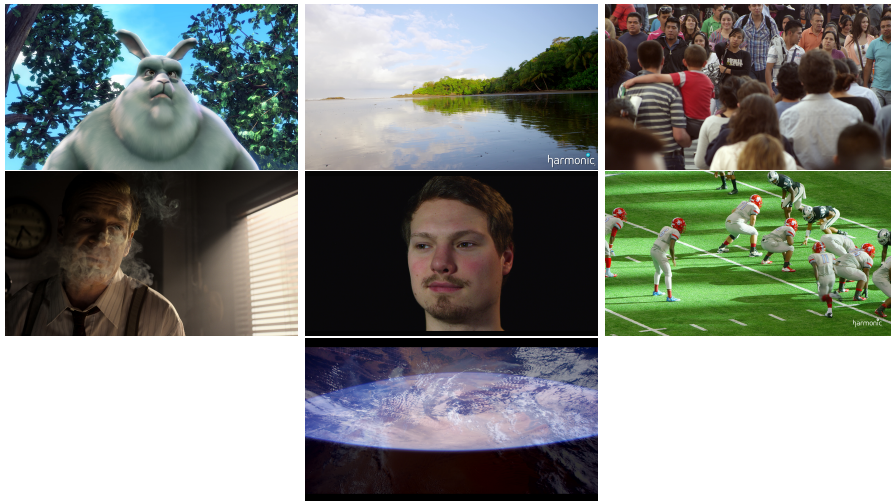


Figure 3.10: Overview of the source videos used in the AV1 dataset.

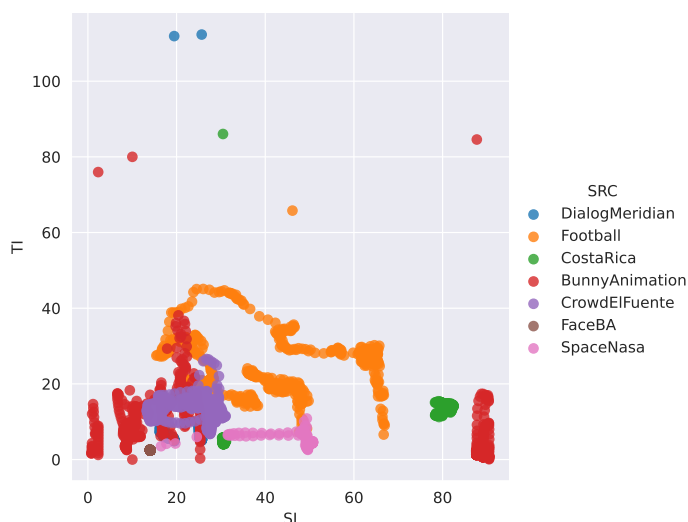


Figure 3.11: SI and TI of all the source contents used in the AV1 dataset.

Table 3.10: Source details for the AV1 dataset.

Content Name	Duration	Details	Source of the content
Animation	10s	3840x2160@60fps	Blender Foundation [Ble]
Landscape	10s	3840x2160@60fps	Harmonic Inc. [Har]
Crowd	10s	3840x2160@60fps	Netflix Inc.
Dialog	10s	3840x2160@60fps	Netflix Inc.
Face	10s	3840x2160@60fps	TU Ilmenau
Sport	10s	3840x2160@60fps	Harmonic Inc. [Har]
Space	10s	3840x2160@30fps*	NASA

3.1.1.5 Test Design

The details of the bitrate-resolution settings used in this test are presented in Table 3.11. Furthermore, two different codecs, namely, H.265 and AV1 are considered. The seven different source contents were encoded with H.265 and AV1 (version released in April 2018) using FFmpeg 4.0. The encoding followed a 2-pass scheme with a 10-bit color depth and a color sub-sampling of 4:2:2. The preset used for H.265 was the default **medium** and the corresponding **cpu-used** parameter for AV1 was

3.1 Lab-based Subjective Quality Assessment

4. It was decided to use the `cpu-used=4` parameter due to performance reasons in the encoding process and to have a comparable preset as for H.265. Each source was encoded in 4 resolutions and 3 bitrates per resolution resulting in a total of 168 PVS's. The detailed test design in terms of bitrate and resolution is described in Table 3.11. In total, 27 subjects participated in the test.

Table 3.11: Test Design - AV1 dataset.

Resolution	Bitrate [kbit/s]				
360p	512	1024	2048		
720p	1024		2048	4096	
1080p			2048	4096	8192
2160p			4096	8192	16384

3.1.1.6 Test Results

For checking the reliability of the users, outlier detection was performed during the analysis. The criterion for outlier detection was based on PCC. PCC was computed between the raw scores of each user and the mean opinion score (MOS). A threshold of 0.8 PCC was used as a criterion to detect outliers. This was slightly different than the thresholds considered for the AVT-PNATS-UHD-1 and AVT-VQDB-UHD-1 datasets. Based on this threshold, there were two outliers. The MOS and the associated confidence interval (95% CI) were computed after removing these outliers. Figure 3.12 shows the overall distribution of the MOS in this test and it can be observed like the other tests described in this chapter, the tendency is toward higher ratings. This is due to the fact that the test was designed to address the quality assessment of high-quality videos.

As with the other datasets, the SOS analysis was conducted on this dataset too. From the result illustrated in Figure 3.13, it can be seen that the value of the SOS parameter “a” is of the same order of magnitude as the other tests.

Furthermore, a comparison of the MOS for the AV1 and H.265 encoded videos was performed and it can be observed from Figure 3.14 that on an average AV1 performs slightly better than H.265.

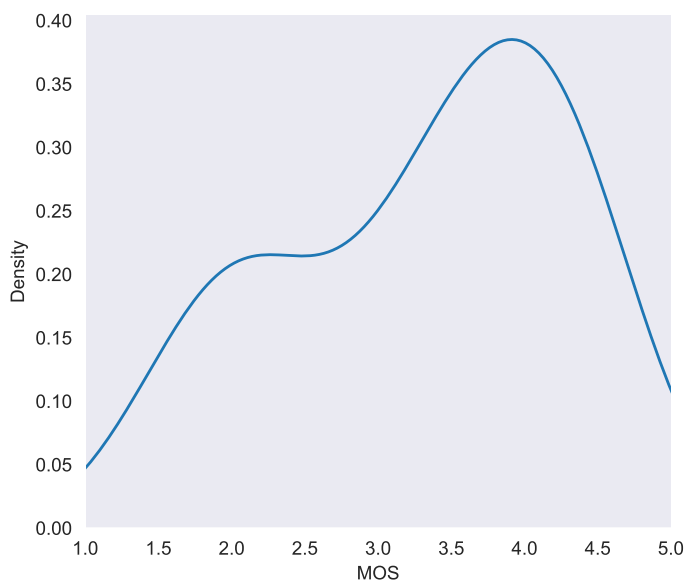


Figure 3.12: MOS distribution of AV1 dataset.

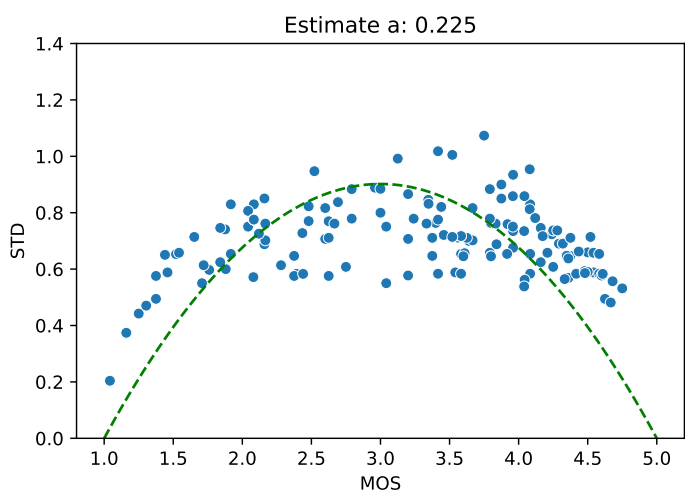


Figure 3.13: SOS analysis of AV1 dataset.

3.1.2 Overall Quality Assessment of a HAS Session

Similar to the short-term video quality assessment, lab-based tests have been used to subjectively assess the overall quality of a HAS session. The tests described in this section consist of tests designed and conducted as part of the “P.NATS Phase 2” competition.

3.1 Lab-based Subjective Quality Assessment

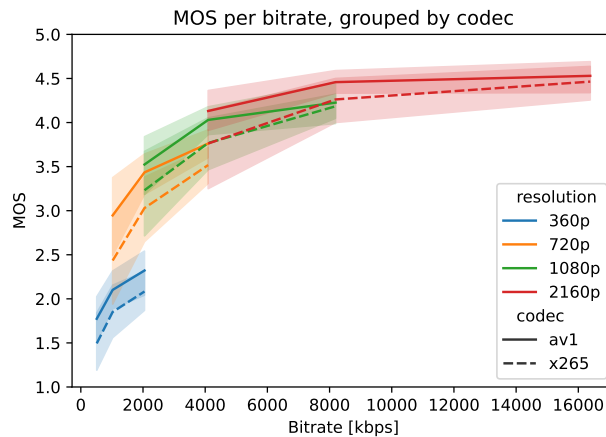


Figure 3.14: MOS comparison between AV1 and H.265.

This consists of a total of six different subjective tests. Two out of these six tests were designed and conducted before the training phase of the competition while the remaining four were designed and conducted after the training phase of the competition. In addition to the variation in video encoding settings, the long tests also involved variation in the duration of the PVSs, initial loading delay, and stalling events. The range of these parameters are summarized in Table 3.12.

Table 3.12: Range of parameters used in the long-duration tests in the PNATS Phase 2 competition.

Parameter	Range
Duration of the PVS	60 seconds - 5 min
Initial loading delay	0 - 30 seconds
Total stalling duration	0 - 26 seconds
Number of stalling events	0 - 5
Total number of quality level switches	0 - 39

Like the short-term tests, these long-term tests were also assigned to different proponents for test conduction. In this section, only five out of the six tests will be described as formal permission was only received for these tests to be described as part of this work. These five tests will be referred to as the “PNATS-UHD-1-Long” hereon.

This dataset consists of five different tests with videos of duration between 1 and 5 min. The tests were designed based on the “immersive” paradigm [PSC14] in which the participants never view the same source stimulus more than once. Each

test included a short training phase in which 3 different videos which included typical HAS-specific degradations such as video quality changes, initial loading delay, and stalling were shown to the participants to familiarize them with these kinds of degradations and consider them in the final assessment of quality.

test_1 and test_2 involved rating videos of 1 min duration. For this purpose, 60 different SRCs in each test were encoded with different HRCs with each HRC consisting of a combination of different HAS specific quality related effects such as quality switches, initial loading delay, and stalling. This resulted in a total of 60 PVSs in both tests. In test_1, 24 participants rated the 60 PVSs and in test_2, 37 participants rated the 60 PVSs with 6 outliers being detected following the criterion of $PCC = 0.7$. The PVSs were displayed on a mobile screen in test_1 with a viewing distance of 6-8H and on a TV in test_2 with a viewing distance of 1.5H. The highest resolution of the PVS used in test_1 was restricted to 2560×1440 as this was the display resolution of the mobile whereas for test_2 the highest resolution of the PVS was kept at the SRC resolution of 3840×2160 .

In test_3 and test_4, the objective was to assess the overall quality of videos of 2 min duration. 30 different SRCs in each test were used and in combination with different HRCs resulted in 30 PVSs. The number of PVSs was adapted to keep the test duration to within 60 min duration. The PVSs in test_3 were presented on a mobile screen and as with test_1 the highest resolution of the PVS used in test_1 was restricted to 2560×1440 and in test_4 it was kept at 3840×2160 as it was a TV test. 24 participants rated 30 PVSs in test_3. In test_4, the 30 PVSs were rated by a total of 31 participants with no outliers being detected.

test_5 involved quality assessment of videos of 5 min duration with 14 different SRCs being used for this purpose. In total 14 PVSs were rated by 31 participants with 5 outliers being detected. As the videos were presented on a mobile screen, the highest resolution of the PVS was again restricted to 2560×1440 .

The following laboratories and companies were involved in conducting the subjective tests. test_1 and test_2 were conducted by Netscout in England, test_3 by SwissQual in Switzerland, test_4 by TU Ilmenau in Germany, and test_5 by Ericsson in Sweden.

3.2 Out-of-the-lab Subjective Quality Assessment

The MOS distribution of all the four tests is shown in 3.15 and reflects a similar tendency of having more PVSs in the higher quality range as in AVT-PNATS-UHD-1 due to a similar test design philosophy used in the design of the dataset.

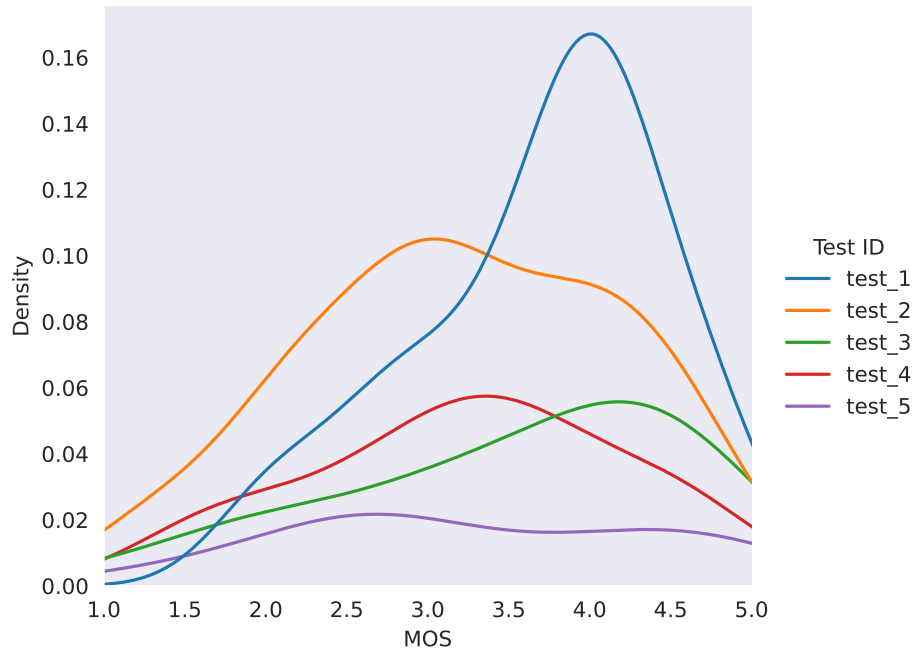


Figure 3.15: MOS distribution of PNATS-UHD-1-Long dataset.

3.2 Out-of-the-lab Subjective Quality Assessment

Although lab-based subjective testing provides an ideal platform for perceptual video quality assessment in a controlled environment, it is both time-consuming and expensive. In addition to this, unforeseen circumstances can make it infeasible to conduct lab tests. For example, the Covid-19 pandemic minimized person-to-person contact to a significant extent and hence made it impractical to conduct lab tests. Hence, other out-of-the-lab approaches have to be developed and investigated for conducting such studies.

Conducting high-resolution video quality assessment studies outside the traditional lab-based settings has its own challenges. Apart from an uncontrolled and non-standardized environment, the devices used for playing out content also vary widely

in an out-of-the-lab setting. Hence, any testing approach should take into account both of these factors. For this, a method to overcome such problems and conduct an out-of-the-lab test for higher resolution content by using a pre-defined crop cut out from the center of the original high resolution ($\geq 2160p$) video is proposed in this section.

3.2.1 Short-term Video Quality Assessment

To test the proposed approach of using a pre-defined crop cut out from the center of the original high-resolution video, quality assessment of 4K/UHD-1 videos is used as a first use case. This section describes the dataset and platform used as well as the necessary pre-processing of the encoded videos to conduct a short-term quality assessment of high-resolution videos in an out-of-the-lab setting.

3.2.1.1 Dataset

The videos from test_1 of the AVT-VQDB-UHD-1 [Rao+19a] dataset are used in this study. Accordingly, six different source videos of a duration of 10 s each were used. The source videos have a resolution of 3840×2160 pixels and a framerate of 60 fps. They were encoded with three different codecs, namely, H.264, H.265, and VP9. For each of the codecs, multiple (bitrate, resolution) conditions were used to encode the videos, resulting in a total of 180 processed video sequences (PVS). The framerate of the encoded videos was kept at the source sequence framerate of 60 fps. In the original lab test, a total of 29 participants took part. As described above, there were no outliers, based on the criterion of 0.75 Pearson correlation between individual subjects and the remaining group of participants.

3.2.1.2 Test Platform

To conduct the crowdsourcing test, a modified version of the publicly available tool avrateNG⁵ was used. According to the developers, the tool avrateNG was originally designed for lab-based tests with a rating by a single user. In a crowdsourcing

⁵<https://github.com/Telecommunication-Telemedia-Assessment/avrateNG>

scenario, it is desirable to enable multiple participants to take part in the test with a self-selected timing and hence possibly simultaneously. To this aim, avrateNG was extended in the following ways: allowing multiple participants to take the test simultaneously, presenting the video stimuli in the client browser, and adding a token-based system to ensure participants do not repeat the same test. This extended version of avrateNG is made publicly available as AVrateVoyager [Gör+21b]⁶.

3.2.1.3 Pre-processing

The encoded video segments were decoded as described in the publicly available implementation of AVT-VQDB-UHD-1 [Rao+19a], which involves a lossless upscaling of the encoded videos to the source sequence resolution and framerate (referred to as the AVPVS in the remainder of the section). In a typical lab test, hardware capable of seamlessly playing out the AVPVS can be ensured. Whereas, in an out-of-the-lab context, neither appropriate playout hardware nor a UHD-1 capable display device can be guaranteed. Since a variety of screen sizes may be used across the participants in an out-of-the-lab test, the fixed 4K/UHD-1 screen and target resolution used in the AVT-VQDB-UHD-1 tests by Rao et al. [Rao+19a] will exceed the available resources in many cases. Hence, different approaches like displaying the most salient regions in a scene are required for quality assessment.

As a consequence, it was decided to display a 540p center crop of the AVPVS which is $(\frac{1}{16})$ th the number of pixels of the AVPVS in the lab test. This is based on the results by Bosse et al. [Bos+16], who concluded that a 128×128 pixels patch out of a 512×512 pixels image is sufficient for subjective image quality assessment and the observations by Göring, Krämmer, and Raake [GKR19] on different pre-defined center crops for full reference model evaluation. However, there still exists the issue of playing out the 540p center-cropped AVPVS seamlessly in the browser. To reduce the data rate of the AVPVS and thus ensure a smooth playout in the browser, the 540p center-cropped version was encoded using H.264 with a CRF of 22. A CRF of 22 guarantees seamless playout in the browser while entailing negligible loss in the visual quality of the AVPVS. In the context of the P.1204 competition [Raa+20a],

⁶<https://github.com/Telecommunication-Telemedia-Assessment/AVrateVoyager>

a similar CRF encoding was used for the playout of stimuli in the case of mobile devices.

3.2.1.4 Test Procedure

The out-of-the-lab test was designed with the intention of restricting the total duration of the test to below 15 minutes. At the beginning of the test, each participant is asked to fill out a form consisting of information regarding the age range, self-judged visual acuity on an ACR scale, the device type being used in the test, and also about the environment the participant is in when doing the test. Only a minimal number of questions were asked to limit annoyance, and all data is stored in an anonymized manner to ensure data protection. Desktops, laptops, tablets, and mobile phones with a recommended minimum resolution of 720p were the devices allowed for the test. The extension of avrateNG included a check for a minimum height and width of the used browser, indicating to the subject to enlarge the window when width < 1100 or height < 500. Three choices were provided to describe the test environment: “Alone in a quiet room”, “Some noise and distractions” and “Significant noise and distractions”.

For the proof-of-concept, a pragmatic approach of asking each participant to rate 30 PVSs that were randomly selected out of the overall number of 180 PVSs was chosen. These 30 PVSs were pre-loaded while the participants answered the pre-test questionnaire and this data had a size of 73MB on average. There was no training phase since the question remains whether training in perceptual quality studies only helps the workers to understand the rating task, or whether it also implicitly suggests the notion of perception/quality of the researchers themselves, and thus, may lead to biased results [SH16]. Foregoing the training phase also has an added effect of keeping the test duration to within 15 minutes.

3.2.1.5 Results and Evaluation

In the first part of this section, the results of the out-of-the-lab study are discussed, and in the second part, the results are compared with the lab test by Rao et al. [Rao+19a]

3.2 Out-of-the-lab Subjective Quality Assessment

to demonstrate the validity and reliability of the proposed crowdsourcing approach for quality assessment of high quality/resolution videos.

Out-of-the-Lab Test Results and Analysis: The crowd panel was mostly recruited from the university body. As the participants were not compensated, this test falls into the category of online testing. Figure 3.16 summarizes the results of the responses to the pre-test questionnaire (cf. Section 3.2.1.4). It can be seen that most of the participants self-reported a good vision and that they took part in the test in an environment with “less distractions”. In addition, most of the participants carried out the test either on a laptop or desktop PC. Most of the participants were in the age range from 18 – 39 years.

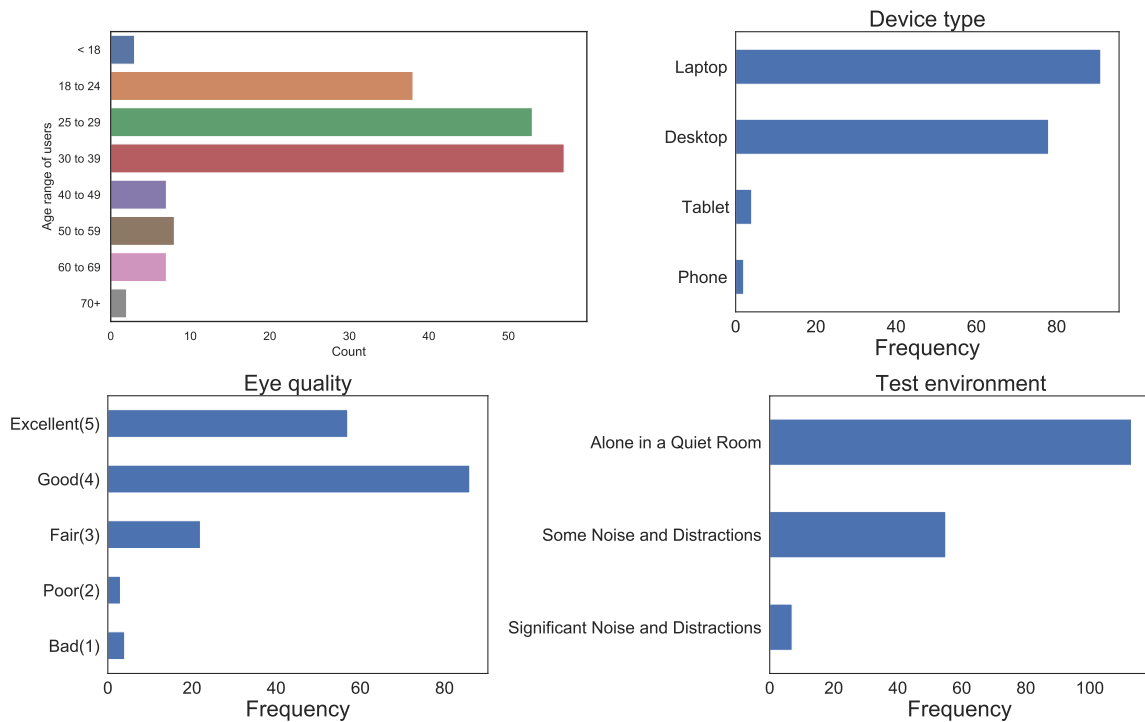


Figure 3.16: Responses to the pre-test questionnaire.

While the participant filled in the questions, the videos were pre-cached and dimensions of the used browser window were collected, see Section 3.2.1.4. The distribution of the extracted height of the window in which the video was viewed is shown in Figure 3.17.

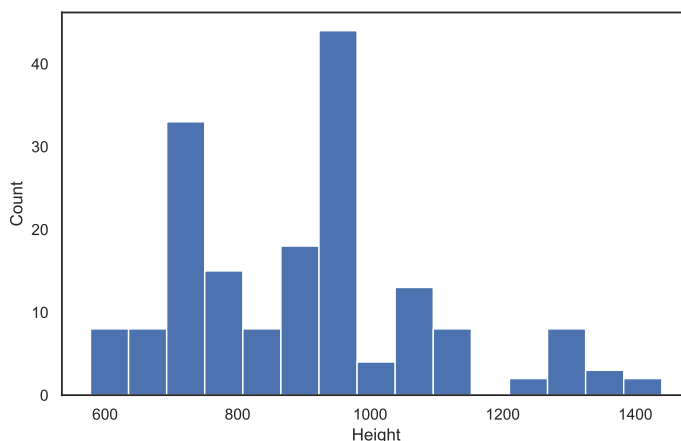


Figure 3.17: Distribution of browser window height across crowd participants.

It can be seen that most of the subjects used the recommended screen resolution of 720p to watch and rate the videos. An interesting observation is that there are very few subjects, $\approx 18\%$, who used a device with a resolution of Full-HD or higher. This indicates that running an out-of-the-lab study for quality assessment of higher-resolution videos is challenging. The device distribution substantiates the need for a test method such as the centre-crop approach used in the presented out-of-the-lab study.

A total of 175 subjects participated in the online study. The participants in this study consisted of people recruited from the university body via email reflectors (reaches students and staff). To determine the outliers in the test, a criterion based on Pearson correlation coefficient (PCC) was used. In the case of a PCC lower than 0.75 of the individual subject’s ratings to the mean ratings across all subjects, that subject was considered as an outlier. Based on a threshold of $PCC = 0.75$, 19 outliers were detected and the ratings from these participants were removed from further analysis. A total of 3987 ratings were obtained after outlier removal, with an average of 22.15 ratings per PVS. In addition, an analysis of how often each PVS is rated was conducted, and created a histogram of these counts is shown in Figure 3.18.

Furthermore, since each participant rated only 30 randomly selected PVSs out of the 180 total PVSs, further analysis was performed to determine the minimum number of subjects needed to have each PVS rated at least once. For this purpose, an analysis of the test results with 64 different randomizations of the order of participants’ ratings

3.2 Out-of-the-lab Subjective Quality Assessment

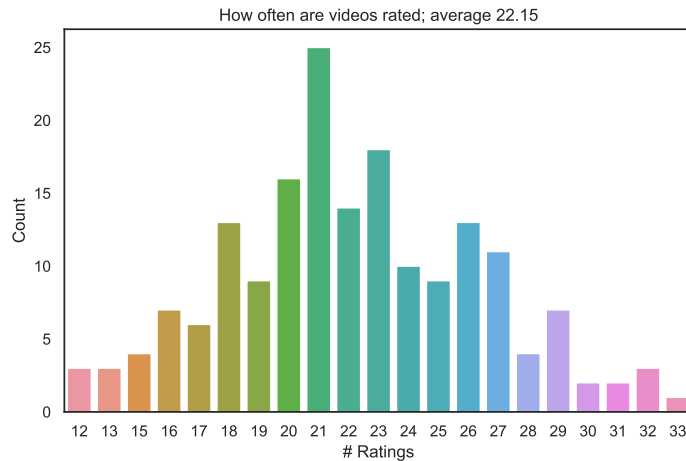


Figure 3.18: Count distribution of how often PVSs were rated; e.g. $x=24$ and $y=10$ means that 10 PVSs were rated 24 times in the crowd test, etc.

was performed and the results were averaged. This analysis indicated that for the given test, it took 39 participants to have each PVS rated at least once, and 144 participants to have each PVS rated at least ten times.

Lab versus Out-of-the-Lab Comparison: In this part, the comparison of the results of the out-of-the-lab and the lab tests is presented.

The distributions of the MOS values of both lab [Rao+19a] and out-of-the-lab tests are shown in Figure 3.19. From the more negative ratings, it can be observed that participants in the out-of-the-lab test are more critical as compared to the participants in the lab test while rating the videos. This can likely be attributed to the fact that in the out-of-the-lab test, the 540p center-cropped versions of the video were rated by the participants and not the full UHD-1 version as in the lab test. As a consequence, the participants in the out-of-the-lab test focused on a smaller area of the video, and hence may have been more sensitive to any kind of distortions. Further, since only a small portion of the video was shown, semantic information and a full understanding of the sequences were not enabled, so the test subjects may have had a stronger focus on the video-signal quality than with the full video frame being shown. In future work, such hypotheses can be investigated by assessing the same sequences within an eye-tracking study, comparing the cropped versus the lab-based full-screen presentation.

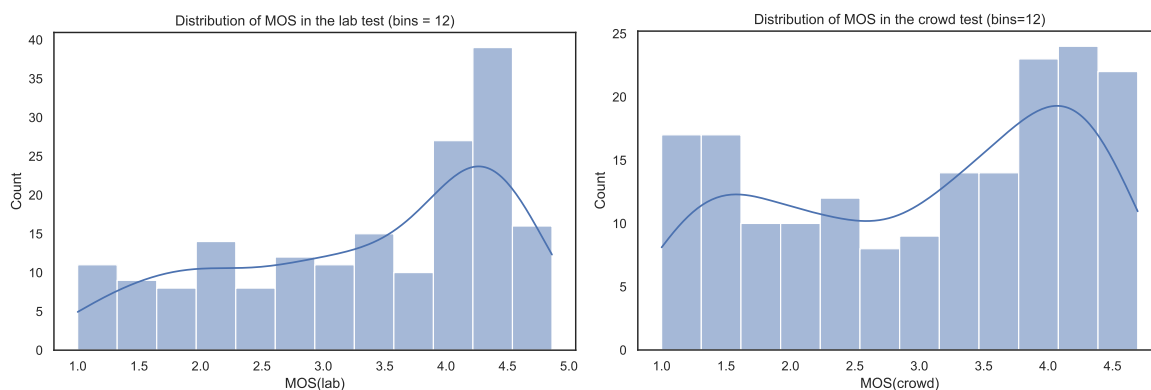


Figure 3.19: Distribution of MOS in the lab and out-of-the-lab tests.

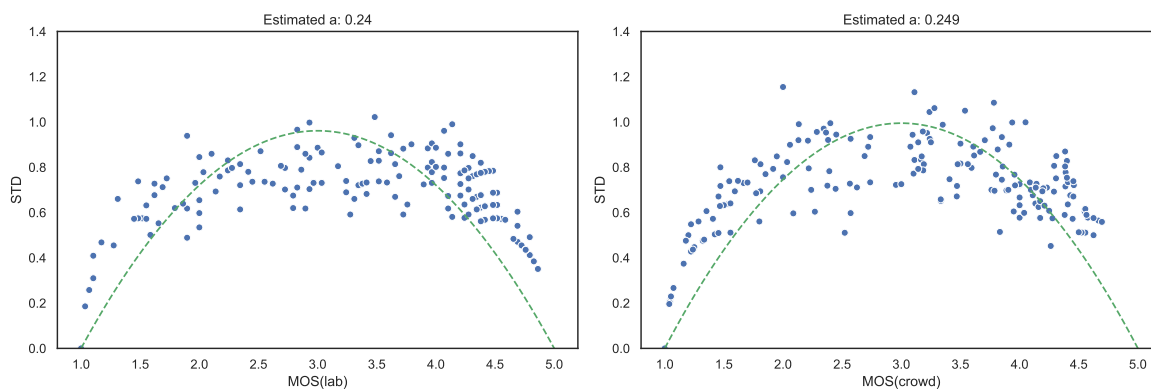


Figure 3.20: SOS analysis of the lab and out-of-the-lab tests.

Furthermore, an analysis of the distribution of standard deviations over the MOS (SOS analysis) as described in [HSE11] is performed to estimate the similarity between the lab and out-of-the-lab tests. The results of the SOS analysis are illustrated in Figure 3.20. For the lab test, the value of the SOS parameter was found to be $a_{lab} = 0.240$, and for the out-of-the-lab test of $a_{crowd} = 0.249$, indicating the same order of magnitude and hence a strong similarity of both tests in this regard.

Figure 3.21 shows the comparison of the MOS from the lab and out-of-the-lab tests. It can be seen that there is a very high correlation between the two tests, with a Pearson correlation of 0.96, which is comparable to the performance achieved for cross-lab testing for video quality assessment [PW]. This indicates the validity and reliability of the out-of-the-lab approach of using a 540p center-cropped version of a UHD-1 upscaled video to evaluate the video quality.

3.2 Out-of-the-lab Subjective Quality Assessment

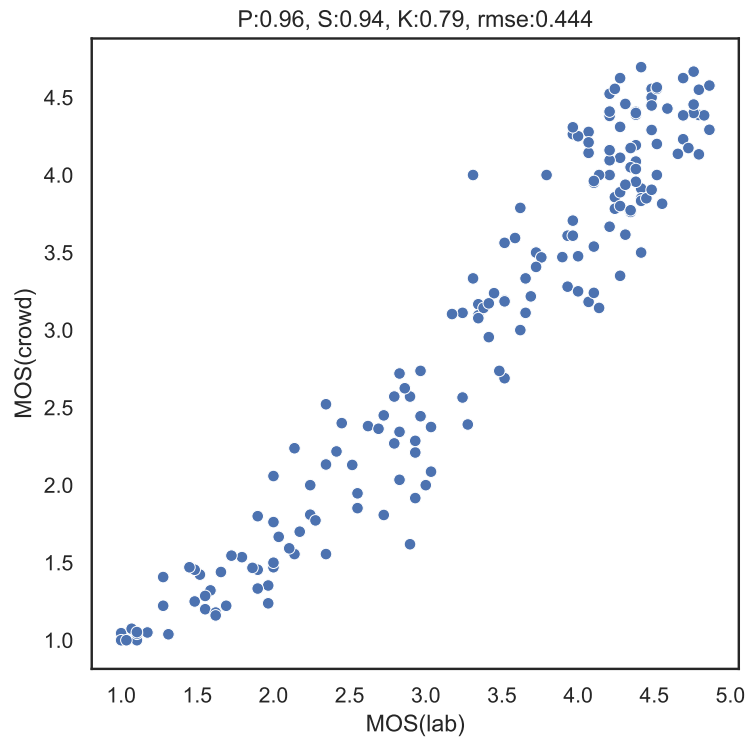


Figure 3.21: Scatter plot of the MOS values from lab [Rao+19a] and out-of-the-lab tests.

Table 3.13: Per-source comparison of lab [Rao+19a] and out-of-the-lab test results.

source	pearson	spearman	kendall	rmse
american_football_harmonic	0.98	0.96	0.85	0.535
bigbuck_bunny_8bit	0.99	0.95	0.84	0.305
cutting_orange_tuil	0.96	0.89	0.75	0.276
surfing_sony_8bit	0.97	0.96	0.86	0.689
vegetables_tuil	0.96	0.88	0.73	0.346
water_netflix	0.99	0.98	0.89	0.444

Furthermore, a comparison of the performance of the two test paradigms on a per-source basis is conducted. As can be seen from Table 3.13, also at a per-source level there is a very high correlation between the two tests.

In addition, the correlation between the lab and out-of-the-lab tests as a function of the number of participants in the out-of-the-lab test is analysed. The reason for this is to evaluate how many participants are required in such a non-full-factorial out-of-the-lab test. For this analysis, the same approach of randomizing the participant order 64 times as described in Section 3.2.1.5 was used and computing the average overall

randomizations. Figure 3.22 shows the variation of the correlation between the lab and out-of-the-lab tests with regard to the number of crowd participants. It should be noted that it took 39 participants to rate each PVS at least once as described earlier. From the figure it can be seen that for a correlation between the tests greater than 0.92, a minimum of 39 participants is required, then leading to a similar correlation as found for cross-lab test comparisons [PW]. It should be noted that the required participants in both cases being the same at 39 is a mere coincidence.

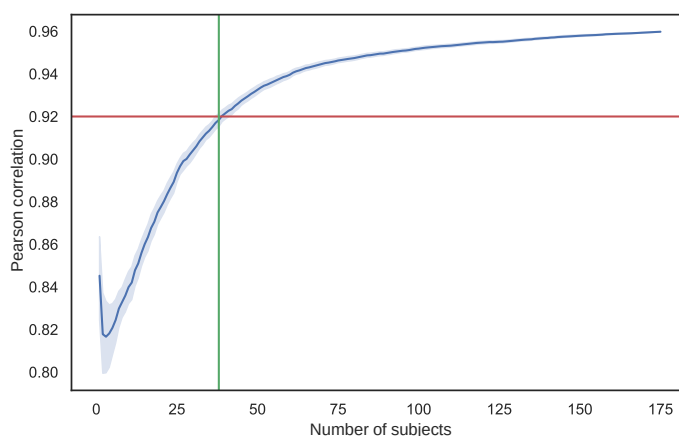


Figure 3.22: Correlation between lab and out-of-the-lab tests as a function of the number of participants in the out-of-the-lab test.

3.2.2 Overall Quality Assessment of a HAS Session

The crowd test to assess overall HAS session quality uses the approach described in Section 3.2.1.3. As with the short-term video quality assessment studies, overall HAS session quality assessment studies can be conducted in an out-of-the-lab setting. However, this comes with additional challenges. One major challenge to conducting such tests with videos of longer duration is the number of PVSs that each participant in a out-of-the-lab setting is asked to rate. Unlike short-term video quality assessment where it is still possible to subsample the PVSs to ensure that each test subject views and rates videos covering the overall quality range, it becomes more difficult when using videos of longer duration as the overall number of PVSs that is rated by a participant is limited. Hence, it is needed to compare the subjective ratings between lab- and crowd-based tests to investigate the rating behavior of the subjects in these

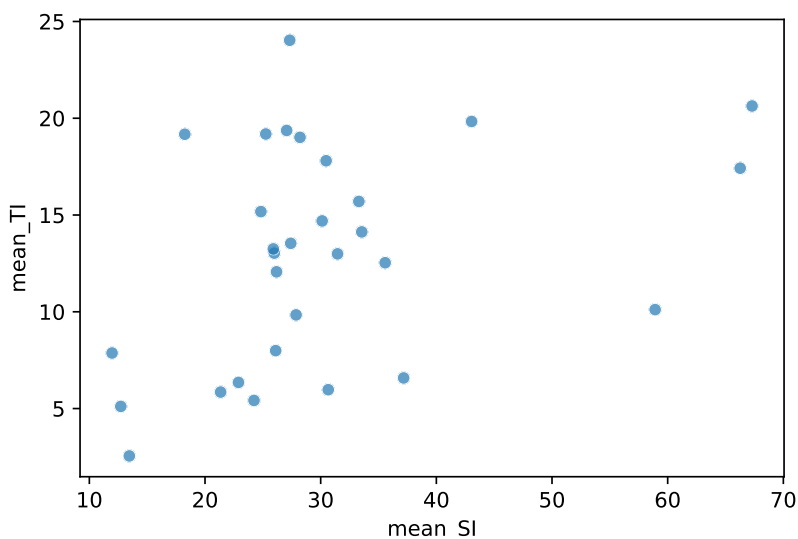


Figure 3.24: SI-TI of all the sources used for the long-term audiovisual quality evaluation in the crowd.

3.2.2.2 Crowdsourcing Platform

The study was conducted using the Clickworker⁷ platform. The countries from which the participants were recruited were restricted to Europe. For the rating task, AVrateVoyager [Gör+21b] was used. All the checks mentioned in Section 3.2.1.4 were also repeated in this test.

3.2.2.3 Pre-processing

The encoded videos were decoded as done for the short-video segments along with lossless upscaling of the encoded videos to SRC resolution and framerate. Furthermore, a 720p centre-crop of the video was extracted to be played out in the rating task. The decision for using a larger centre-crop as compared to the short-duration video quality crowdsourcing test was to provide more context in terms of the video content as the duration of the content was longer. This centre-crop version of the video was then encoded with a CRF of 22 using H.264 with a yuv420 8-bit pixel format. Also, the decision of using the centre-crop and the chosen CRF-based encoding was primarily motivated by the challenges outlined in Section 3.2.1.3.

⁷<https://www.clickworker.com/>

3.2.2.4 Test Procedure

As with the short-duration video quality test, the overall test duration was restricted to 15 min. This included the time required to fill in the pre-test questionnaire which consisted of the same questions asked in short duration video quality assessment study. Unlike the short-duration video test, this test had a training phase consisting of one training video with all the possible degradations related to a HAS session such as initial loading delay, quality switches, and stalling events to familiarize the test participants with these degradations while evaluating the video quality. Furthermore, the subjects were provided explicit instructions to consider only the degradations and not the content to evaluate the overall quality of a session. There were no distortions introduced to the audio. A total of 100 crowdworkers were recruited via the Clickworker platform and as a pragmatic approach, each crowdworker was asked to rate 5 PVSs that were randomly selected out of the overall number of 30 PVSs. The crowdworkers were compensated with an appropriate payment for this task thus making this study a traditional crowdsourcing study. These 5 PVSs were pre-loaded while the subjects answered the pre-test questionnaire.

3.2.2.5 Results and Evaluation

The results are presented in two parts. The results of the crowdsourcing study are presented in the first part. In the second part, the comparison results of the lab and crowdsourcing tests to demonstrate the applicability and reliability of extending the centre-crop-based video quality assessment for long-duration videos with HAS related impairments.

Crowdsourcing Test Results and Analysis: As mentioned above, the crowdworkers were recruited using Clickworker, and hence unlike the short-duration video test, the subjects were paid. Figure 3.25 shows the results of the pre-test questionnaire. It can be observed that most of the participants did the test alone in a quiet room with a significant proportion of them doing it on their laptop or desktop and self-reporting good to excellent visual acuity. It should be noted that visual acuity determination is

based on self-reporting on a 5-point ACR scale. Also, there is a good distribution in the age range of participants taking part in the study.

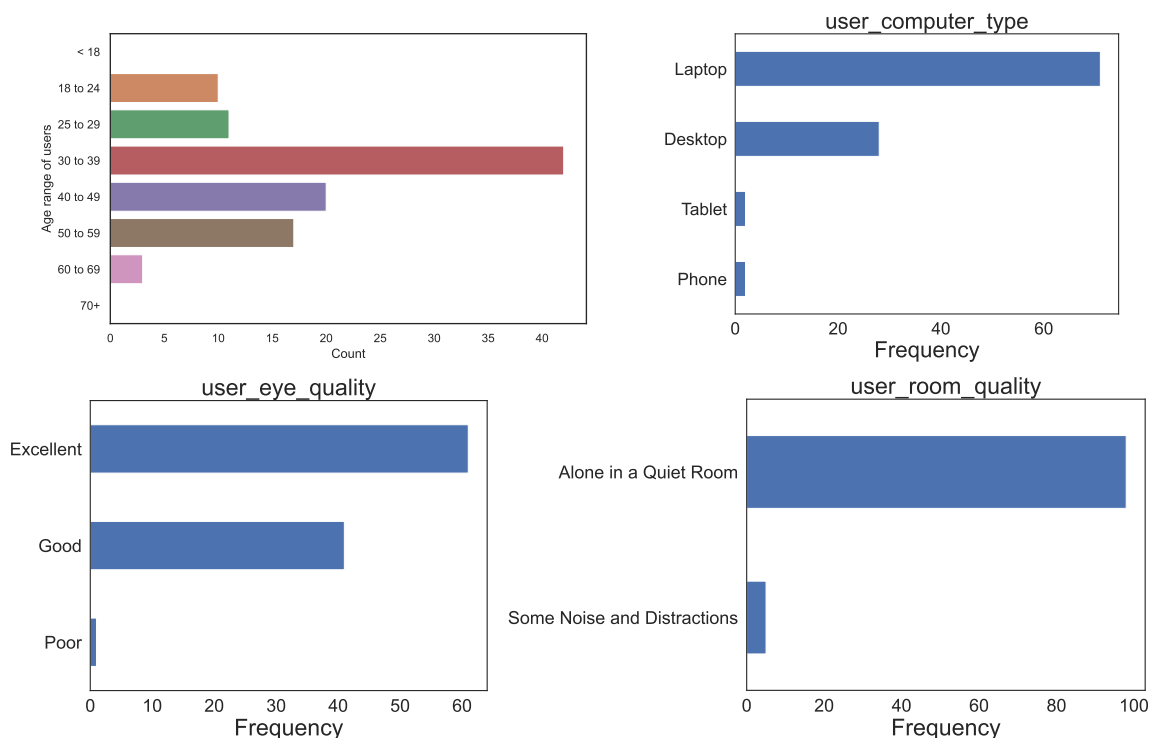


Figure 3.25: Responses to the pre-test questionnaire.

In addition to gathering responses of the participants using the pre-test questionnaire, the dimensions of the used browser window were also collected in parallel to the subjects answering the questionnaire. From Figure 3.26, it can be observed that like in the short-duration video test, very few subjects (<10%) used a screen with 1080p or higher resolution, thus justifying the decision to use a 720p center-crop for quality assessment.

In addition to this, an analysis of how often each PVS was rated was performed and is illustrated in 3.27 and the average number of ratings for each PVS was 17.2.

Lab versus Crowd Comparison: A comparison of the lab and crowd tests is described in this section to show that the centre-crop approach can be used for the assessment of long-duration videos with HAS-related impairments. Figure 3.28 shows the distribution of the MOS in both lab and crowd tests. As with the short-

3.2 Out-of-the-lab Subjective Quality Assessment

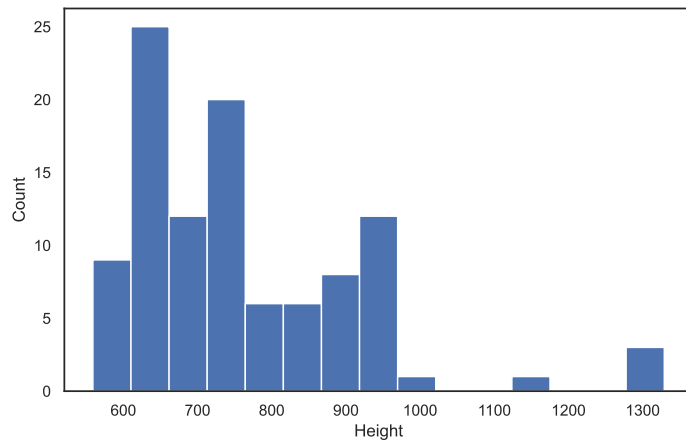


Figure 3.26: Distribution of browser window height across crowd participants.

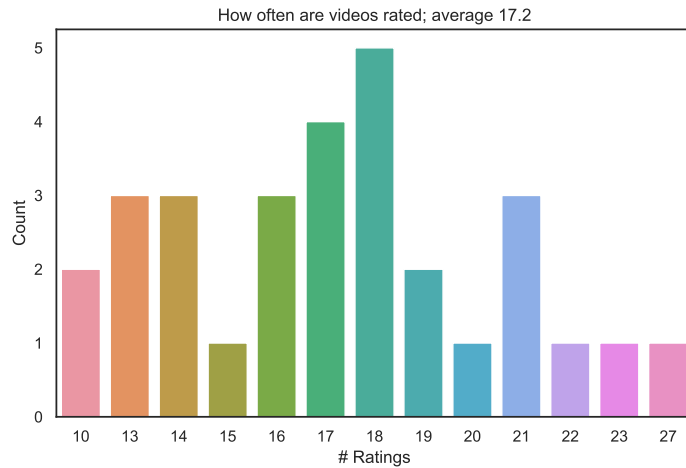


Figure 3.27: Count distribution of how often PVSs were rated; e.g. $x=18$ and $y=5$ means that 5 PVSs were rated 18 times in the crowd test, etc.

duration video test, the crowd participants are more critical than the lab subjects also most likely because they had a smaller region to focus on and hence would have been more critical to the video-related degradations. As with the short test, this hypothesis has to be further investigated.

Furthermore, an SOS analysis of both the lab and crowd tests, respectively, was conducted to estimate the similarity between the two tests. There exists a high similarity between the tests which is confirmed by the SOS plots illustrated in Figure 3.29 and same order of the magnitude of the SOS parameter with $a_{lab} = 0.221$ and $a_{crowd} = 0.226$.

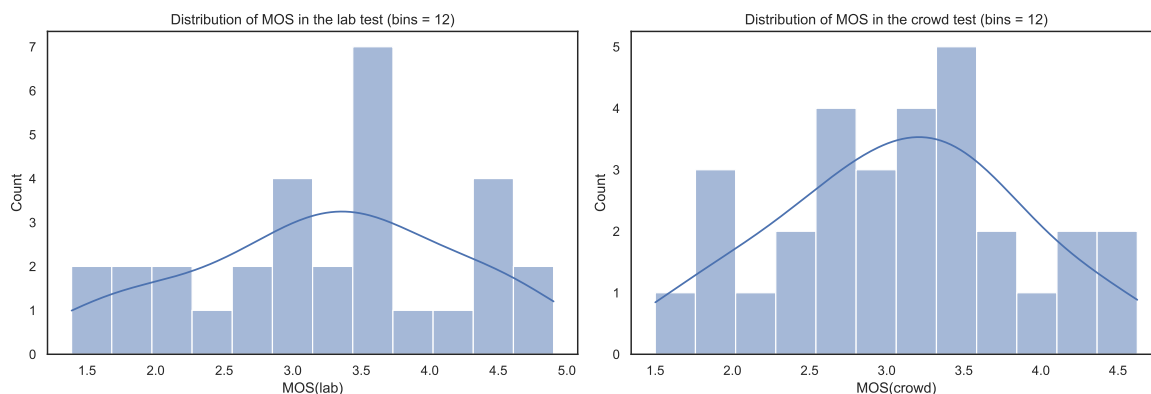


Figure 3.28: Distribution of MOS for the lab and crowd tests.

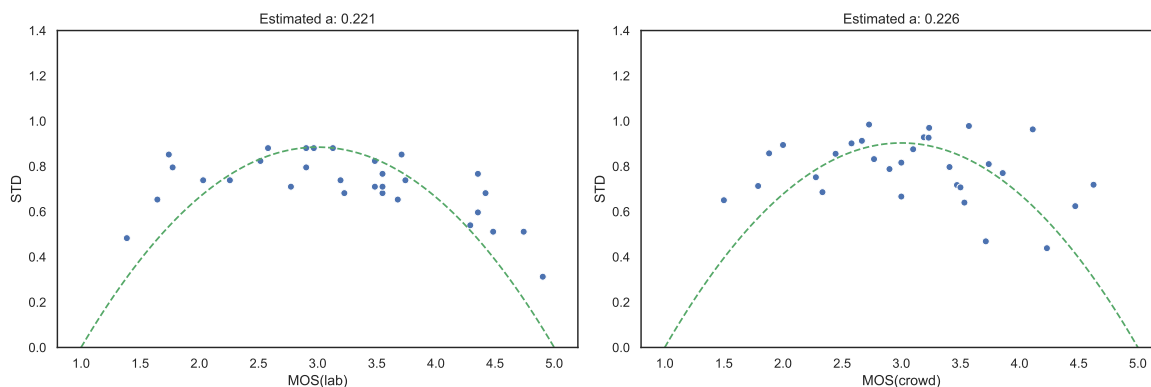


Figure 3.29: SOS analysis of the lab and crowd tests.

In addition to these analyses, a comparison of the MOS from the lab and crowd tests was performed and is shown in Figure 3.30. A high correlation can be observed between the two tests with a PCC of 0.96 thus indicating the validity and reliability of extending the crowdsourcing approach to assess the overall quality of a HAS session. Also, it can be observed from the high value of the Spearman correlation of 0.94 that the rank order of the PVS is similar in both the tests, and the general agreement in assessing the cases related to stalling events further establishes that the instructions provided to the participants were sufficient.

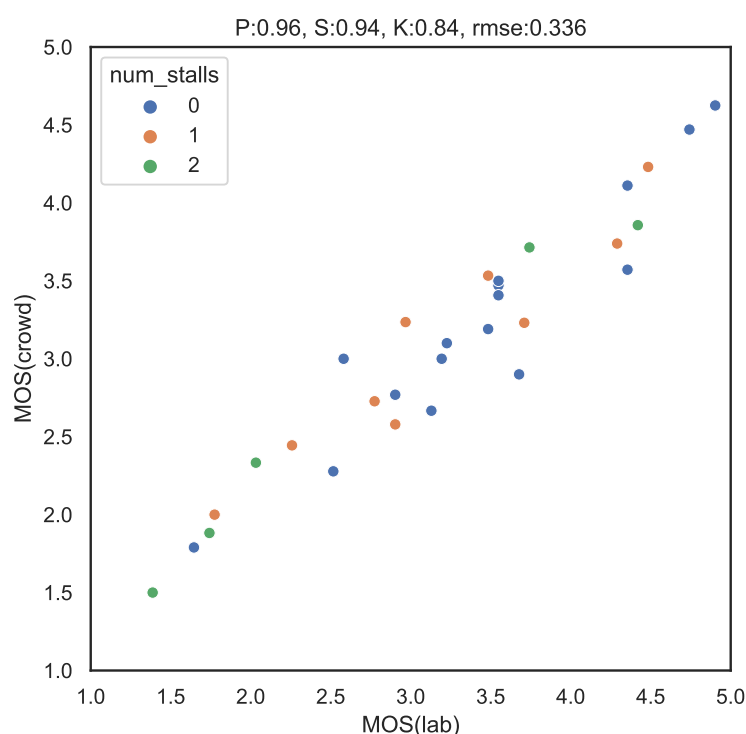


Figure 3.30: Scatter plot of the MOS values from lab [Rao+19a] and crowd tests.

3.3 Summary

To summarize, this chapter focused on subjective quality assessment of short-term video quality for high-resolution videos and the overall integral quality of a HAS session in both lab and out-of-the-lab contexts. Hence, a series of short-duration video quality assessments and overall HAS session quality assessment studies designed and conducted over the course of this thesis have been presented.

Mainly, lab-based tests were used for the quality assessment for both use cases. For this, two different series of short-term video quality assessment tests have been presented. The first set of short-duration video quality tests was designed and conducted as part of the *P.NATS Phase 2* competition. This consisted of four different subjective tests using a wide range of source contents with more than 50 different SRCs in each of the four tests. All these tests follow a non-full-factorial test design due to the high number of SRCs employed in each of the tests. These tests form the AVT-PNATS-UHD-1 dataset. The second set of tests was developed in the context

of this thesis in parallel to the *P.NATS Phase 2* competition, and also consisted of four different tests. A total of 17 different SRCs were used in these tests and all the tests in this set followed a full-factorial design. These four tests together form the publicly available AVT-VQDB-UHD-1 dataset. Furthermore, a brief description of the dataset created as part of the *P.NATS Phase 2* competition, namely, PNATS-UHD-1, is provided.

In addition to these two test series, another test was designed and conducted to compare the performance of AV1 with H.265 in terms of subjective ratings and also generating ground truth to extend existing SoA models for newer codecs and higher resolution and framerate.

Following this, a series of overall HAS session quality assessment studies were designed as part of the *P.NATS Phase 2* competition. These tests used audiovisual sequences of duration 1-5 min across five different tests. These five tests form the PNATS-UHD-1-Long dataset. Out of five tests designed for this purpose, one was conducted as part of this thesis. The remaining four tests were conducted by other proponents involved in the *P.NATS Phase 2* competition, namely, Ericsson, Netscout, and SwissQual. These databases were shared by the respective proponents for model development and evaluation conducted as part of this thesis.

All these datasets were created with the primary objective of generating ground truth for developing models capable of short-term video quality prediction for high-resolution video and overall HAS session quality. Hence, this will be used as ground truth for the model development and evaluation performed as part of this work and will be presented in Chapter 4.

To assess the viability of using out-of-the-lab testing approaches for quality assessments of the short-term video quality of high-resolution videos and overall integral quality of a HAS session, further tests were conducted. For this purpose, an approach based on a pre-defined centre-crop was proposed to conduct short-term quality assessment of high-resolution videos in an online scenario. Using this approach, an online study using the PVSs from test_1 of AVT-VQDB-UHD-1 was conducted. A high correlation between the MOS of the lab and out-of-the-lab tests demonstrated the validity of the out-of-the-lab test and its agreement with the corresponding lab test.

As a next step, the applicability of the proposed centre-crop approach for overall HAS session quality assessment was tested. For this, the PVSs from test_4 of the PNATS-UHD-1-Long dataset consisting of videos of 2 min duration were used. The analysis of the results from the lab and crowd tests in terms of SOS and correlation showed a high similarity between the tests and validity and agreement of the crowd test respectively. From the observations from the two out-of-the-lab studies, it can be concluded that these can be used as a viable alternative to the lab tests for both use cases, thus addressing research question 5.

Using the data from the subjective tests, a family of models referred to as *AVQBits* has been developed and evaluated as part of this thesis and will be presented in the next chapter.

AVQBits: Adaptive Bitstream-based Video Quality Model

With the aim of accurate video quality monitoring either at the client side, in the network, or directly after encoding, which is critical to assess and improve the Quality of Experience perceived by the end-user, a versatile, bitstream-based video quality model namely *AVQBits* is presented in this chapter. It can be applied in several contexts such as service monitoring, evaluation of encoding quality, gaming video QoE, and even omnidirectional video quality assessment. At its core, *AVQBits* encompasses the standardized ITU-T P.1204.3 model, with further model instances that can either have restricted or extended input information, depending on the application context. Four different instances of *AVQBits* are proposed and investigated: (1) P.1204.3 as a bitstream-based model with full access to encoded bitstream information, (2) a “Mode 0” variant, with access only to metadata; (3) a “Mode 1” instance using frame-level data and metadata; (4) a Hybrid no-reference Mode 0 (two variants) extension which has access to pixel data and metadata. For training, the AVT-PNATS-UHD-1 subjective test dataset presented in Chapter 3 is used. The evaluation is performed with the publicly available AVT-VQDB-UHD-1 [Rao+19a] dataset. All *AVQBits* variants are made publicly available for the community for further research.

With this, research questions 1 and 2 as outlined in Chapter 1 will be tackled in this chapter.

This chapter is based on the following publications:

- [Raa+20a] Alexander Raake, Silvio Borer, Shahid Satti, Jörgen Gustafsson, **Rakesh Rao Ramachandra Rao**, Stefano Medagli, Peter List, Steve Göring, David Lindero, Werner Robitza, Gunnar Heikkilä, Simon Broom, Christian Schmidmer, Bernhard Feiten, Ulf Wüstenhagen, Thomas Wittmann, Matthias Obermann, and Roland Bitto. “Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204”. In: *IEEE Access* 8 (2020)
- [Rao+20a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Peter List, Werner Robitza, Bernhard Feiten, Ulf Wüstenhagen, and Alexander Raake. “Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analysis and Open Source Implementation”. In: *Twelfth IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020
- [RGR22] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications”. In: *IEEE Access* 10 (2022)
- [Rao+19b] **Rakesh Rao Ramachandra Rao**, Steve Göring, Patrick Vogel, Nicolas Pachatz, Juan Jose Villamar Villarreal, Werner Robitza, Peter List, Bernhard Feiten, and Alexander Raake. “Adaptive video streaming with current codecs and formats: Extensions to parametric video quality model ITU-T P.1203”. In: *Electronic Imaging* (2019)
- [RGR21a] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Enhancement of Pixel-based Video Quality Models using Meta-data”. In: *Electronic Imaging, Human Vision Electronic Imaging*. 2021

The base model, *AVQBits|M3* P.1204.3, was developed as part of the “P.NATS Phase 2” competition conducted in ITU-T SG12/Q14 which resulted in the ITU-T P.1204 series of Recommendations. To enable a detailed understanding of the model development and validation in the standardization context, an overview of the “P.NATS Phase 2” competition is provided before presenting the details of the model algorithms and the extensive evaluation of the models.

4.1 P.NATS Phase 2 Competition

The details related to the “P.NATS Phase 2” competition presented in this thesis is based on Raake et al. [Raa+20a]. The overall procedure was designed by the proponents and was formalized in the of form contributions and temporary documents presented at ITU-T SG12/Q14.

4.1.1 General Details

With the aim of developing more accurate short-term video quality models than the ones recommended as part of the ITU-T P.1203 series of Recommendations, ITU-T SG 12 in collaboration with the Video Quality Experts Group (VQEG) launched the *P.NATS Phase 2* competition. Initially, a total of 11 proponents agreed to be part of the competition. Over the course of the competition, two out of the 11 proponents dropped out and did not submit any models for further evaluation.

Models in the following categories were allowed to be submitted as part of the competition for consideration to be standardized based on the criteria that will be outlined later: Bitstream Mode 0 (BSM0), Bitstream Mode 1 (BSM1), Bitstream Mode 3 (BSM3), Pixel Full-Reference (PXFR), Pixel Reduced-Reference (PXRR), Pixel No-Reference (PXNR), Hybrid Full-Reference Mode 0 (HYF0), Hybrid Full-Reference Mode 1 (HYF1), Hybrid Full-Reference Mode 3 (HYF3), Hybrid Reduced-Reference Mode 0 (HYR0), Hybrid Reduced-Reference Mode 1 (HYR1) and Hybrid Reduced-Reference Mode 3 (HYR3), Hybrid No-Reference Mode 0 (HYN0), Hybrid No-Reference Mode 1 (HYN1) and Hybrid No-Reference Mode 3 (HYN3).

As part of the competition, 13 training databases were jointly created by the proponents for model development. After the models were trained using these databases and submitted to the ITU-T TSB, 13 validation databases were jointly designed by the proponents. The details of the test design and overview in terms of the average confidence interval, average correlation, target display, number of test participants, and the number of PVSs rated by each participant for these training and validation databases have been presented in Chapter 3.

As part of performing the validation using unknown databases, the subjective scores of the validation tests were submitted to the ITU-T TSB. A period of bug-fixing was allowed for the models before the final validation of the submitted models and a well-defined bug-fixing procedure was designed for this purpose. Following the bug-fixing procedure, each proponent had to compute the scores using the models that were submitted and uploaded these scores to a dedicated folder only accessible to a given proponent on the ITU-T TSB.

Once the prediction scores of all the models were uploaded, a model verification procedure was carried out. The objective of this procedure was to ensure that the submitted predictions were produced by the submitted models that were uploaded to the ITU-T TSB. The process consisted of proponents reproducing scores for a dedicated number of PVSs under the supervision of one other proponent.

Following the verification of all the models, the subjective scores of the validation databases were disclosed to all the proponents to validate the models and compute the performance numbers in terms of the root mean square error (RMSE) for each model to determine either a winning model or a winning group. In total, 35 model candidates spanning 10 different model categories were submitted as part of the competition. The determination of the winning candidates was done by following a well-defined statistical evaluation procedure.

The following section is mainly focused on describing the used statistical evaluation. In-depth information on the databases used to determine the winning candidates is presented in Chapter 3.

4.1.2 P.NATS Phase 2 Statistical Evaluation

The first step of the statistical evaluation procedure consisted of data cleaning and mapping.

4.1.2.1 Data Cleaning and Mapping:

In this step, each database from both training and validation was inspected to identify problematic cases related to errors in applying the specific encoding settings,

unsuitable sources, or improper subjective test conduction. This process involved both removing individual PVSs and entire databases if issues were found.

To analyse the validity of a database, the common set PVSs were used. The rank order and absolute scores of these common PVSs were investigated to determine if the different subjective tests had a similar range in terms of quality scores. During this analysis, it was found that one training database did not comply with the agreed-upon subjective testing procedure and hence was removed from further analysis and winning model determination procedure.

Furthermore, to remove the bias between subjective tests across different labs, a per-database linear mapping was applied to the predicted scores before computing the performance evaluation metrics as recommended in ITU-T Rec. P.1401 [ITU14a]. Following this, the performance evaluation metrics were computed for each submitted model.

4.1.2.2 Performance Measure:

As aforementioned, the performance measure used to evaluate the submitted models is RMSE aggregated across all databases [ITU14a]. For the determination of model performance and subsequent comparison of the models, both training and validation databases were used with different weights. For the final performance calculation, the training databases had a weight of 0.1 ($w_{training} = 0.1$) and the validation databases had a weight of 0.9 ($w_{validation} = 0.9$). Following the per-database RMSE computation, the aggregated error across all the databases was computed as a weighted sum of the RMSE per database and is defined by Equation (4.1).

$$p_v = \frac{1}{W} \sum_{k=1}^M w_k \cdot RMSE_{k,v}^2 \quad (4.1)$$

where M represents the total number of (training and validation) databases, w_k is the weight of each database, and $RMSE_{k,v}$ is the root mean square error of model v for database k . The normalization constant W is given by $W = \sum_{k=1}^M w_k$. The model achieving the lowest p_v value is the best model.

In addition to computing the performance evaluation measure for each of the submitted models, a minimum performance criterion based on a baseline model was determined.

4.1.2.3 Minimum Requirement:

A simple model which uses bitrate to predict the quality of a video was chosen as a baseline model. The baseline model is as described in Equation (4.2).

$$Q_{baseline} = a \cdot \log(\text{bitrate} + b) + c \quad (4.2)$$

The coefficients of the model are both codec and target display device dependent and the values were determined by training against the corresponding training data. Using the determined coefficients, the performance of the model was calculated as described in Section 4.1.2.2 and the p_v associated with this $p_{baseline}$ was used as a minimum threshold which each of the submitted models had to better to be considered for winning model determination.

Following the determination of the $p_{baseline}$, the performances of the models were compared to determine the winning candidate models.

4.1.2.4 Model Performance Comparison

The models passing the minimum required performance criterion were considered for winning models determination. The comparison of the model performances was not done based on absolute-RMSE and was instead tested for statistical significance. The statistical significance test was applied to the aggregated error p_v . p_v is approximately χ^2 -distributed according to the Welch-Satterthwaite approximation [Net+96] with the degrees of freedom θ given by Equation (4.3).

$$\theta \approx \frac{(\sum_{k=1}^M w_k)^2}{\sum_{k=1}^M \frac{(w_k)^2}{\theta_k}}, \quad (4.3)$$

where w_k represents the weight of the database k and θ_k denotes the degrees of freedom of $RMSE_{k,v}^2$ and is given by $\theta_k = N_k - 2$, with N_k the number of PVSs in the database k . For the aggregated error p_v of model v , the statistical significance test takes the form shown in Equation (4.4).

$$t_v = \max \left(0, \frac{p_v}{p_{v_{min}}} - F(0.95, \theta, \theta) \right) \quad (4.4)$$

Here, v_{min} denotes the model with lowest error $p_{v_{min}}$ in the evaluation, $F(0.95, \theta, \theta)$ denotes the 0.95-quantile of the F -distribution with θ degrees of freedom [RC19]. If $t_v = 0$, the model v is considered to be statistically equivalent to the model v_{min} . In case that $t_v > 0$, the difference in performance between the model v_{min} and model v is termed "statistically significant".

Following this significance test, the winning models were determined.

4.1.2.5 Model Selection Procedure

For a model to be regarded as a winning candidate it had to pass the following three criteria. Firstly, the considered model should significantly outperform the baseline model described in Equation (4.2). Secondly, the model has to either perform significantly better or be statistically equivalent to the other models in the category to which the models belong. Finally, the model under consideration should perform significantly better than the models that are simpler than themselves in complexity¹.

4.1.2.6 Winning Models

Following the described procedure, winning models in five different categories were determined. The five categories in which the winning models were determined were BSM0, BSM1, BSM3, HYN0, and PXRR. In the BSM0 category, three different models were part of the winning group and in the BSM1 category, there were two different winning models. For the remaining three categories, namely, BSM3, HYN0, and

¹Complexity was defined in terms of either additional information that are required (e.g. a pixel-based NR model vs. a pixel-based hybrid NR model) or referring to the complexity of input information of similar type (e.g. RR vs. FR, with FR being more complex.)

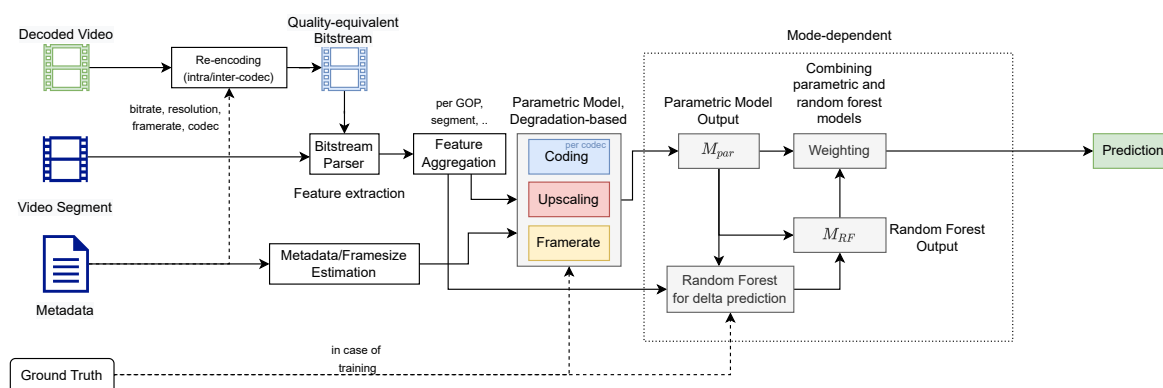


Figure 4.1: General model structure of AVQBits including all four model instances.

PXRR, there was only one winning model per category. As per the rules set out for the competition, all winning models of a certain category were required to be merged and optimized to create one model that could be finally standardized. As no agreement could be reached by the winning proponents for the BSM0 and BSM1 categories, no model was standardized in these two categories. Hence, the final outcome of the competition was the three standardized models for the categories of BSM3, PXRR, and HYN0 which resulted in ITU-T Rec. P.1204.3, P.1204.4, and P.1204.5 respectively.

The models developed by the author during the course of this work were part of the winning groups in the BSM0, BSM1, and BSM3 categories.

In the next section, a detailed description of the AVQBits models along with the training procedure and validation results is presented.

4.2 Short-term Video Quality Models: Model Description

This section presents a detailed description of the algorithms of the different instances of AVQBits focused on demonstrating the versatility of the AVQBits model in terms of scalability and adaptability regarding the available input information, starting with the standardized ITU-T P.1204.3 model. As aforementioned, in this chapter, model instances of two different types are introduced, namely, bitstream-based and hybrid. For the bitstream domain, the focus is on the ITU-T P.1204.3 standard,

4.2 Short-term Video Quality Models: Model Description

which is a Mode 3 model with access to full bitstream information, referred to as $AVQBits|M3$ in the following. Two further $AVQBits$ instances are considered for application scenarios where the full bitstream information is not available. For these cases, Mode 0 and Mode 1 variants of $AVQBits$ are proposed ($AVQBits|M0$, $AVQBits|M1$). To describe all $AVQBits$ instances, the $AVQBits|M3$ algorithm with its full Mode 3 bitstream access forms the starting point. The Mode 0 and 1 instances are implemented by synthetically generating missing model input information based on the Mode 0 or 1 type information available, as will be outlined in subsequent sections. For the case that only Mode 0 type metadata is available, but a more accurate video quality estimation is sought than what can be achieved with a Mode 0 model, a hybrid no-reference Mode 0 model instance of $AVQBits$ is proposed ($AVQBits|H0$). It has access to Mode 0 metadata and the decoded pixel information. The pixel information is used as an additional input by converting the degraded video into a “quality-equivalent bitstream” using an external video encoder and then applying the existing and unchanged full-bitstream-based $AVQBits|M3$.

The general model structure of the proposed $AVQBits$ model is shown in Figure 4.1. The approach is centred around the full-bitstream-based video quality model by the authors [Rao+20a] standardized as ITU-T P.1204.3, i.e. $AVQBits|M3$. For example, in the case of a Mode 0 or Mode 1 model, the required parts of the full-bitstream $AVQBits|M3$ model are adapted to handle the input and use the underlying other components for the final prediction. For the hybrid case, in the first iteration a quality-equivalent video bitstream mimicking the original bitstream is created.

To enable reproducibility, an open-source reference implementation of all the proposed models is made publicly available².

4.2.1 $AVQBits|M3$ / P.1204.3

All $AVQBits$ instances are based on the Mode 3 $AVQBits|M3$ model (ITU-T Rec. P.1204.3). Hence, its algorithm is described here first, followed by the different further $AVQBits$ instances. An overview of ITU P.1204.3 model is shown in Figure 4.2, which highlights the individual components of the $AVQBits$ general structure. It should be

²https://github.com/Telecommunication-Telemedia-Assessment/p1204_3_extensions

noted that the model is developed for two different target device categories, namely, PC/TV and Mobile/Tablet (MO/TA). This and all further models presented here are applicable to videos encoded with the H.264, H.265 and VP9 codecs. An extension to AV1 is currently underway. For all codecs, a corresponding bitstream parser is used to extract the relevant bitstream information as input to *AVQBits*|M3. The model consists of two components, a traditional curve-fitting-based component (referred to as the “Core Model”) and a machine-learning component, which are described in more detail in the following sections.

4.2.1.1 Core Model

The “Core Model” is based on the principle of degradation-based modeling, similar to ITU-T Rec. P.1203.1 [Raa+17]. It is initially inspired by the so-called E-model for speech quality [Joh97; Möl00; ITU09] and also based on the work on modeling television picture quality by Allnatt [All75]. In the core model, three different degradations expressed on a [0, 100] scale are considered: quantization degradation D_q , upscaling degradation D_u and temporal degradation D_t . Values on the 100-scale can be mapped to the 5-point ACR-scale used in the subjective test (i.e. the resulting mean opinion score, MOS) using the S-shaped transformation from the E-model [ITU09], as further described in Appendix C. This way, scale-compression effects of the ACR-scale at the scale ends can be avoided [Möl00], improving predictions, especially for the higher-quality range of the scale.

Quantization Degradation: D_q The observable degradation that results from the chosen quantization settings during the encoding process is termed “Quantization degradation” (D_q), see also [ITU19e; Raa+17]. This type of degradation manifests itself as blockiness or deblocking-filter-related blurring to the end-user. Since this type of degradation is dependent on the specific encoding settings, the “Core Model” handles D_q separately per codec. The number of codec categories is extended from the initial three (H.264, H.265, VP9) to five, by including the bit-depth information and splitting H.264 and H.265 into 8- and 10-bit variants.

D_q is a function of the quantization parameter (QP) used to encode the video, which is extracted as model input information using the respective bitstream parser. To

4.2 Short-term Video Quality Models: Model Description

calculate D_q , firstly, $quant$, which is the normalized value of the QP is defined, cf. Equation (4.5).

$$quant = \frac{QP_{non-Iframes}}{QP_{max}} \quad (4.5)$$

Here, QP_{max} is codec and bit-depth dependent.

$$QP_{max} = 51, \text{ if codec is H.264-8-bit or H.265-8-bit} \quad (4.6)$$

$$QP_{max} = 63, \text{ if codec is H.264-10-bit or H.265-10-bit} \quad (4.7)$$

$$QP_{max} = 255, \text{ if codec is VP9} \quad (4.8)$$

$QP_{non-Iframes}$ is the average of the QP for all non-I frames for an entire segment.

The resulting $quant \in (0, 1]$. This $quant$ value is then used to estimate mos_q , see Equation (4.9).

$$mos_q = a + b \cdot \exp(c \cdot quant + d) \quad (4.9)$$

mos_q is used to estimate D_{q_raw} , that uses $RfromMOS$ as the mapping function to map the 5-point ACR scale to a 100-point scale similar to the one recommended in ITU-T G.107 [ITU09].

$$D_{q_raw} = 100 - RfromMOS(mos_q) \quad (4.10)$$

The final D_q value is the result of constraining D_{q_raw} to $[0, 100]$ as shown in Equation (4.11).

$$D_q = \max(\min(D_{q_raw}, 100), 0) \quad (4.11)$$

Upscaling Degradation: D_u In addition to the degradations resulting from the chosen encoding settings, there are observable degradations resulting from upscaling the distorted video to the screen resolution during playback, which can be perceived by an end-user as blurriness. This kind of degradation is termed the upscaling degradation (D_u). Hence, the ‘‘Core Model’’ should be able to account for this upscaling degradation and it is assumed that this degradation is codec-independent. Due

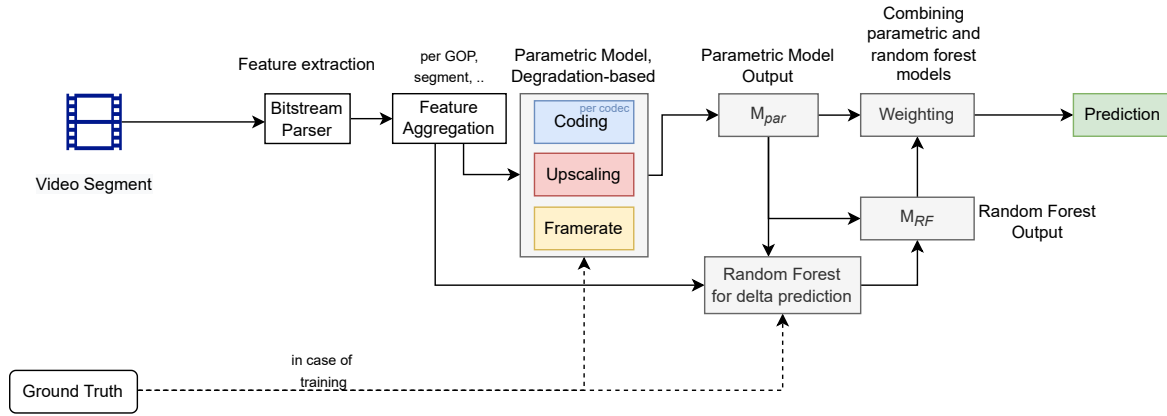


Figure 4.2: General model structure of the AVQBits|M3 / P.1204.3 model.

to the fact that in real-world streaming scenarios, upscaling is typically performed by the player software, where streaming resolutions lower than the target screen resolution typically are a result of the adaptive streaming of bandwidth-dependent representations, this degradation is assumed to be codec-independent.

$$D_{u_raw} = x \cdot \log(y \cdot scale_factor) \quad (4.12)$$

$$D_u = \max(\min(D_{u_raw}, 100), 0) \quad (4.13)$$

Equation (4.12) shows how D_u is estimated, where D_u is the $[0, 100]$ constrained value of D_{u_raw} , with \log being the natural logarithm. The $scale_factor$ is calculated according to Equation (4.14)

$$scale_factor = \frac{coding_res}{display_res} \quad (4.14)$$

as the ratio of coding and display resolution, with $display_res = 3840 \times 2160$ for PC/TV and 2560×1440 for mobile/tablet. $coding_res$ is the resolution of the encoded video and is expressed in terms of $height \times weight$. The $scale_factor$ is always limited to values $\in (0, 1]$.

Temporal Degradation: D_t Finally, the “Core Model” handles the degradations due to the adjustment of the lower framerate representations to the display framerate

4.2 Short-term Video Quality Models: Model Description

as temporal degradation (D_t). This type of degradation may be perceivable as jerkiness. Similar to upscaling D_u , it is handled in a codec-independent fashion and is estimated as follows:

$$D_{t_{raw}} = z \cdot \log(k \cdot \text{framerate_scale_factor}) \quad (4.15)$$

$$D_t = \max(\min(D_{t_{raw}}, 100), 0) \quad (4.16)$$

The temporal degradation D_t is mainly a function of the *encoded* and the *display* frame rates (the latter assumed to be constant with 60) that are combined in a *framerate_scale_factor*, cf. Equation (4.17), a value scaled in the range (0, 1]:

$$\text{framerate_scale_factor} = \frac{\text{coding_framerate}}{60} \quad (4.17)$$

4.2.1.2 Prediction

The Equation (4.18) describes the final prediction of the “Core Model”, M_{par} . Here, the described degradation-based approach is shown, using the 100-scale, $M_{p_{[0,100]}}$. The final prediction is further rescaled to a 5-point MOS-scale, Equation (4.19). Here, MOS_{fromR} is the inverse mapping from the 100-point scale to the 5-point scale, similar to the one recommended in ITU-T G.107 [ITU09].

$$M_{p_{[0,100]}} = 100 - (D_q + D_u + D_t) \quad (4.18)$$

$$M_{p_{[1,4.5]}} = MOS_{fromR}(M_{p_{[0,100]}}) \quad (4.19)$$

$$M_{par} = \text{scaleTo5}(M_{p_{[1,4.5]}}) \quad (4.20)$$

During the training of the model, the subjective scores were linearly mapped to a 4.5-point scale from the 5-point scale to avoid information loss due to the $R_{fromMOS}$ and MOS_{fromR} computations, since both of these mapping functions assume that the highest MOS that can be reached is 4.5. Hence, the coefficients predict the video quality scores on a 4.5-scale, denoted as $M_{p_{[1,4.5]}}$. Consequently, as a final step, the predictions on the 4.5-point scale are mapped back to the full 5-point scale

range using a simple linear transformation, denoted as *scalet05* (see Appendix C), Equation (4.20), resulting in the final prediction of the parametric core model M_{par} .

4.2.1.3 Machine-learning-based Video Quality Model

The second part of the model uses a machine-learning approach to estimate video quality. It is used to estimate the “residual”, that is, the part of the MOS that the parametric “core model” part is unable to predict. Hence, the target for the training for the machine-learning part of the model is the difference between MOS and the “core model” output, see Equation (4.21).

$$target_residual = MOS - M_{par} \quad (4.21)$$

This machine-learning part of the model uses Random Forest (RF) regression as the underlying machine-learning algorithm and is referred to as M_{RF} in the following. Two different RF models, one each for the PC/TV and MO/TA cases are trained.

In addition to the features the “Core Model” uses, bitstream features such as the average motion per frame, motion in the x-direction (horizontal motion), and frame sizes with frame types are extracted with the bitstream parser and employed as model input. The rationale behind this is that the parametric part is not able to fully incorporate the spatio-temporal content complexity of the video sequences. The RF model also uses the “Core Model” prediction M_{par} as an additional feature. These features are aggregated according to different functions and used as input to the random forest. The required aggregations are presented in Table 4.1. The Random Forest model used 20 trees with a fixed depth of 8. The final output was calculated as shown in Equation (4.22):

$$M_{RF} = M_{par} + predicted_residual \quad (4.22)$$

Hence, the RF-based quality prediction is the addition of the predicted residual value *predicted_residual* to the M_{par} value predicted by the core model.

4.2 Short-term Video Quality Models: Model Description

Table 4.1: Aggregated features for RF model.

Aggregated Feature	Type
Framerate	float
Resolution (<i>width · height</i>) of the distorted video	int
Codec (H.264, H.264_10bit, H.265, H.265_10bit, VP9)	boolean
M_{par}	float
Mean bitrate per segments	float
Maximum frame size	int
Kurtosis of the non-I frame sizes	float
Standard deviation of frame size of non-I frame in bits	float
Quant	float
IQR of the average QP of non-I frames	float
IQR of the minimum QP per frame	float
Kurtosis of the average QP of non-I frames	float
Mean of the average QP of non-I frames	float
Standard deviation of maximum QP of non-I frames	float
Kurtosis of the average motion per frame over all frames in a segment	float
Minimum standard deviation of motion in the x-direction (horizontal motion) per frame	float

4.2.1.4 Overall Video Quality Prediction

The overall final video quality prediction is the convex linear combination of the predictions from the parametric M_{par} and machine learning parts M_{RF} . In this case, equal weights, thus $w = 0.5$, are assigned to both of the predictions, shown in Equation (4.23). Considering Equation (4.22), it is shown that the RF residual part overall has a weight of 0.5, with the core model prediction being weighted with $0.5 + 0.5 = 1$.

$$Prediction = w \cdot M_{par} + (1 - w) \cdot M_{RF} \quad (4.23)$$

To enable reproducibility, an open-source reference implementation of the model along with the FFmpeg-based bitstream parser for all three codecs H.264, H.265, and VP9 is made available³, including also the trained random forest model.

³https://github.com/Telecommunication-Telemedia-Assessment/bitstream_mode3_videoparser

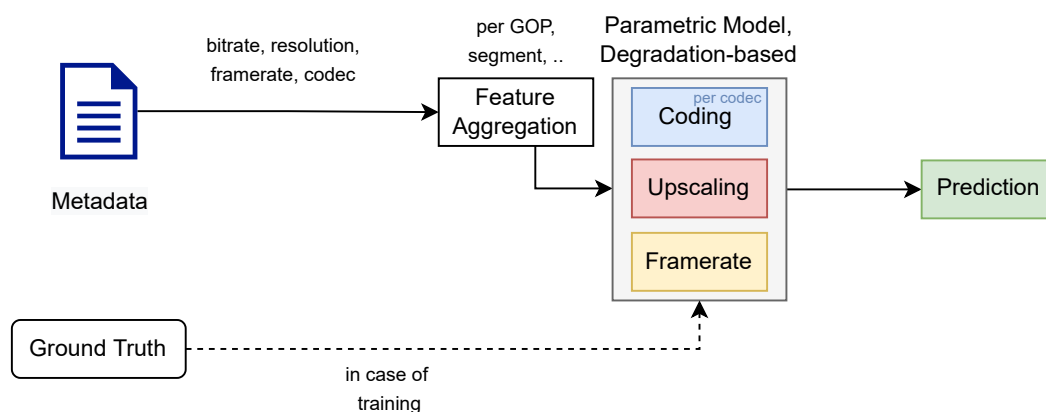


Figure 4.3: General model structure of the AVQBits|M0 model.

4.2.1.5 Per-1-Second Score Prediction

In addition to the overall per-segment video quality score, the model also outputs per-1-second scores. The per-1-second score is calculated using Equation (4.24).

$$\text{per-1-sec-score} = \frac{QP_{non-I,per-seg}}{QP_{non-I,per-sec}} \times Prediction \quad (4.24)$$

where,

- ▷ $QP_{non-I,per-seg}$ is the average QP of all non-I frames in a segment
- ▷ $QP_{non-I,per-sec}$ is the average QP of all non-I frames for each second
- ▷ $Prediction$ is per-segment video quality score described in Equation (4.23)

It should be noted that the per-1-second scores are calculated with a non-overlapping 1-sec window.

4.2.2 AVQBits|M0

A Mode 0 model is the least complex of bitstream models, both in terms of available input information and computational complexity. It has access to metadata such as bitrate, resolution, framerate, and codec information as available input for video quality estimation. The proposed Mode 0 model AVQBits|M0 instantiates the

4.2 Short-term Video Quality Models: Model Description

AVQBits model using the same general model structure as outlined for *AVQBits|M3* above and underlying ITU-T Rec. P.1204.3, with some key modifications which are indicated in Figure 4.3. The traditional curve-fitting-based part of *AVQBits|M3*, referred to as the “Core Model” in Sec. 4.2.1, is exclusively used in *AVQBits|M0*, due to the limited numbers of features available for a Mode 0 model. The RF-based model component of *AVQBits|M3* for the residual is not used in *AVQBits|M0*, and with the purely metadata-based input information, the model is not content-aware. The “Core Model” is made up of three different degradations, namely, coding/quantization degradation, upscaling degradation, and temporal degradation. For the Mode 0 instance *AVQBits|M0*, the focus is only on the quantization degradation, because this part is the only one affected by the lack of full-bitstream information.

The quantization degradation in the case of *AVQBits|M3* is a function “quantization parameter” (QP), which is codec dependent. Since in a Mode 0 model, there is usually no access to the bit-depth information as input, only three codec categories are defined, namely, H.264, H.265, and VP9, in contrast to five codec categories in *AVQBits|M3*, see Sec. 4.2.1. Accordingly, QP_{max} which is required to define *quant* as proposed in Equation (4.5) is restricted to one of the following two values based on the used codec.

$$QP_{max} = 63, \text{ if codec is H.264 or H.265} \quad (4.25)$$

$$QP_{max} = 255, \text{ if codec is VP9} \quad (4.26)$$

Because QP is not accessible as direct input information in the case of a Mode 0 model, it is approximated using the available metadata information, namely, bitrate, resolution, and framerate, see Equation (4.27).

$$\begin{aligned} QP_{pred} = & a_{qp_m0} + b_{qp_m0} \cdot \log(\text{bitrate}) \\ & + c_{qp_m0} \cdot \log(\text{resolution}) \\ & + d_{qp_m0} \cdot \log(\text{framerate}) \end{aligned} \quad (4.27)$$

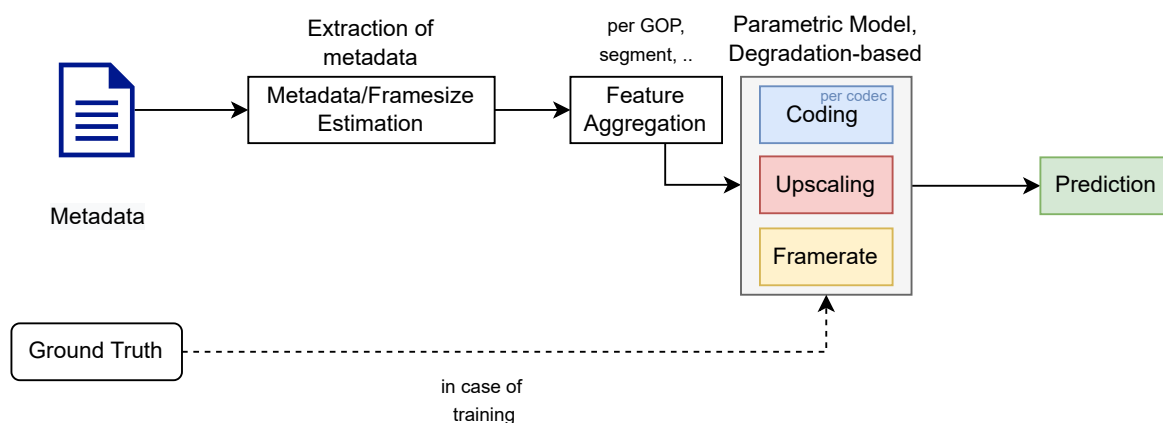


Figure 4.4: General model structure of the AVQBits|M1 model.

The resulting $quant$ is defined as in Equation (4.2.2) and is content agnostic, due to the lack of content-specific features.

$$quant = \frac{QP_{pred}}{QP_{max}} \quad (4.28)$$

Using $quant$ as defined in Equation (4.2.2), quantization degradation is calculated as described in Sec. 4.2.1.1. As a result of using QP_{pred} instead of the actual QP value as in AVQBits|M3, the coefficients related to the quantization degradation should also be re-trained by taking into account the QP_{pred} values. The training procedure and the resulting coefficients are detailed in Section 4.3.2.

4.2.2.1 Per-1-Second Score Prediction

For the AVQBits|M0 model, no separate windowing approach is used unlike in the case of AVQBits|M3 / P.1204.3 and hence the per-1-second scores are just equal to the per-segment scores.

4.2.3 AVQBits|M1

In addition to the metadata such as bitrate, resolution, framerate, and codec information, a Mode 1 model has access to frame size and frame type information. This

4.2 Short-term Video Quality Models: Model Description

information enables the inclusion of source-, and hence, content-specific features into the model. Like the Mode 0 model $AVQBits|M0$, the Mode 1 model $AVQBits|M1$ introduced in this thesis is based on the same general model structure as that of $AVQBits|M3$ (i.e. ITU-T P.1204.3) and just modifies the information pre-processing for the “Core Model”, as seen in Figure 4.4 which is a reduced variant of the general model structure shown in Figure 4.1 for clarity. Here too, the focus is on the quantization degradation D_q , as it is the only Mode-dependent part of the model.

As in Mode 0, the $AVQBits|M1$ model has been developed for three codecs, namely, H.264, H.265, and VP9, for one single bit-depth, as the exact profile usually cannot be known based on the available input information. Hence, the same QP_{max} values are used for the three categories as for $AVQBits|M0$, cf. Equations (4.25) and (4.26).

For the purpose of QP estimation, two new features using the framesize and frame-type information are defined in Equations (4.29) and (4.30).

The feature $fsratio$ represents the ratio between the average sizes of I-frames and non-I-frames for a given segment under consideration:

$$fsratio = \frac{1/N_I \sum_i (S_{I,i})}{1/N_{nI} \sum_j (S_{nI,j})} \quad (4.29)$$

Here, $S_{I,i}$ is the size of I-frame i , $S_{nI,j}$ is the size of a non-I-frame, that is, P- or B-frame, which are treated alike for this calculation, with index j . N_I is the overall number of I-frames, N_{nI} is the overall number of non-I-frames. Like for $AVQBits|M3$, all I-frames i and non-I-frames j belonging to a given segment under consideration are used.

The second feature introduced is the mean size of non-I-frames ms_{nI} .

$$ms_{nI} = 1/N_{nI} \sum_j S_{nI,j} \quad (4.30)$$

As in $AVQBits|M1$, QP_{pred} is calculated according to Equation (4.31).

$$\begin{aligned}
 QP_{pred} = & a_{qp_m1} \\
 & + b_{qp_m1} \cdot \log(ms_nI) \\
 & + c_{qp_m1} \cdot \log(resolution) \\
 & + d_{qp_m1} \cdot \log(framerate) \\
 & + e_{qp_m1} \cdot \log(fsratio)
 \end{aligned} \tag{4.31}$$

Considering the QP estimation of qp_m1 following $AVQBits|M0$, the quantization degradation is retrained to be Mode 1 specific. The details of the training procedure and the final coefficients are described in Section 4.3.3.

4.2.3.1 Per-1-Second Score Prediction

Like the $AVQBits|M0$ model, the per-1-second scores for the $AVQBits|M1$ model are just equal to the per-segment scores.

4.2.4 AVQBits|H0

As mentioned earlier, a hybrid no-reference Mode 0 model has access to both the metadata and the decoded pixel information of the distorted video to estimate video quality.

The main idea of the proposed model is to create a “quality equivalent bitstream” (QEB) which is similar to the original bitstream using the decoded pixels and the provided metadata. After the QEB is created, the $AVQBits|M3$ model (i.e. ITU-T P.1204.3) is applied with slight changes. A somewhat related approach has been used in ITU-T Rec. P.563 [ITU04] to provide a more general description of the received speech quality, which is given by comparing the input signal with a pseudo reference signal generated by a speech enhancer.

The process of creating the QEB is shown in Figure 4.1 and Figure 4.5, wherein the provided metadata such as bitrate, resolution, framerate, and codec information is used. The distorted video is re-encoded with the encoding settings corresponding

4.2 Short-term Video Quality Models: Model Description

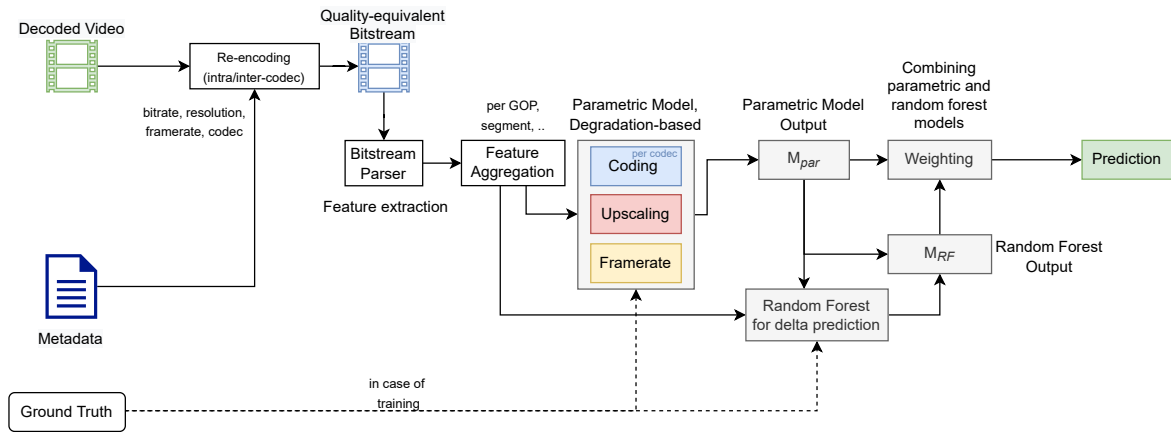


Figure 4.5: General model structure of the $AVQBits|H0$ model.

to the metadata following a 1-pass encoding strategy. This is based on the results reported in Stankowski et al. [Sta+13] that the quality loss across different QP values remains constant for the second round of encoding, which can be compensated by the model due to the use of QP as the feature for quality estimation in the Core Model. Furthermore, the QP that an encoder chooses for a bitrate-resolution during the QEB generation process will be in the same range as that of the initial encoding due to the same bitrate and resolution settings.

In the following, two variants of the hybrid no-reference Mode 0 model $AVQBits|H0$ are proposed. These variants are based on the codec used to re-encode the video and are referred to as the “same” and “fixed” codec variants. In the case of the “same” codec variant, hereafter referred to as $AVQBits|H0|s$, the QEB is created using the codec specified by the metadata. After the QEB is created, the $AVQBits|M3$ model ITU-T P.1204.3 is directly applied to estimate the video quality without any modifications.

The second, “fixed” codec variant, is referred to as $AVQBits|H0|f$ in the following. By using a fixed, pre-defined codec to create the QEB, no stream-specific encoding and then bitstream parsing is needed, reducing the complexity of the implementation. After the creation of the QEB, the $AVQBits|M3$ model ITU-T P.1204.3 estimates the video quality. For the proof-of-concept of $AVQBits|H0|f$ presented in this thesis, H.265 is selected as the codec to create the QEB, irrespective of the codec used to generate the original bitstream. Different codecs have a different impact on quality

for a given specific setting. Since H.265 is used to create the QEB irrespective of the codec to generate the original bitstream, the initially estimated quality of this QEB may not optimally reflect the impact of the original codec. As a result, to estimate the final quality score, a simple linear mapping function is proposed that takes into account the impact of the original codec on quality to map the initial prediction of the ITU-T P.1204.3 model to the respective codec characteristics.

$$Prediction = a_{cmap} \cdot Prediction_{M3} + b_{cmap} \quad (4.32)$$

where, $Prediction_{M3}$ is the prediction from the AVQBits|M3 P.1204.3 model and a_{cmap} and b_{cmap} are codec-specific mapping coefficients. The coefficient values are provided in Section 4.3.4.

It is noted that besides this instance of the “fixed” codec hybrid model variant presented as a proof-of-concept, other realizations can be conceived. For example, a more sophisticated codec-specific mapping function can be developed instead of the simple linear mapping as proposed in this work. Further, in principle also another of the three encoders and hence bitstream-parsers can be used to create and analyze the QEB. With H.265, the newest of the three codecs was selected, and currently developed updates of the proposed models can use even newer codecs such as AV1 or VVC.

The impact of the “fixed” instead of the “same” codec variant on quality prediction accuracy is extensively analyzed in Sec. 4.4.2.

4.3 Short-term Video Quality: Model Training

This section details the training procedure that was performed to obtain the coefficients for the different AVQBits models.

4.3.1 AVQBits|M3 / P.1204.3

The presented model coefficients in Tables 4.2, 4.3, 4.4 and 4.5 are based on the 26 databases that were designed as part of the P.NATS Phase 2 competition and presented as PNATS-UHD-1 in Chapter 3.

During the training of the model, the subjective scores were linearly mapped to a 4.5-point scale from the 5-point scale in order to avoid information loss due to the $R_{fromMOS}$ and MOS_{fromR} computations, since both of these mapping functions assume that the highest MOS that can be reached is 4.5. Hence, the coefficients predict the video quality scores on a 4.5-scale, denoted as $M_{p_{[1,4.5]}}$. As a final step, the predictions on the 4.5-point scale were mapped back to the full 5-point scale range using a simple linear transformation, denoted as “scaleto5”, Equation (4.20), resulting in the final prediction of the parametric core model M_{par} .

In the following, the required coefficients for all parts are summarized. The quantization-degradation-related coefficients are presented for the PC/TV case in Table 4.2, and for the mobile/tablet case in Table 4.3. The upscaling- and temporal-degradation-related coefficients are reported in Table 4.4 for the PC/TV case and in Table 4.5 for the mobile/tablet case.

Table 4.2: AVQBits / P.1204.3 Quantization-degradation coefficients, PC/TV case.

Codec	a	b	c	d
H.264	4.4344	-1.7058	4.9654	-4.1203
H.264-10bit	4.6467	-0.8091	5.9835	-4.4398
H.265	4.3789	-1.0208	5.7572	-4.5625
H.265-10bit	4.5458	-0.866	6.1116	-3.3828
VP9	4.3404	-0.9961	4.5282	-3.9641

Table 4.3: AVQBits / P.1204.3 Quantization-degradation coefficients, MO/TA case.

Codec	a	b	c	d
H.264	4.4365	-1.4909	5.4251	-4.5198
H.264-10bit	4.5399	-0.414	6.2249	-4.2599
H.265	4.3089	-0.6685	6.0551	-4.6974
H.265-10bit	4.9999	-2.6821	1.5069	-1.7664
VP9	4.4024	-1.2504	2.9268	-3.0087

Table 4.4: Upscaling- and temporal-degradation coefficients, PC/TV case.

x	y	k	z
-9.5497	1.1999	4.1696	-8.3084

Table 4.5: Upscaling- and temporal-degradation coefficients, MO/TA case.

x	y	k	z
-8.4690	1.1999	4.2701	-6.3648

4.3.2 AVQBits|M0

For the AVQBits|M0 model instance, a two-step training procedure was implemented to estimate the coefficients related to QP_{pred} and *quantization degradation*. In the first step, the QP_{pred} prediction module as described in Equation (4.27) was trained using the true QP values extracted from the 764 PVSs of AVT-PNATS-UHD-1 as ground-truth. The resulting coefficients for determining QP_{pred} are detailed in Table 4.6.

Following this, the coefficients in Table 4.6 were used to estimate QP values, and the resulting estimates QP_{pred} were used as input to the *quantization degradation*. The new coefficients of the core model were obtained by training the model using the subjective MOS from AVT-PNATS-UHD-1 as ground truth. The resulting new coefficients of the core model are as shown in Table 4.7.

To estimate the *upsampling degradation* and *temporal degradation* component of the “Core Model”, the coefficients for the AVQBits|M3 / P.1204.3 model reported in Table 4.4 are used.

Table 4.6: QP-Prediction coefficients for AVQBits|M0, PC/TV case.

Codec	a_{qp_m0}	b_{qp_m0}	c_{qp_m0}	d_{qp_m0}
H.264	-5.7284	-5.3586	4.1965	5.6231
H.265	-7.6866	-6.0256	4.8298	4.0869
VP9	-140.8384	-46.5290	37.5395	27.5876

As there were no MO/TA databases that were part of the training dataset, AVT-PNATS-UHD-1, a synthetic dataset consisting of AVQBits|M3 / P.1204.3 was devel-

4.3 Short-term Video Quality: Model Training

Table 4.7: Quantization-degradation coefficients for $AVQBits|M0$, PC/TV case.

Codec	a	b	c	d
H.264	4.7342	-0.9469	4.0831	-2.0624
H.265	4.5731	-0.6835	3.3163	-1.4604
VP9	4.2624	-0.6135	3.2368	-2.2657

oped to determine the coefficients for the MO/TA case. These coefficients can be found in the reference implementation that is publicly available.

4.3.3 $AVQBits|M1$

For the $AVQBits|M1$ model instance, the same two-step training approach as for Mode 0 was used to determine the coefficients related to QP_{pred} and subsequently the *quantization degradation* component D_q of the “Core Model”, cf. Equation (4.9) to Equation (4.11). The coefficients related to QP_{pred} and the *quantization degradation* D_q (i.e. mos_q at first, cf. Equation (4.9)) are presented in Tables 4.8 and 4.9, respectively. Similar to the Mode 0 model $AVQBits|M0$, the coefficients for the $AVQBits|M3$ / P.1204.3 model detailed in Table 4.4 are used to estimate the *upsampling degradation* and *temporal degradation*.

Table 4.8: QP-Prediction coefficients for $AVQBits|M1$, PC/TV case.

Codec	a_{qp_m0}	b_{qp_m0}	c_{qp_m0}	d_{qp_m0}	e_{qp_m0}
H.264	28.4333	-7.3951	5.7821	0.2479	-5.4537
H.265	22.3936	-6.5297	5.1573	-0.8999	-2.2889
VP9	92.1245	-51.1209	40.6832	-10.2195	-18.7809

Table 4.9: Quantization-degradation coefficients for $AVQBits|M1$, PC/TV case.

Codec	a	b	c	d
H.264	4.6602	-1.1312	4.2268	-2.4471
H.265	4.5375	-0.6829	3.5053	-1.6074
VP9	4.5253	-1.2635	2.0732	-1.8051

A similar approach to determine the MO/TA coefficients for the $AVQBits|M0$ model was used for the $AVQBits|M1$ model and the corresponding coefficients can be found in the publicly available reference implementation.

4.3.4 *AVQBits|H0*

As discussed in Section 4.2.4, two different variants of the *AVQBits|H0* model are proposed in this thesis. The first, *AVQBits|H0|s*, applies the same encoder used for initially encoding the video to be evaluated, plus Mode 0 data for a quality-equivalent re-encoding of the video. The *AVQBits|M3 / P.1204.3* model is then directly applied to the resulting bitstream, without any further modifications to the model. Hence, no additional training of the *AVQBits|H0|s* model is needed.

Instead of the original video codec, that has been used for encoding the distorted video, the *AVQBits|H0|f* model has a fixed video encoder for generating the quality-equivalent bitstream. For the model instance presented in this thesis, H.265 is selected. As a result, the prediction from the *AVQBits|M3 / P.1204.3* model requires a codec-specific mapping of the predicted score to represent the quality that would be provided by the originally applied encoder. Hence, a simple mapping function as described in Equation (4.32) is proposed. MOS data from AVT-PNATS-UHD-1 are applied as training targets to determine the coefficients of the mapping function. The resulting coefficients are presented in Table 4.10.

Table 4.10: Codec mapping coefficients for *AVQBits|H0|f*, PC/TV case.

Codec	a_{cmap}	b_{cmap}
H.264	0.9053	0.0931
VP9	0.8530	0.6979

4.4 Short-term Video Quality: Model Evaluation

This section focuses on the evaluation of the different models making up *AVQBits* and is divided into two parts. In the first part, details of the validation of *AVQBits|M3 / P.1204.3* as part of the *PNATS Phase 2* competition is described. Following this, a detailed evaluation of all four models of *AVQBits* using the AVT-VQDB-UHD-1 dataset described in Chapter 3 and a comparison with SoA models is presented.

4.4.1 Validation of AVQBits|M3 / P.1204.3

The models submitted to the *P.NATS Phase 2* competition were validated using the 13 validation databases of PNATS-UHD-1 described in Chapter 3. The results of the validation process and the performance of both the submitted and finally standardized BSM3 model in comparison with the other two standard models are described in Table 4.11.

Table 4.11: Aggregated RMSE on validation and on all databases (training and validation databases according to (Equation (4.1))) of the models submitted to the competition, and the standardized (re-trained) versions of the models [Raa+20a].

Model	Validation DBs	All DBs
Submitted Bistream Model	0.429	0.421
P.1204.3 Standard	0.397	0.394
Submitted Pixel RR Model	0.448	0.444
P.1204.4 Standard	0.415	0.418
Submitted Hybrid NR Model	0.451	0.452
P.1204.5 Standard	0.442	0.440

Following the statistical significance test outline in Section 4.1.2.4, it can be concluded from Table 4.11 that the ITU-T P.1204.3 (BSM3) model significantly outperforms both the ITU-T P.1204.4 (PXRR) and ITU-T P.1204.5 (HYN0) models.

In addition to this, a comparison of the submitted BSM3 model with the submitted PXRR and HYN0 models and other SoA FR models is presented in Table 4.12. The procedure of per-database linear mapping before determining RMSE as used in the *P.NATS Phase 2* competition was used to determine the RMSE of the three submitted models and VMAF. As both PSNR and SSIM show a non-linear relationship to MOS, a 3rd-order polynomial mapping is applied per database before computing the RMSE values. In addition to RMSE, PCC values were calculated for each of the models for comparison purposes. It can be seen from the table that the ITU-T P.1204.3 (BSM3) model significantly outperforms all the considered models.

Table 4.12: Overall model performance of different models on P.NATS Phase 2 validation databases only. Left: All HRCs. Right: Only HRCs where the HRC framerate corresponds to that of the SRC [Raa+20a].

Model	All HRCs			HRCs using SRC fps		
	RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
PSNR	0.716	0.630	0.615	0.688	0.625	0.609
SSIM	0.648	0.609	0.704	0.580	0.665	0.725
VMAF	0.611	0.761	0.773	0.548	0.794	0.790
P.1204.3	0.422	0.899	0.883	0.429	0.891	0.875
P.1204.4	0.441	0.889	0.872	0.440	0.884	0.864
P.1204.5	0.448	0.885	0.880	0.447	0.880	0.871

4.4.2 Evaluation of AVQBits Model Instances

For evaluating the model instances of *AVQBits*, the publicly available AVT-VQDB-UHD-1 dataset [Rao+19a] consisting of 756 PVSs is used, see also Sec. 3.1.1.3. Note that only 432 PVSs are publicly available due to source copyright issues. However, in this thesis, the evaluation is performed on the entire dataset consisting of 756 PVSs, as the author has access to the complete set. Table 4.13 provides a detailed overview of the performance of the model instances *AVQBits*|M3 / P.1204.3, *AVQBits*|M0, *AVQBits*|M1 and the two versions of the Hybrid Mode 0 model *AVQBits*|H0. Performance is given in terms of RMSE, Pearson Correlation Coefficient (PCC), Spearman Rank Order Correlation Coefficient (SROCC), Kendall correlation, and R^2 score for the four tests individually and all databases together. As is expected, *AVQBits*|M3 / P.1204.3 outperforms all other model instances for all databases combined as it has access to the entire bitstream to estimate video quality. An interesting observation is that the other model instances perform slightly better than *AVQBits*|M3 / P.1204.3 for *test_4*. This specific test considers a wide range of framerate variations. It should be noted that such a high variation in framerate between SRC and PVS is rather unrealistic for HAS applications. However, it can be seen from Table 4.13, that *AVQBits*|M3 / P.1204.3 performs significantly better across all databases.

Figure 4.6 show the scatter plots for all models. It can be observed that *AVQBits*|M3 / P.1204.3 leads to very few outliers as compared to the subjective tests, whereas results for the other instances show a larger number of outliers. Most notably, it can be observed that for Mode 1 *AVQBits*|M1, the *Surfing* sequence suffers from

4.4 Short-term Video Quality: Model Evaluation

Table 4.13: Performance of the *AVQBits* instances on the AVT-VQDB-UHD-1 dataset (*The RMSE and R^2 numbers for *AVQBits*|M3 / P.1204.3 may differ from the ones reported in [Rao+20a], as here the RMSE and R^2 values after linear mapping whereas in [Rao+20a] the RMSE and R^2 values were calculated on raw predictions).

Database	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
test_1	<i>AVQBits</i> M3 / P.1204.3*	0.280	0.968	0.953	0.822	0.937
test_1	<i>AVQBits</i> M1	0.614	0.836	0.851	0.677	0.699
test_1	<i>AVQBits</i> M0	0.507	0.891	0.888	0.703	0.795
test_1	<i>AVQBits</i> H0 s	0.298	0.964	0.954	0.817	0.929
test_1	<i>AVQBits</i> H0 f	0.324	0.957	0.946	0.805	0.916
test_2	<i>AVQBits</i> M3 / P.1204.3*	0.287	0.966	0.960	0.830	0.934
test_2	<i>AVQBits</i> M1	0.441	0.918	0.930	0.780	0.844
test_2	<i>AVQBits</i> M0	0.511	0.889	0.895	0.714	0.790
test_2	<i>AVQBits</i> H0 s	0.394	0.936	0.934	0.782	0.875
test_2	<i>AVQBits</i> H0 f	0.451	0.915	0.916	0.752	0.837
test_3	<i>AVQBits</i> M3 / P.1204.3*	0.324	0.957	0.935	0.785	0.917
test_3	<i>AVQBits</i> M1	0.363	0.946	0.924	0.766	0.895
test_3	<i>AVQBits</i> M0	0.464	0.911	0.896	0.712	0.830
test_3	<i>AVQBits</i> H0 s	0.395	0.936	0.920	0.756	0.877
test_3	<i>AVQBits</i> H0 f	0.450	0.916	0.908	0.737	0.840
test_4	<i>AVQBits</i> M3 / P.1204.3*	0.485	0.876	0.853	0.681	0.767
test_4	<i>AVQBits</i> M1	0.366	0.931	0.911	0.756	0.867
test_4	<i>AVQBits</i> M0	0.443	0.897	0.851	0.673	0.805
test_4	<i>AVQBits</i> H0 s	0.460	0.889	0.876	0.699	0.790
test_4	<i>AVQBits</i> H0 f	0.405	0.915	0.898	0.734	0.837
All	<i>AVQBits</i> M3 / P.1204.3*	0.370	0.942	0.927	0.768	0.887
All	<i>AVQBits</i> M1	0.476	0.901	0.900	0.730	0.811
All	<i>AVQBits</i> M0	0.499	0.890	0.877	0.684	0.792
All	<i>AVQBits</i> H0 s	0.408	0.928	0.919	0.755	0.861
All	<i>AVQBits</i> H0 f	0.433	0.919	0.909	0.743	0.844

under-prediction in a few cases due to a large error in the method used for QP estimation in this case. In addition, it can also be seen that *AVQBits*|M0 suffers from slightly more over-prediction, which is a result of the lack of source-specific information for quality estimation, which *AVQBits*|M3 and hence also *AVQBits*|H0 partly handle in the random forest model part. Also, *AVQBits*|H0|f would benefit from a more sophisticated codec mapping than the linear one defined in Section 4.2.4 to better take into account codec-specific differences.

The performance of the *AVQBits* instances is also compared with that of SoA models. For this purpose, the performance numbers for SoA models on the AVT-VQDB-UHD-1 dataset reported in [Raa+20a] are used. In Tables 4.14 and 4.15, different FR and NR models are compared with the proposed models for tests without and with framerate variation respectively. As can be seen from the results, *AVQBits*|M3 / P.1204.3 is the best performing model across all tests, with VMAF being the best performing

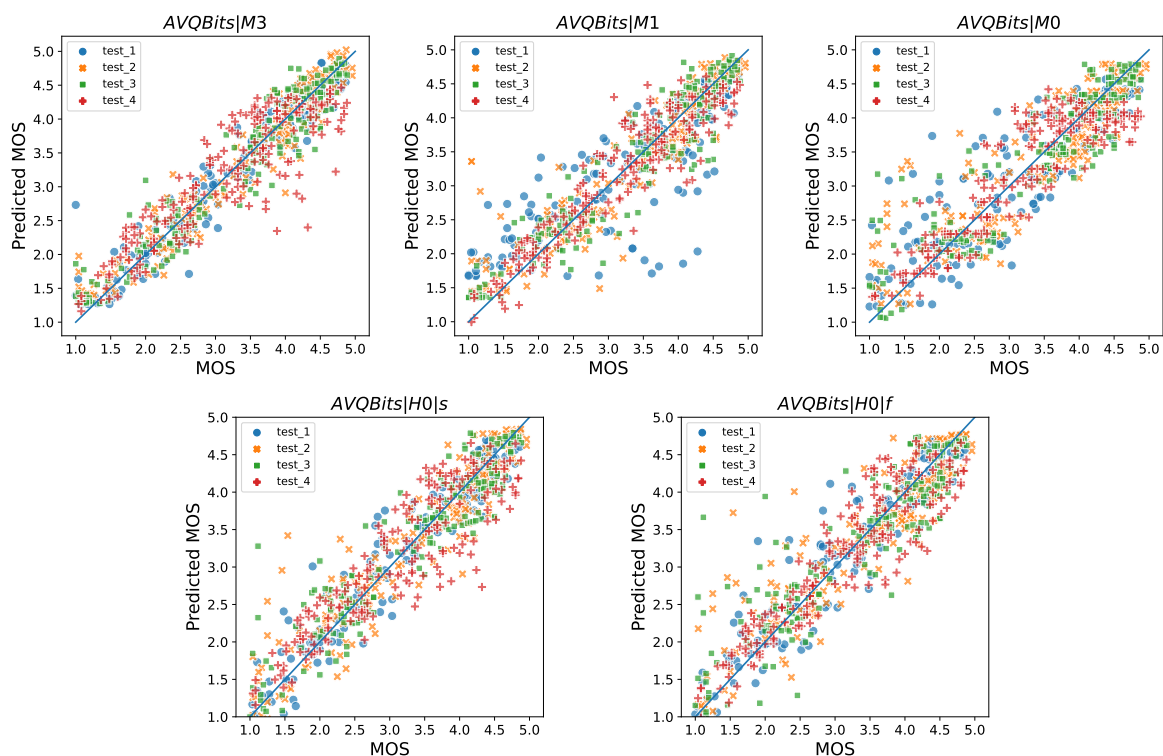


Figure 4.6: Scatter plot of AVQBits instances for AVT-VQDB-UHD-1 dataset.

FR model. Despite the reduced input data for these models, the other AVQBits instances are still able to outperform a number of the SoA models. For example, AVQBits|M0 shows a better performance than Brisque and SSIM, or AVQBits|M1 shows a better performance than VMAF. The hybrid models also outperform VMAF and generally are surpassed only by AVQBits|M3 / P.1204.3 in performance. It is noted that the good performance of the AVQBits instances other than AVQBits|M3 / P.1204.3 and AVQBits|H0 may be due to the selected specific encoding settings and their range. In general, test_4 seems to be the most difficult test in terms of estimating video quality, due to the wide range of framerates included in this test. The comparatively bad performance of VMAF for test_4 can be attributed to the lack of a sophisticated motion-related feature in the model.

Table 4.14: Performance comparison of the *AVQBits* instances with SoA models for tests in the AVT-VQDB-UHD-1 dataset without framerate as dependent variable (*The RMSE and R^2 numbers for *AVQBits*|M3 / P.1204.3 may differ to the ones reported in [Rao+20a], as here the RMSE and R^2 values after linear mapping are shown, whereas in [Rao+20a] the RMSE and R^2 values were calculated on raw predictions).

Model	RMSE	PCC	SROCC	Kendall	R^2 Score
VMAF [Net18]	0.531	0.880	0.889	0.721	0.774
Brisque [MMB12]	0.653	0.815	0.838	0.653	0.660
NIQE [MSB13]	1.009	0.432	0.445	0.301	0.187
PSNR	1.109	0.131	0.682	0.531	0.017
SSIM [Wan+04]	0.956	0.520	0.761	0.569	0.270
MS-SSIM [WSB03]	0.896	0.599	0.752	0.563	0.358
ADM2 [Li+11]	0.580	0.855	0.874	0.698	0.731
VIFP [SB06]	0.757	0.736	0.756	0.562	0.542
<i>AVQBits</i> M3 / P.1204.3*	0.306	0.962	0.948	0.804	0.925
<i>AVQBits</i> M1	0.486	0.901	0.904	0.738	0.812
<i>AVQBits</i> M0	0.503	0.894	0.891	0.701	0.799
<i>AVQBits</i> H0 s	0.373	0.943	0.935	0.778	0.889
<i>AVQBits</i> H0 f	0.439	0.920	0.914	0.749	0.846

Table 4.15: Performance comparison of *AVQBits* instances with SoA models for tests with framerate as independent variable in the AVT-VQDB-UHD-1 dataset (*The RMSE and R^2 numbers for P.1204.3 may differ to the ones reported in [Rao+20a], as here the RMSE and R^2 values after linear mapping are shown, whereas in [Rao+20a] the RMSE and R^2 values were calculated on raw predictions).

Model	RMSE	PCC	SROCC	Kendall	R^2 Score
VMAF [Net18]	0.592	0.807	0.811	0.624	0.652
Brisque [MMB12]	0.641	0.813	0.833	0.646	0.657
NIQE [MSB13]	1.006	0.393	0.387	0.265	0.154
PSNR	1.004	0.313	0.491	0.352	0.000
SSIM [Wan+04]	0.871	0.497	0.580	0.418	0.247
MS-SSIM [WSB03]	0.832	0.559	0.581	0.421	0.312
ADM2 [Li+11]	0.598	0.803	0.806	0.615	0.644
VIFP [SB06]	0.789	0.618	0.612	0.449	0.381
<i>AVQBits</i> M3 / P.1204.3*	0.485	0.876	0.853	0.681	0.767
<i>AVQBits</i> M1	0.366	0.931	0.911	0.756	0.867
<i>AVQBits</i> M0	0.443	0.897	0.851	0.673	0.805
<i>AVQBits</i> H0 s	0.460	0.889	0.876	0.699	0.790
<i>AVQBits</i> H0 f	0.405	0.915	0.898	0.734	0.837

4.5 Other Prototype Models

In addition to the different instances of *AVQBits*, other models focusing on mainly Mode 0 and Hybrid models have been developed as part of this work. These models are to be seen as pre-cursors to the corresponding instances of *AVQBits* and are partly based on other previously reported models such as VMAF [Net18] and ITU-T Rec. P.1203.1 Mode 0 [ITU19e].

4.5.1 ITU-T P.1203.1 Mode 0 Extension

The ITU-T P.1203.1 Mode 0 [ITU19e] model was developed for videos encoded with H.264 and resolutions up to 1080p and framerates up to 30 fps. However, due to the proliferation of newer codecs such as H.265, VP9, and AV1 and also a widespread capture and streaming of videos of resolutions higher than 1080p and framerates higher than 30 fps, it was necessary to extend the P.1203.1 Mode 0 model. With the motivation of predicting video quality for this newer parameter space, before the newly expected ITU-T P.1204 series of Recommendations were finalized, the following extension was proposed.

4.5.1.1 Proposed P.1203.1 Mode 0 Extension

The subjective ratings from test_1 of AVT-VQDB-UHD-1 and the AV1 dataset described in Chapter 3 were used to derive a mapping/correction function for the ITU-T P.1203.1 Mode 0 model to handle new codecs, resolution, and framerate. As a pre-processing step, MOS were obtained by averaging the individual ratings overall individual sources to eliminate content dependency as the mode 0 model is unable to handle this content dependency in a meaningful way. Various extensions were developed using different parameters such as bitrate, resolution, and codec as input parameters in several combinations. In addition to these parameters, the output of P.1203.1 mode 0 model, referred as to *mode0_output*, was also used as an additional input to the mapping/correction function. While computing the P.1203.1 mode 0 output for the subjective data, appropriate changes in terms of codec and resolution handling were made to the existing model to take into account newer codecs and resolution. Different possible input parameters were analyzed, e.g. only {codec}, only {resolution, codec}, only {bitrate} and {resolution, codec, bitrate}. It was found that {resolution, codec, bitrate} parameters are required for good performance of such a correction function since modern codecs are designed to handle lower and higher resolutions with varying bitrates differently than H.264.

Curve fitting was used for training. The software is based on Python 3 and uses LmFit⁴. Several candidate functions were checked to determine the best perform-

⁴<https://lmfit.github.io/lmfit-py/>

ing function. Out of these candidates, the below mentioned candidate, see Equation (4.33), was the best performing extension.

The final correction/mapping function is given by Equation (4.33):

$$\text{predicted_mos} = a + b * \text{mode0_output} + c * \log(\text{bitrate}) + d * \log(\text{resolution}) \quad (4.33)$$

For each video codec, a different set of coefficients is used. In Table 4.16 all coefficients are summarized. To determine the coefficients, the AV1 dataset and test_1 of AVT-VQDB-UHD-1 described in Chapter 3 are used.

Table 4.16: Correction Mapping - Coefficients per codec.

Codec	a	b	c	d
H.264	-0.19	0.04	0.72	-0.18
H.265	0.05	0.47	0.40	-0.09
VP9	-3.55	-0.008	0.43	0.25
AV1	-7.38	-0.18	0.46	0.54

4.5.1.2 Evaluation of the proposed P.1203.1 Mode 0 extension

In this section, a comparison of the correction method with non-adjusted P.1203.1 Mode 0 is conducted. At first, the non-adjusted P.1203 Mode 0 Pv model was used to obtain the predicted MOS for the two subjective tests that were conducted. This was done to ascertain if the existing model works well for all the codec, resolution, and framerate extensions or if there is a need for a correction mapping to handle these extensions. The evaluation of the unadjusted model showed that the performance in terms of rmse is as follows for the three new codecs: $\text{rmse}_{h264} = 0.66$, $\text{rmse}_{h265} = 0.65$, $\text{rmse}_{vp9} = 0.69$ and $\text{rmse}_{av1} = 0.9$. These values were obtained by using the coefficients reported in the corresponding standard [ITU19e] for all three codecs. These values diverge considerably from the rmse of 0.465 for mode 0 as reported in the standard [ITU19e] even considering only H.264. Even for the case of H.264, the difference between the rmse reported in the standard and the one for our test is also

considerably large. This deviation in performance is expected as the existing model was trained for H.264 for only resolutions up to 1080p and framerates up to 24fps. For the model to take into account the new input data, either the model has to be re-trained or else a correction mapping can be done.

The decision to go for a simple correction mapping over re-training the model was based on the rationale that such a correction mapping would keep the structure and coefficients of the original model intact and ensures that the development process can rely on the well-developed P.1203 models. Moreover, it was planned as an intermediate solution, as the *P.NATS Phase 2* competition was underway to develop the ITU-T P.1204 series of recommendations. Keeping in view this rationale, a correction mapping that takes the output of the original mode 0 model as an input and performs the correction based on bitrate, resolution, and framerate is proposed. Figure 4.7 shows the performance of the correction mapping for all the four codecs. It can be seen that the RMSE is lower than 0.3 for all the codecs for the mapped version, and hence it can be concluded that the developed method is better in terms of RMSE than the original non-adjusted model for the new application scenarios.

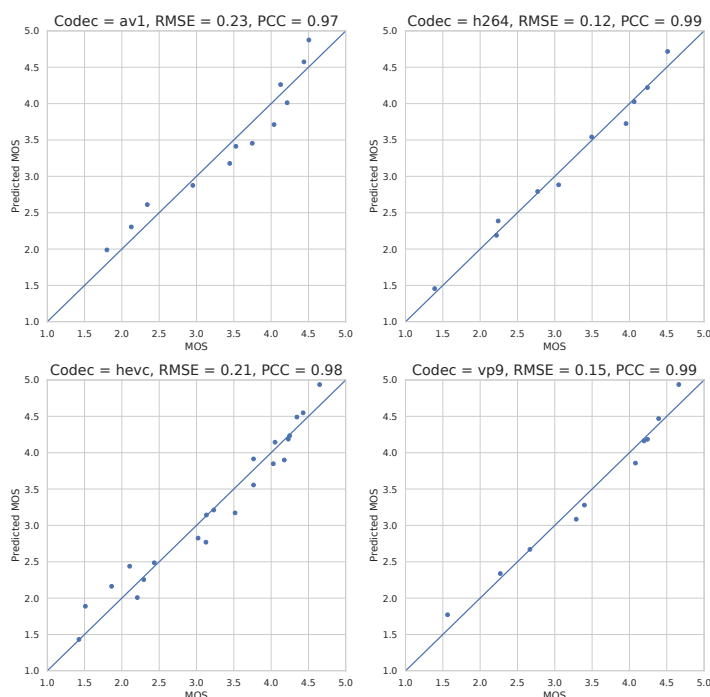


Figure 4.7: Performance of the correction mapping for all considered video codecs.

4.5.2 Hybrid-VMAF

With the objective of proposing an approach to extend existing SoA models using additional metadata that can be extracted from a video to improve prediction accuracy, a hybrid version of VMAF using metadata such as video resolution and framerate was developed and is presented in the following section.

4.5.2.1 Proposed Approach

The general model structure is shown in Figure 4.8. The approach starts with extracting pixel-based features from the video input. In the case of an underlying full-reference model, the video inputs are the source video and the distorted video. Whereas, in the case of other model types, the video input can be only the distorted video or different variants of reference video information and distorted video. In this prototype, a full-reference model is considered as the underlying pixel model for the hybrid extension. To prove the validity of the concept, the popular full-reference model, Netflix’s VMAF, is used as a feature extractor. Similar extensions can also be used for no- or reduced-reference models.

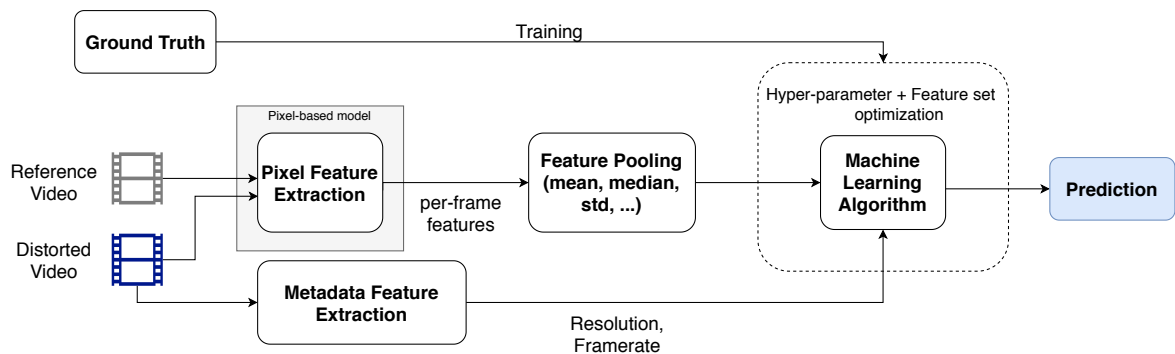


Figure 4.8: General Machine Learning Pipeline, indicating the involved steps from pixel-based feature extraction based on any kind of pixel-based model, to temporal feature pooling to the final training of the included machine learning algorithm.

After feature extraction, the per-frame feature values are temporally pooled to obtain the final per-video features. Temporal feature pooling is a well-known approach to remove the time dependency of short video sequences and therefore to provide a constant number of features to the underlying machine learning model. Similar methods have been used in other models, e.g. nofu [GRR19; Gör+21a]. Such a

temporal feature pooling can range from simple arithmetic mean to more complex methods such as harmonic mean, Minkowski summation, percentile, etc. For the presented model, the focus is on the widely used and simple, arithmetic mean.

In addition to the pooled features, resolution and framerate are included as metadata features, resulting in a hybrid extension of the considered FR model. FFprobe which is part of FFmpeg is used to estimate the required metadata.

In the hyperparameter and feature optimization step, a large combination of (*hyperparameter, feature set*) values are considered while training the machine learning algorithm for enhancing the pixel-based model using metadata. The combination which performs best in terms of Pearson correlation coefficient was chosen for the final model.

The AVT-VQDB-UHD-1 dataset is used to train and validate the proposed hybrid-VMAF model.

4.5.2.2 Evaluation

In total, three machine learning algorithms, namely, support vector regression (SVR), random forest (RF), and extreme gradient boosting trees (XGBoost), are considered for evaluation. Besides the evaluation of different machine learning approaches, the hyper-parameters and feature sets of the different hybrid-FR models are further optimized. These machine learning algorithms were selected because they have been used already in other models, e.g. SVRs in VMAF, RF in P.1203/1204.3 [ITU19e; ITU19b] and extreme gradient boosting trees to predict the number of video encoding passes in [GRR20] and for video quality modeling in [Gör+21a]. Other models based on neural networks are possible, however, due to the low number of training samples within the used databases, that choice is out of scope in this context.

In the case of RF and XGBoost-based models, only the *number of trees* parameter was considered for hyper-parameter optimization. Otherwise, default values were used for all other parameters as included in the scikit-learn and xgboost implementation of RF and XGBoost, respectively. For the proof-of-concept, no hyperparameter optimization for the SVR-based approach was employed. For SVR, the radial basis function (RBF) is used as the kernel and default values for all the parameters

that are included in scikit-learn; this is similar to VMAF, BRISQUE [MMB12] and NIQE [MSB13].

For optimizing the feature set, a full grid search covering all possible combinations of features is performed. Finally, the combination of features that results in the best performance in terms of PCC and RMSE was selected. This approach resulted in 255 possible feature combinations, where 6 features are based on VMAF's pixel-based calculations and two result from added metadata, namely resolution, and framerate. The hyper-parameter and feature set optimization is a joint optimization process where for every value of the *number of trees parameter*, all possible feature combinations were tested, and finally, the *(number of trees; feature set)* that results in the best performance is selected. 20 different values for the *number of trees* parameter ranging between 1 and 101 with a step size of 5 are considered.

Other parameters of RF and XGBoost can be optimized in a similar manner. Some of the parameters were checked with several pre-tests, and it was finally decided to use the default values, as the optimization of additional parameters did not show any significant improvement in the overall performance of the models. For each of the aforementioned variations of parameters, a separate model is trained and evaluated, considering the prediction performance of the validation set.

The training-validation ratio for all three machine learning algorithms was chosen to be 50:50, ensuring that none of the common sources are used in training so that the model variants are validated with completely unknown videos.

Table 4.17 shows the results using the best combination of *(number of trees; feature set)*. In addition, Table 4.18 summarizes the performance using a *(number of trees; feature set)* set with as few values as possible for both parameters, with comparable performance with the best case. For the case of SVR, even with just 4 features, the performance is comparable to the best case.

The best combination is 26 trees with 4 features for the RF case, and 101 trees with 4 features for the XGBoost-based model. Even with only 6 trees and 4 features, the RF model has comparable performance with the best RF case. Similarly, a model with 71 trees and 4 features shows comparable performance with the best case for the XGBoost-based model. All these best-performing cases included resolution and framerate as features. Furthermore, a detailed feature relevance analysis by counting

the number of occurrences of features is performed which is detailed in Figure 4.9. It can be observed that all hybrid-VMAF instances outperform the retrained VMAF significantly, in terms of all applied performance criteria, namely PCC, SROCC, Kendall rank correlation coefficient, and RMSE. This demonstrates the validity of the proposed approach and also the suitability of the used machine learning algorithms to develop such models. In addition to the performance metrics, a significance analysis was performed according to ITU-T P.1401 [ITU14a].

Besides VMAF, model performance is compared with a number of further metrics, as shown in Tables 4.17 and 4.18. It should be noted that for BRISQUE and NIQE, the performance numbers reported in Tables 4.17 and 4.18 result after retraining, as described in the AVT-VQDB-UHD-1 study [Rao+19a].

Table 4.17: Performance comparison between Hybrid-VMAF and other SoA video quality models (considering the best performing feature combination).

Metric	RMSE	PCC	SROCC	Kendall	#Tree	#Feature
VMAF [Net18]	0.592	0.807	0.811	0.624	NA	NA
BRISQUE [MMB12]	0.641	0.813	0.833	0.646	NA	NA
NIQE [MSB13]	1.006	0.393	0.387	0.265	NA	NA
PSNR	1.004	0.313	0.491	0.352	NA	NA
SSIM [Wan+04]	0.871	0.497	0.580	0.418	NA	NA
MS-SSIM [WSB03]	0.832	0.559	0.581	0.421	NA	NA
ADM2 [Li+11]	0.598	0.803	0.806	0.615	NA	NA
VIFP [SB06]	0.789	0.618	0.612	0.449	NA	NA
VMAF (50:50 retraining)	0.588	0.849	0.870	0.690	NA	6
Hybrid-VMAF (SVR)	0.397	0.939	0.929	0.774	NA	5
Hybrid-VMAF (RF)	0.434	0.921	0.918	0.756	26	4
Hybrid-VMAF (XGBoost)	0.433	0.924	0.927	0.772	101	4

In addition to the performance analysis in terms of correlation and RMSE, an analysis of the number of times of occurrence of all the features for the top 100 performing (*number of trees; feature set*) set is performed. It can be seen from Figure 4.9 that for all three cases (SVR, RF, XGBoost), resolution and framerate occurred the highest number of times. Also, in the best performing (*number of trees; feature set*) set in both Tables 4.17 and 4.18, resolution and framerate were part of the feature set for all three machine learning algorithms. This further shows that the additional metadata-based input features, resolution, and framerate, indeed play an important role in improving the performance of FR models. It further indicates that only a small

Table 4.18: Performance comparison between Hybrid-VMAF and other SoA video quality models (considering lowest number of trees and features with comparable performance as the best case).

Metric	RMSE	PCC	SROCC	Kendall	#Tree	#Feature
VMAF [Net18]	0.592	0.807	0.811	0.624	NA	NA
BRISQUE [MMB12]	0.641	0.813	0.833	0.646	NA	NA
NIQE [MSB13]	1.006	0.393	0.387	0.265	NA	NA
PSNR	1.004	0.313	0.491	0.352	NA	NA
SSIM [Wan+04]	0.871	0.497	0.580	0.418	NA	NA
MS-SSIM [WSB03]	0.832	0.559	0.581	0.421	NA	NA
ADM2 [Li+11]	0.598	0.803	0.806	0.615	NA	NA
VIFP [SB06]	0.789	0.618	0.612	0.449	NA	NA
VMAF (50:50 retraining)	0.588	0.849	0.870	0.690	NA	6
Hybrid-VMAF (SVR)	0.438	0.930	0.913	0.744	NA	4
Hybrid-VMAF (RF)	0.442	0.919	0.920	0.751	6	4
Hybrid-VMAF (XGBoost)	0.438	0.921	0.925	0.769	71	4

amount of additional data is required to develop a hybrid model variant with good overall performance. As mentioned before, the inclusion of bitrate and video codec as metadata features is also evaluated, but no improvement was observed; hence these metadata features were removed for the proposed hybrid model framework.

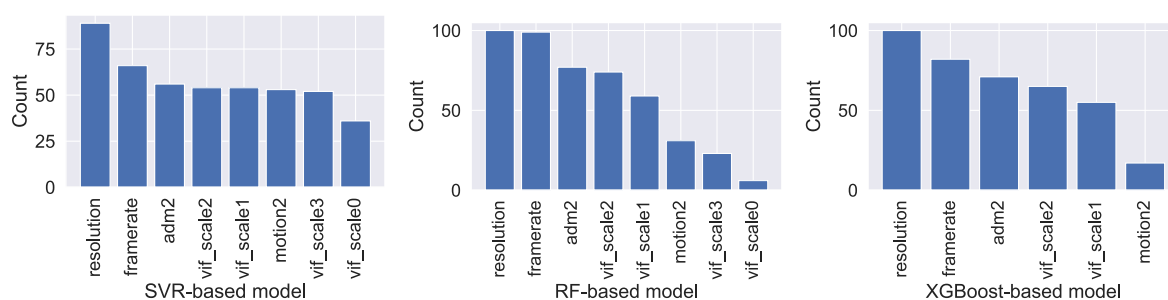


Figure 4.9: Frequency of occurrence of features in top 100 performing cases.

4.6 Summary

With the objective of developing quality models for both short-term video quality assessments for high-resolution videos and addressing research questions 1 and 2, different types of models are described in this chapter. In this regard, *AVQBits*, a versatile, bitstream-based video quality model that can be used in different scenarios

based on the available input information is presented. It consists of four different model types, namely, Mode 3, Mode 1, Mode 0, and Hybrid No-reference Mode 0.

The Mode 3 model developed as part of this work was submitted as a candidate model for the *P.NATS Phase 2* competition. Hence, firstly, a detailed overview of the *P.NATS Phase 2* competition including the winning model selection procedure was outlined. The work of the author as part of this thesis resulted in being part of three winning categories in the *P.NATS Phase 2* competition. Among them, the winning Mode 3 model is standardized as ITU-T Rec. P.1204.3 and forms the base for all the other models *AVQBits*.

Following this, the Mode 3 model was described in detail. A detailed evaluation of the models as part of the competition showed that the proposed Mode 3 model performed the best in comparison to all the models submitted to the competition. In addition to this, the submitted Mode 3 model is further compared with the SoA models using the *P.NATS Phase 2* validation database. The comparison showed that the Mode 3 model outperformed all considered SoA models.

Subsequently, four different extensions of the ITU-T Rec. P.1204.3 model were developed. These extensions consist of a Mode 0 model, a Mode 1 model, and two variants of a hybrid NR Mode 0 model. The AVT-PNATS-UHD-1 dataset was used to train these models. All five instances of *AVQBits* were evaluated using the AVT-VQDB-UHD-1 dataset. Furthermore, all five models were compared with the SoA FR and NR models using the AVT-VQDB-UHD-1 dataset. This comparison showed that ITU-T P.1204.3 and hybrid NR Mode 0 models significantly outperform the SoA FR and NR models with the Mode 0 and Mode 1 extensions performing significantly better than the considered NR models.

After this, two other models developed as part of the overall model development process were presented. The first one is an extension of the ITU-T P.1203.1 Mode 0 model to handle newer codecs and videos of resolution up to 4K/UHD-1 and framerates up to 60 fps. This extension was developed as a precursor to the P.1204.3-based Mode 0 model. This development also resulted in the creation of a new dataset which is described in Chapter 3 as the “AV1 Dataset”. The second model is a hybrid version of the VMAF model. With the objective of developing approaches to extend existing SoA pixel-based models using metadata to enhance prediction accuracy,

a generalized approach to developing hybrid models is presented. VMAF is used as a candidate FR model to validate the proposed approach. The hybrid extension of VMAF was developed using video resolution and framerate as additional input information. The AVT-VQDB-UHD-1 dataset was used to train and validate this extension. Results indicate that the proposed approach to developing hybrid models does increase prediction accuracy significantly.

This chapter primarily focuses on the development and evaluation of models for the prediction of the short-term video quality of traditional video content mainly streamed on platforms such as YouTube, Netflix, Amazon Prime Video, etc. However, the developed models may not be limited to this particular use case. Hence, all four instances of *AVQBits* are tested for their applicability to other use cases to address research questions 3 and 4. For this purpose, firstly, the four models instances of *AVQBits* are tested for the prediction of the overall quality of a HAS session using the model instances as the video quality component in a long-term integration model in Chapter 5. Following this, all four model instances are evaluated for other application scopes such as gaming, 360°, HFR videos, and images in Chapter 6.

Overall Integral Quality

Usually, there is a tendency to treat QoE as a static event, and the QoE measured for a stimulus of delimited length is assumed to be stable along its duration. However, this rarely happens for stimuli extending over several minutes [Wei+14]. This can be well observed in a typical HAS session lasting several minutes, which may include different quality-related events, for example, quality switching, initial loading delay, and stalling. Hence, any model designed to estimate the overall integral quality of a HAS session has to take into account the impact of these events.

ITU-T Rec. P.1203.3 [ITU20] is the first standardized model that takes into account all these factors to predict the QoE of a HAS session [Rob+18a]. This model takes per-1-second video and audio quality scores, stalling-related information, and the device type (either “PC/TV” or “Mobile/Tablet”) as input to calculate the QoE of a HAS viewing session. The main model output (referred to as $O.46$ in [ITU20]) is a final media session quality score on the 5-point “MOS-scale”. Further, besides the parametric input information, the model produces intermediate values that can be used for HAS-system diagnostics, such as a perceptual stalling indication, audiovisual segment coding quality per output sampling interval, and a final audiovisual coding quality score. The design of the subjective tests conducted to gather ground truth for model training and validation used a retrospective rating by the participants on a 5-point ACR scale given at the end of an audiovisual stimulus lasting between 1 – 5 min. Due to this test design, it becomes pertinent to address cognitive effects such as the *recency effect* and *primacy effect* (see, e.g., [Wei+14] for more details and references around these effects). Accordingly, ITU-T Rec. P.1203.3 considers these cognitive effects as part of the model.

An adaptation of the ITU-T Rec. P.1203.3 which takes into account the more accurate short-term video quality prediction obtained from the different instances of *AVQBits* is proposed in this chapter. This proposed extension has been extensively evaluated using the PNATS-UHD-1-Long dataset described in Chapter 3.

This chapter is based on the following publication:

[RGR22] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications”. In: *IEEE Access* 10 (2022)

5.1 Model Description

In this section, a long-term integration model specifically designed for the four types of *AVQBits* models is presented, which is based on ITU-T Rec. P.1203.3 [ITU20]. It relies on the same model structure as P.1203.3, adapting the final audiovisual coding quality estimation, using more accurate short-term video quality models such as ITU-T P.1204.3. The final audiovisual coding quality score $O.35$ in ITU-T Rec P.1203.3 is estimated following Equation (5.1).

$$\begin{aligned}
 O.35 &= O.35_{baseline} - negBias - oscComp \\
 &\quad - adaptComp \\
 O.35_{baseline} &= \frac{\sum_t w_1(t) \cdot w_2(t) \cdot O.34[t]}{\sum_t w_1(t) \cdot w_2(t)} \\
 w_1(t) &= t_1 + t_2 \cdot \exp\left(\frac{t-1}{t_3}\right) \\
 w_2(t) &= t_4 - t_5 \cdot O.34[t]
 \end{aligned} \tag{5.1}$$

Here, $O.34$ is the audiovisual segment coding quality per output sampling interval. The values w_i are weighting coefficients specified in the standard [ITU20; Rob+18a]. The three factors *negBias*, *oscComp* and *adaptComp* are used to take into account certain temporal effects related to video-quality fluctuations. In the proposed model, these three factors are ignored, reflecting two assumptions:

5.2 Evaluation of the Overall Integral Quality Model

1. The per-second and per-segment scores of the *AVQBits* model instances are generally more accurate than those from the short-term video-quality module variants of ITU-T Rec. P.1203.1 [ITU19e; Raa+17], where these were re-engineered from the final, retrospective and longer-session media session quality score (*O.46*) (see, e.g., [Raa+17] for more details).
2. This re-engineering may have been impacted by specific ITU-T P.1203.1 implementations, and thus be very specific for data created as part of the ITU-T P.1203 development process, and not optimally suited for the *AVQBits* variants proposed in this thesis.

The *AVQBits* model instances are specifically trained on short-term videos and hence are capable of more accurately estimating both per-segment and per-1-second video quality scores. As a result, the new, simplified *O.35* is given by Eq (5.2).

$$O.35 = O.35_{baseline} \quad (5.2)$$

It should be noted that no other changes to the model algorithm or coefficients inherited from ITU-T P.1203.3 have been applied.

5.2 Evaluation of the Overall Integral Quality Model

The proposed long-term integration model, a simplified version of ITU-T Rec. P.1203.3 [ITU20] is evaluated on PNATS-UHD-1-Long consisting of five tests with PVSs ranging from 1-5 min in duration (cf. Ch.. 3). As explained in Section 5.1, for estimating the overall integral quality, the proposed model follows the same architecture as ITU-T Rec. P.1203.3 and takes per-1-second video and audio scores as input, along with stalling-related information. In this evaluation, to estimate the per-1-second video quality scores, the different *AVQBits* instances are considered. The per-1-second audio quality scores are assumed to be 4.5, which is the highest quality estimated by ITU-T Rec. P.1203.2 [ITU17]. This assumption is based on the fact that the audio quality is not varied and the best possible audio quality is used in all the five tests considered for evaluation.

Table 5.1: Performance of *AVQBits* instances on the PNATS-UHD-1-Long dataset.

Database	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
test_1	<i>AVQBits</i> M3 / P.1204.3	0.353	0.892	0.823	0.649	0.795
test_1	<i>AVQBits</i> M1	0.432	0.831	0.822	0.655	0.691
test_1	<i>AVQBits</i> M0	0.399	0.859	0.863	0.694	0.738
test_1	<i>AVQBits</i> H0 s	0.357	0.888	0.870	0.706	0.789
test_1	<i>AVQBits</i> H0 f	0.352	0.891	0.868	0.705	0.798
test_2	<i>AVQBits</i> M3 / P.1204.3	0.559	0.813	0.801	0.610	0.661
test_2	<i>AVQBits</i> M1	0.624	0.760	0.744	0.562	0.577
test_2	<i>AVQBits</i> M0	0.658	0.728	0.707	0.514	0.529
test_2	<i>AVQBits</i> H0 s	0.640	0.745	0.689	0.511	0.556
test_2	<i>AVQBits</i> H0 f	0.650	0.736	0.679	0.498	0.541
test_3	<i>AVQBits</i> M3 / P.1204.3	0.485	0.888	0.798	0.630	0.788
test_3	<i>AVQBits</i> M1	0.552	0.852	0.809	0.644	0.725
test_3	<i>AVQBits</i> M0	0.641	0.794	0.749	0.570	0.630
test_3	<i>AVQBits</i> H0 s	0.514	0.873	0.804	0.635	0.762
test_3	<i>AVQBits</i> H0 f	0.540	0.858	0.818	0.658	0.737
test_4	<i>AVQBits</i> M3 / P.1204.3	0.377	0.917	0.899	0.748	0.842
test_4	<i>AVQBits</i> M1	0.517	0.838	0.798	0.627	0.702
test_4	<i>AVQBits</i> M0	0.534	0.826	0.782	0.609	0.683
test_4	<i>AVQBits</i> H0 s	0.392	0.910	0.878	0.729	0.829
test_4	<i>AVQBits</i> H0 f	0.370	0.920	0.878	0.715	0.847
test_5	<i>AVQBits</i> M3 / P.1204.3	0.386	0.934	0.922	0.796	0.872
test_5	<i>AVQBits</i> M1	0.700	0.762	0.796	0.641	0.581
test_5	<i>AVQBits</i> M0	0.832	0.638	0.594	0.464	0.407
test_5	<i>AVQBits</i> H0 s	0.502	0.855	0.842	0.684	0.732
test_5	<i>AVQBits</i> H0 f	0.500	0.857	0.825	0.658	0.734
All	<i>AVQBits</i> M3 / P.1204.3	0.479	0.864	0.844	0.660	0.747
All	<i>AVQBits</i> M1	0.596	0.780	0.787	0.602	0.608
All	<i>AVQBits</i> M0	0.694	0.686	0.683	0.500	0.471
All	<i>AVQBits</i> H0 s	0.570	0.797	0.768	0.584	0.635
All	<i>AVQBits</i> H0 f	0.582	0.787	0.756	0.572	0.619

Table 5.1 shows the performance numbers for all the tests for the proposed models. It can be concluded that using *AVQBits*|M3 / P.1204.3 to estimate the per-1-second scores results in a very good performance of the proposed long-term integration model. This is due to the high accuracy of the ITU-T P.1204.3 model. The estimation of per-1-second and per-segment quality scores is better as compared to the other instances of *AVQBits*, which use less complex input information without full bitstream access for video quality prediction. Furthermore, it can be observed that the *AVQBits*|H0|s and *AVQBits*|H0|f variants show similar performance to the *AVQBits*|M3 / P.1204.3 in terms of PCC but have worse performance in terms of RMSE for each of the five tests. This is unlike the short-term video quality prediction where the *AVQBits*|H0|s and *AVQBits*|H0|f variants have a similar performance to *AVQBits*|M3 / P.1204.3 both in terms of PCC and RMSE. This can be attributed to

5.2 Evaluation of the Overall Integral Quality Model

the fact that in the case of short-term video quality prediction a simple linear mapping according to ITU-T P.1401 before computing the RMSE would accommodate for the difference in prediction due to the usage of the QEB instead of the original bitstream. Whereas in the case of overall integral quality prediction, the input consists of per-1-sec scores, and these per-1-sec scores are computed on the QEB which may not reflect the true quality of the original bitstream. A dedicated linear mapping of the per-1-sec scores to take into account the effect of QEB at the per-1-sec level could alleviate such a problem and hence result in a lower RMSE value. Despite this, the overall performance of the $AVQBits|H0|s$ and $AVQBits|H0|f$ variants are significantly better than the $AVQBits|M0$ and $AVQBits|M1$ models.

The scatter plots illustrated in Figure 5.1 show that both $AVQBits|M0$ and $AVQBits|M1$ seems to over-predict in the lower-quality range, which can be attributed to the less accurate per-1-second score estimation by these models. This is assumed to reflect that the quality impact due to more encoder-demanding video content is less well captured by these models.

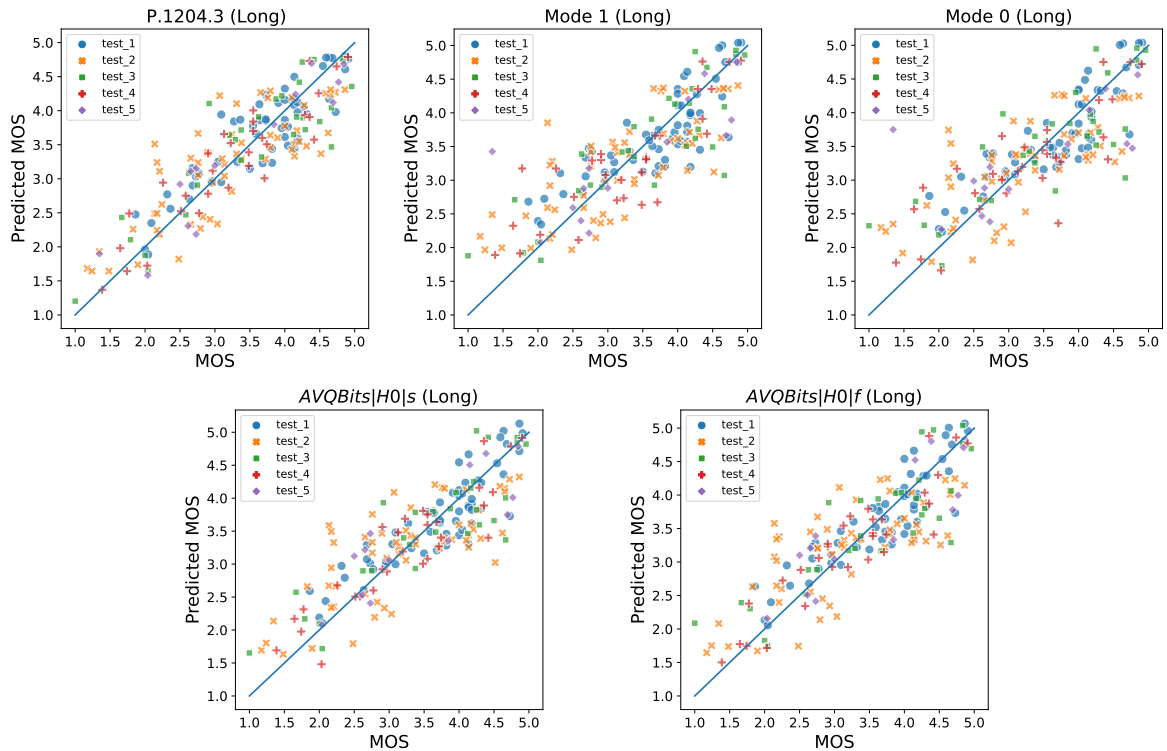


Figure 5.1: Scatter plot of $AVQBits$ instances for PNATS-UHD-1-Long dataset.

5.3 Summary

To address the need for the short-term video quality models proposed as *AVQBits* in Chapter 4 to be more ecologically valid, the scope of their usage has been extended to include typical viewing sessions in real-world in this chapter. For this purpose, a long-term integration model based on ITU-T Rec. P.1203.3 has been proposed. This model takes into account the effects of the typical degradations encountered in a HAS session such as audio and video quality switches, initial loading delay, and stalling events. As the video quality prediction module in this long-term integration model, all five model instances of *AVQBits* are considered. The analysis of the results shows that the Mode 3 and HYN0 variants perform better than the Mode 0 and Mode 1 variants owing to their more accurate per-1-second quality predictions. From this usage of the *AVQBits* variants, it can be seen that they can be applied to real-world HAS viewing sessions with an appropriate long-term integration model, thus addressing the objectives outlined in research question 4.

Following the successful demonstration of the usage of the *AVQBits* variants for assessment of the overall quality of a HAS session, an investigation of their applicability to other use cases such as gaming video, 360° video, HFR content, and images will be conducted in the next chapter.

Extended Application Scopes of *AVQBits*

As described in Chapter 1, there has been an increase in not only traditional 2D video streaming but also in other application areas such as gaming video streaming and 360° videos. This can be attributed to multiple factors. Some of those factors include an increase in affordable capture equipment and dedicated display devices such as head-mounted displays (HMDs) in case of 360° videos, and better and affordable equipment to stream video game play. In addition to this, significant advancements in transmission technologies focus on reducing the overall data, e.g. tiled-based streaming in the case of 360° videos, and also dedicated encoding strategies have made this expansion into newer application scopes feasible. Even in traditional 2D videos, apart from VoD services, there has been a considerable increase in the amount of live-streamed content and also HFR-related content mainly in live gaming streaming scenarios. Furthermore, all these use cases have been increasingly using HAS for delivering the content to the end-user. Hence, there is a need for automated assessment of the overall quality of such services and therefore the need for video quality models that can be used for these different use cases. There can either be dedicated models focused on a particular application scenario or one model that can be satisfactorily used across these different use cases with minimal adjustments. Of all the SoA models, VMAF has been shown to perform satisfactorily across these different scenarios of gaming [Bar+18], 360° video [Fre+20], etc. This chapter is focused on investigating the applicability of the *AVQBits* instances for different use cases such as gaming video, 360° video, HFR content, live-streamed content, user-generated content (UGC), and images.

This chapter is partially based on the following publications.

[Rao+20b] **Rakesh Rao Ramachandra Rao**, Steve Göring, Robert Steger, Saman Zadtootaghaj, Nabajeet Barman, Stephan Fremerey, Sebastian Möller, and Alexander Raake. “A Large-scale Evaluation of the bitstream-based video-quality model ITU-T P.1204.3 on Gaming Content”. In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020

[RGR22] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications”. In: *IEEE Access* 10 (2022)

6.1 Gaming

As a first extended application scope to evaluate the *AVQBits* instances, gaming video quality evaluation is considered. In the following, a brief survey of gaming video quality evaluation SoA, the details of the datasets considered for this purpose, and the performance evaluation are described.

6.1.1 Related Work for Gaming Video Quality Assessment

Besides traditional 2D video, there has been a significant increase in gaming video streaming. As a result, several video quality models dedicated to gaming video quality evaluation have been proposed in the literature. The focus has mainly been on the development of no-reference video quality models, due to the lack of high-quality reference videos in a gaming video streaming session. NR-GVQM is an example of machine-learning-based NR-models specifically developed for video quality estimation of gaming videos [Zad+18]. The model is based on support vector regression (SVR) and includes nine frame-level features indicating impairments such as blockiness, naturalness, etc. VMAF was used as the ground truth for model training, and hence this model can be viewed as a no-reference counterpart to VMAF. The model was trained and validated with the GamingVideoSet [Bar+18]. 408 out of the 576 processed video sequences (PVS) were used for training and the remaining 144 PVSs for validation.

Another NR-based gaming video quality model was proposed by Göring, Rao, and Raake [GRR19] referred to as “nofu”. It is based on a number of features that are integrated using a random forest model. There are two instances of the nofu model, with the first developed to predict VMAF scores and the second to predict subjective MOS. Both model instances were trained on the GamingVideoSet [Bar+18] dataset and shown to perform well based on 10-fold cross-validation.

Using a similar approach, Barman et al. [Bar+19] develop two NR model instances, namely, “NR-GVQSI” and “NR-GVQSE”, with NR-GVQSI using subjective MOS as training target and NR-GVQSE using VMAF as the training ground truth. The models were trained and tested using two different datasets: KUGVD [Bar+19] and GVS [Bar+18]. NR-GVQSI is based on neural networks and extracts seven features for quality prediction and was trained on one dataset and tested on the other and vice versa. NR-GVQSE is based on SVR and also uses seven (different from NR-GVQSI) parameters. The model was trained on GVS using a 10-fold cross-validation strategy. Additionally, it was tested on KUGVD for its performance using MOS data.

In addition to other machine-learning-based models, deep learning approaches have been explored to develop gaming video quality models. One example of such a model is the NDNetGaming model proposed by Utke et al. [Utk+20], which shows a good performance in terms of PCC on the KUGVD [Bar+19] dataset. A further extension of the NDNetGaming model called “DEMI” has been presented by Zadtootaghaj et al. [Zad+20a]. “DEMI” incorporates a more sophisticated pooling of the per-frame quality scores to obtain the per-segment quality score, in addition to other improvements. This model is developed to be applicable to non-gaming videos, too. The performance of the model was evaluated on the CGVDS [Zad+20b] dataset, showing the model to be on par with or better than SoA models.

Although a bigger focus has been on pixel-based NR models for gaming video quality assessment, some studies have investigated bitstream-based models for gaming video quality prediction. One example of a gaming-specific bitstream-based model is the BQGV proposed by [Zad+20b], which is described in Section 2.2.2.1. A 5-fold cross-validation approach using the CGVDS [Zad+20b] dataset was performed for performance evaluation, and it has been reported to outperform the ITU-T Rec. P.1203.1 Mode 1 and Mode 3 models. However, it should be noted that the ITU-T

Rec. P.1203.1 Mode 1 and Mode 3 models were not retrained for gaming videos in that study.

Moreover, a gaming-specific planning model called GamingPara has been presented by Zadtootaghaj et al. [Zad+20b]. It is based on a multidimensional approach, where overall video quality is a combination of impairments related to video discontinuity, video fragmentation, and video unclearness. The model is shown to outperform ITU-T P.1203.1 Mode 0 on gaming data.

In addition, ITU-T Rec. G.1072 comprises a video quality component that can be used to evaluate gaming video quality. It is based on retraining the video quality component of the IPTV-related planning model described in ITU-T G.1071 [ITU16c].

6.1.2 Datasets

Four different datasets, namely the three publicly available datasets GamingVideoSet [Bar+18], KUGVD [Bar+19], CGVDS [Zad+20b], and a self-developed proprietary Twitch dataset [Rao+20b] are considered for the evaluation of the AVQBits model instances. In the following, the datasets are described in more detail.

6.1.2.1 GamingVideoSet (GVS)

GVS [Bar+18] consists of 24 SRCs that have been extracted from 12 different games. The SRCs are of 1920×1080 pixels resolution, 30 *fps* framerate and have a duration of 30 s. The HRCs included 24 different bitrates across three different resolutions, namely, 480*p*, 720*p* and 1080*p*. H.264 was selected to encode the videos with the defined bitrate-resolution pairs, resulting in a total of 576 PVSs. A Constant Bitrate (CBR) encoding approach with *veryfast* preset was selected to encode the videos. Out of the 576 PVSs, a reduced sample of 90 PVSs with six SRCs and 15 bitrate-resolution pairs was chosen for subjective evaluation. A total of 25 participants rated all the 90 PVSs.

6.1.2.2 Kingston University Gaming Video Dataset (KUGVD)

Six SRCs out of the 24 SRCs from the GamingVideoSet were used to develop KUGVD [Bar+19]. The same bitrate-resolution pairs from GamingVideoSet were included to define the HRCs. Following the encoding approach in GamingVideoSet, 144 PVSs were created and 90 PVSs out of these were selected for the subjective evaluation, with 17 participants taking part in the test. This dataset was created mainly for the development of the NR-GVSQI and NR-GVSQE models.

6.1.2.3 Cloud Gaming Video Dataset (CGVDS)

Compared to the aforementioned datasets, CGVDS [Zad+20b] consists of a larger number of games, i.e. 15, and also includes videos captured at 60fps. Similar to the previously discussed two datasets, three different resolutions, namely, 480p, 720p, and 1080p are considered at three different framerates of 20, 30, and 60fps. A total of 17 bitrate conditions spread across all the resolutions are used in the design of this dataset. Unlike the GamingVideoSet and KUGVD datasets, this dataset uses a hardware-accelerated implementation of H.264/MPEG-AVC (NVENC) because most cloud providers use this for delay-sensitive cloud gaming services. A CBR mode of encoding with the preset of **llhq** (low latency, high quality) was used to encode the videos. 5 different subjective tests were conducted to make sure all 15 games were addressed, using 3 video sequences as anchor conditions. Each subjective test had a total of 72 PVSs using a display with FHD resolution. Over 100 subjects participated across all tests with a minimum of 20 subjects for each test.

6.1.2.4 Twitch Dataset

The last considered dataset, referred to as Twitch Dataset [Rao+20b], was created with the initial aim of using it for genre classification, hence, a due effort was spent to make sure that the dataset comprises gaming videos of different genres. This dataset consists of a total of 36 different games, with each genre being represented by 6 games. The genres were chosen based on their relevance and popularity on Twitch. Three different streamers were recorded three times per game to maintain high diversity for each game. A total of 351 video sequences with a duration of

approximately 50 s spanning all representation levels were downloaded from Twitch. This was done to ensure the usage of real-world encodings in the subjective test. A subset of 90 sequences out of the 351 sequences was used in the test. Only the first 30 s of each video in the chosen subset were shown to the test subjects to maintain a fixed duration of one hour for the test. All 36 games from the original dataset are represented in the test with either two or three streamers. Resolutions of 160p, 360p, 480p, 720p, 900p and 1080p and framerates of 30 and 60 fps were used. The encoding scheme was the one used in Twitch.tv since the encoded representations were directly downloaded from Twitch. A total of 29 subjects participated in the test. One outlier was detected using a criterion of 0.75 PCC and was removed from further analysis.

The datasets are summarized in Table 6.1 and the MOS distribution of all the four datasets is as shown in Figure 6.1.

Table 6.1: Overview of the used gaming datasets.

Parameter	GVS	KUGVD	CGVDS	Twitch
No. of sources	6	6	15	36
No. of PVS's	90	90	72×5	90
No. of subjects	25	17	>100 (5 tests)	29
Resolution	480p, 720p, 1080p	480p, 720p, 1080p	480p, 720p, 1080p	160p, 360p, 480p, 720p, 900p, 1080p
Framerate (fps)	30	30	20, 30, 60	30, 60
Duration (s)	30	30	30	30
Encoder	ffmpeg x264	ffmpeg x264	ffmpeg NVENC (H.264)	H.264
Encoding mode	CBR	CBR	CBR	Twitch default
Preset	veryfast	veryfast	llhq	Twitch default

6.1.3 Evaluation

It should be noted that all model instances of *AVQBits* are used without any re-training to estimate the video quality for the four gaming datasets considered for performance evaluation. The only difference to the case of “normal” 2D video is that here all databases were created for a Full-HD (1920×1080 pixels) display instead of the 4K/UHD-1 target screen resolution used in the case of the PC databases for

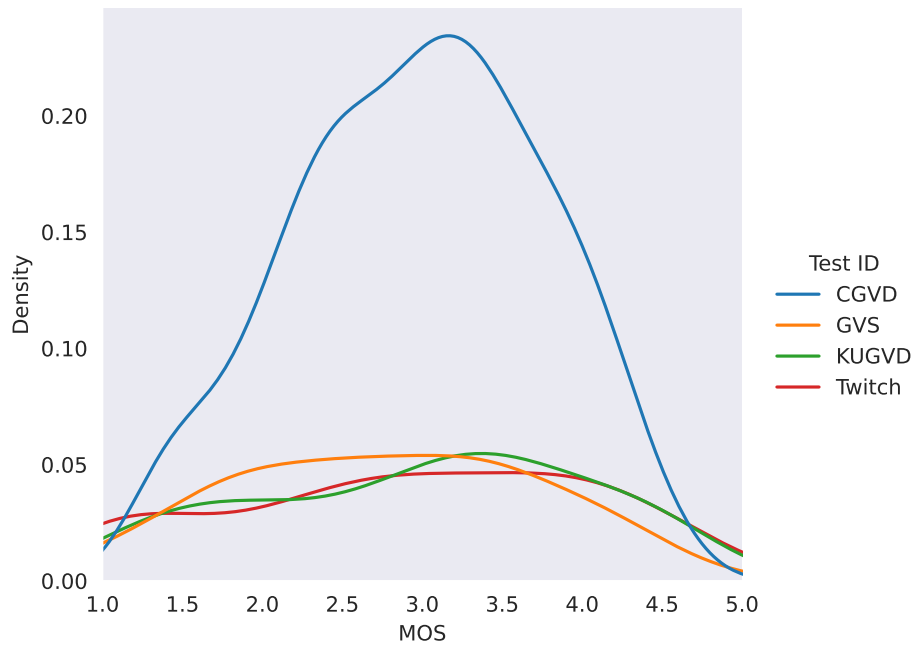


Figure 6.1: MOS distribution of GVS, KUGVD, CGVDS, and Twitch datasets.

“normal” video and the initial development of $AVQBits|M3$ / P.1204.3 in ITU-T SG12. In this evaluation, $AVQBits|M3$ / P.1204.3 and its extensions are used directly with the target resolution of 4K/UHD-1.

Table 6.2 provides a detailed view of the performance of the proposed models on all four considered tests. $AVQBits|M3$ / P.1204.3 and $AVQBits|M1$ perform on par across all datasets, with $AVQBits|M0$ being the least well performing model. The good performance of the $AVQBits|M1$ model indicates that the features related to frame size and frame type can be used to estimate the impact of content improving the estimation of the QP value and bringing it closer to the one of the $AVQBits|M3$ model with its full bitstream access. Although $AVQBits|M1$ performs on par with $AVQBits|M3$ / ITU-T P.1204.3 on average, from the scatter plots shown in Figure 6.2 it can be observed that there is a general tendency of the $AVQBits|M1$ model to slightly over-predict as compared to ITU-T P.1204.3. Furthermore, it can be observed from the scatter plot associated with $AVQBits|M0$ in Figure 6.2 that $AVQBits|M0$ suffers significantly from the lack of content-related features, leading to cases with a larger prediction inaccuracy. In case Mode 0 type data and pixel information can be accessed in a practical monitoring scenario, the $AVQBits|H0$ models are highly

Table 6.2: Performance of AVQBits instances using the considered gaming datasets.

Dataset	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
GVS	AVQBits M3 / P.1204.3	0.45	0.88	0.87	0.69	0.77
GVS	AVQBits M1	0.42	0.89	0.87	0.71	0.79
GVS	AVQBits M0	0.69	0.67	0.65	0.49	0.45
GVS	AVQBits H0 s	0.48	0.86	0.86	0.69	0.74
GVS	AVQBits H0 f	0.62	0.75	0.73	0.56	0.56
KUGVD	AVQBits M3 / P.1204.3	0.39	0.93	0.92	0.77	0.86
KUGVD	AVQBits M1	0.50	0.87	0.86	0.69	0.76
KUGVD	AVQBits M0	0.84	0.59	0.57	0.41	0.35
KUGVD	AVQBits H0 s	0.46	0.90	0.89	0.72	0.80
KUGVD	AVQBits H0 f	0.65	0.78	0.76	0.58	0.61
CGVDS	AVQBits M3 / P.1204.3	0.38	0.85	0.84	0.65	0.72
CGVDS	AVQBits M1	0.36	0.90	0.88	0.70	0.78
CGVDS	AVQBits M0	0.47	0.78	0.75	0.56	0.60
CGVDS	AVQBits H0 s	0.36	0.89	0.88	0.70	0.79
CGVDS	AVQBits H0 f	0.38	0.87	0.87	0.68	0.76
Twitch	AVQBits M3 / P.1204.3	0.40	0.93	0.93	0.77	0.87
Twitch	AVQBits M1	0.37	0.94	0.93	0.77	0.89
Twitch	AVQBits M0	0.43	0.92	0.89	0.71	0.85
Twitch	AVQBits H0 s	0.31	0.96	0.95	0.82	0.92
Twitch	AVQBits H0 f	0.30	0.96	0.95	0.81	0.92
All	AVQBits M3 / P.1204.3	0.41	0.90	0.90	0.73	0.81
All	AVQBits M1	0.41	0.90	0.89	0.73	0.81
All	AVQBits M0	0.60	0.76	0.75	0.56	0.58
All	AVQBits H0 s	0.40	0.90	0.90	0.73	0.82
All	AVQBits H0 f	0.48	0.86	0.85	0.67	0.73

usable. The results show that *AVQBits|H0|s* performs as well as *AVQBits|M3 / P.1204.3* for all the four considered gaming datasets. The *AVQBits|H0|f* model variant with fewer requirements on the set of codecs available during monitoring performs on par with *AVQBits|M3 / P.1204.3* for the CGVDS and Twitch datasets, but less well for GVS and KUGVD. This may be due to the coefficients a_{map} and b_{map} being obtained by training on traditional 2D video datasets. Dedicated retraining of these two coefficients for gaming content may result in improved performance.

In addition to this, the performance of the proposed AVQBits model instances are compared with SoA models and the details are reported in Table 6.3. The performance numbers corresponding to the SoA models relating to the open datasets are directly taken from respective papers. In general, it can be concluded that

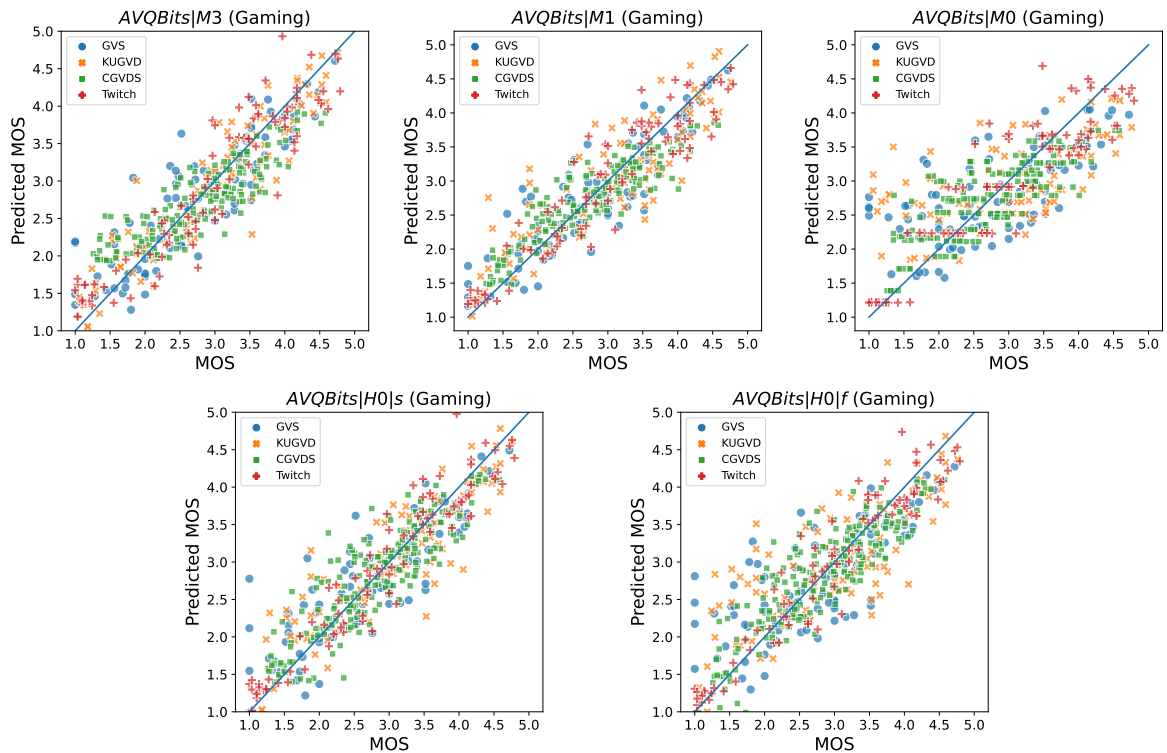


Figure 6.2: Scatter plot of *AVQBits* instances for the considered gaming datasets.

AVQBits|M3 / P.1204.3 and *AVQBits|M1* perform on par with VMAF across all datasets. It should be noted that two out of the four datasets, namely, CGVDS and the Twitch dataset use completely different encoding strategies than the ones these models were trained on. CGVDS uses a hardware-accelerated encoder and the Twitch dataset consists of PVSs with proprietary Twitch encoding. Despite this, the models perform well indicating the generalizability of the model w.r.t different encoder implementations and strategies. Although *AVQBits|M0* is the worst performing bitstream model, it still outperforms all the considered NR models for all datasets. The performance of *AVQBits|M0* can be enhanced by retraining it for the gaming-specific use case. Furthermore, *AVQBits|H0|s* performs as well as both the best-performing pixel and bitstream models. Although *AVQBits|H0|f* suffers from a lower performance for the GVS and KUGVD datasets, the average performance across all four datasets is still competitive in comparison with the SoA models.

Table 6.3: Comparison of performance of AVQBits instances with SoA models using the considered gaming datasets.

Dataset	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
GVS	PSNR	0.63	0.74	0.74	0.57	0.55
GVS	SSIM	0.57	0.80	0.80	0.61	0.62
GVS	VMAF	0.47	0.87	0.86	0.69	0.75
GVS	NIQE	0.64	0.77	0.71	0.53	0.52
GVS	AVQBits M3 / P.1204.3	0.45	0.88	0.87	0.69	0.77
GVS	AVQBits M1	0.42	0.89	0.87	0.71	0.79
GVS	AVQBits M0	0.69	0.67	0.65	0.49	0.45
GVS	AVQBits H0 s	0.48	0.86	0.86	0.69	0.74
GVS	AVQBits H0 f	0.62	0.75	0.73	0.56	0.56
KUGVD	PSNR	0.62	0.80	0.84	0.67	0.64
KUGVD	SSIM	0.48	0.89	0.91	0.74	0.79
KUGVD	VMAF	0.41	0.92	0.92	0.77	0.85
KUGVD	NIQE	0.55	0.85	0.84	0.66	0.72
KUGVD	AVQBits M3 / P.1204.3	0.39	0.93	0.92	0.77	0.86
KUGVD	AVQBits M1	0.50	0.87	0.86	0.69	0.76
KUGVD	AVQBits M0	0.84	0.59	0.57	0.41	0.35
KUGVD	AVQBits H0 s	0.46	0.90	0.89	0.72	0.80
KUGVD	AVQBits H0 f	0.65	0.78	0.76	0.58	0.61
CGVDS	PSNR	0.60	0.64	0.65	0.47	0.41
CGVDS	SSIM	0.59	0.67	0.78	0.60	0.45
CGVDS	VMAF	0.38	0.88	0.87	0.69	0.77
CGVDS	NIQE	0.66	0.54	0.56	0.41	0.29
CGVDS	AVQBits M3 / P.1204.3	0.38	0.85	0.84	0.65	0.72
CGVDS	AVQBits M1	0.36	0.90	0.88	0.70	0.78
CGVDS	AVQBits M0	0.47	0.78	0.75	0.56	0.60
CGVDS	AVQBits H0 s	0.36	0.89	0.88	0.70	0.79
CGVDS	AVQBits H0 f	0.38	0.87	0.87	0.68	0.76
Twitch	NIQE	0.96	0.24	0.11	0.17	0.04
Twitch	AVQBits M3 / P.1204.3	0.40	0.93	0.93	0.77	0.87
Twitch	AVQBits M1	0.37	0.94	0.93	0.77	0.89
Twitch	AVQBits M0	0.43	0.92	0.89	0.71	0.85
Twitch	AVQBits H0 s	0.31	0.96	0.95	0.82	0.92
Twitch	AVQBits H0 f	0.30	0.96	0.95	0.81	0.92

In the next section, an FHD-mapped version of AVQBits|M3 / P.1204.3 is proposed and evaluated.

6.1.3.1 FHD-mapped P.1204.3 model

The standardized P.1204.3 model was trained and validated on two different target devices, namely, a TV/PC monitor with 3840×2160 and a mobile/tablet (MO/TA) with 2560×1440 as the two target resolutions. Hence, the corresponding *scale_factor* that is used in P.1204.3 to determine the “upscaling degradation” is specified differently for PC/TV and MO/TA, as given in Equations (6.1) and (6.2), respectively [Rao+20a]:

$$scale_factor = \frac{coding_resolution}{3840 \cdot 2160} \quad for \text{ PC/TV} \quad (6.1)$$

$$scale_factor = \frac{coding_resolution}{2560 \cdot 1440} \quad for \text{ MO/TA} \quad (6.2)$$

All described gaming datasets use PC/TV as the target device, and hence only the PC/TV case was used for the FHD-mapped P.1204.3 version. As can be seen from Equation (6.1), the normalization of the *coding_resolution* is done w.r.t the display resolution of 3840×2160 . This is expected to lead to over-predicting the upscaling degradation when a lower resolution video is considered, that in the actual test was presented on an FHD screen rather than a 4K/UHD-1 screen. For example: If a Full-HD video is considered, the upscaling degradation should be 0 since the coding resolution of the video matches the display resolution used in the tests. But, if we use the original *scale_factor* definition, this would result in a finite non-zero upscaling degradation which is not the case. Similarly, the relative perception of other lower resolutions changes with the target display resolution.

To develop the FHD-mapped version of P.1204.3, the focus is on developing a dedicated adaptation of P.1204.3 targeting FHD resolution. Here, a correction factor to account for the overly strong handling of the upscaling degradation part by the original model when applying it to FHD resolution is proposed. This correction factor is referred to as $D_{u_corr_fac}$ and is defined in Equation (6.3).

$$D_{u_corr_fac} = a * \log \left(b * \left(\frac{coding_resolution}{1920 * 1080} \right) \right) \quad (6.3)$$

where $coding_resolution = coding_height * coding_width$ and \log is the natural logarithm.

Hence, the final prediction of the P.1204.3 model is adjusted using $D_{u_corr_fac}$ as defined in Equation (6.3) to obtain a final FHD-mapped prediction. This is represented in Equation (6.4).

$$pred_{hd_mapped} = pred_{p1204_3} + D_{u_corr_fac} \quad (6.4)$$

where $pred_{p1204_3}$ is the output of the standardized P.1204.3 model. The additive term preserves the overall architecture of the P.1204.3 model, considering the overly strong handling of the upscaling effect when applying P.1204.3 to FHD. With this approach, the original P.1204.3 model could be kept unchanged.

For training the correction factor $D_{u_corr_fac}$, the four datasets are split into a training and a validation set. GamingVideoSet and KUGVD are considered as the training datasets, which have a total of 24 encoding conditions (i.e. bitrate and resolutions). These two datasets consider 12 different sources in total which are encoded at 3 different resolutions (480p, 720p, and 1080p) to result in a combined total of 180 PVSs (90 + 90). The remaining two datasets, namely, the CGVDS and Twitch datasets were used as validation datasets.

The final coefficient values (cf. Equation (6.3)) after the training procedure are: $a = -0.10756695$ and $b = 0.08303269$.

The performance of the FHD-mapped P.1204.3 for the validation databases is reported in Table 6.4.

Table 6.4: Performance of FHD-mapped P.1204.3 on the validation datasets.

Dataset	RMSE	PCC	SROCC	Kendall	R^2 Score
CGVDS	0.40	0.84	0.83	0.62	0.70
Twitch	0.45	0.91	0.92	0.75	0.83

6.2 360° Video

As the next application scope, 360° video quality evaluation is considered to investigate the applicability of AVQBits on different video formats. For this evaluation 360 Streaming Video Quality Dataset [Fre+20] is considered. This section provides a

brief overview of the SoA for 360° video quality evaluation and describes the dataset considered for this evaluation in detail and also presents the performance results.

6.2.1 State of the Art for 360° Video Quality Assessment

Like for gaming video quality assessment, pixel-based models have been the main focus of the quality assessment of 360° videos. For example, variants of PSNR to take into account the possibility of viewing 360° in all directions have been proposed. S-PSNR [YLG15], a sphere-based PSNR computation, and WS-PSNR [SLY17], a position-weighted PSNR have been proposed as quality metrics to ultimately increase compression efficiency while maintaining a similar quality.

Tran et al. [Tra+17] conduct a performance evaluation of 360° video quality metrics considering different variants of PSNR including S-PSNR and WS-PSNR, among others. They concluded that the traditional approach of calculating PSNR was the most appropriate for 360° video.

More perception-oriented, traditional 2D video quality models such as VMAF have also been evaluated for quality assessment of 360° video. For example, Fremerey et al. [Fre+20] evaluated the applicability of both the original version of VMAF and the centre-cropped version of VMAF [GKR19] for 360° video quality evaluation and reported good performance in terms of PCC. Also, Orduna et al. [Ord+20] report similar results for VMAF as reported by Fremerey et al. [Fre+20] for 360° video quality evaluation. Furthermore, extensions to VMAF to make it more suitable for 360° video quality evaluation have been proposed. To this aim, Croci et al. [Cro+19] present a Voronoi-based extension of VMAF. In addition to the Voronoi-based extension of VMAF, the study also presents Voronoi-based extensions for PSNR, SSIM, and MS-SSIM and reports that the Voronoi-based extensions generally outperform their traditional counterparts for 360° video.

More sophisticated models based on neural network approaches have also been proposed. For example, Li et al. [Li+19a] present a viewport-based convolutional neural network (V-CNN) to estimate 360° video quality and is shown to outperform the SoA models. The model is also capable of predicting viewport saliency.

In addition to the mentioned pixel-based models, bitstream and hybrid models could also be used to estimate 360° video quality. One example is Yao, Fan, and Hsu [YFH19], who propose a series of bitstream-based and hybrid models using QP as the bitstream feature and additional features such as spatial genre (simple versus complex), temporal genre (slow- versus fast-paced) and projection scheme. The described models are reported to outperform S-PSNR-I and V-PSNR based on a three-fold cross-validation approach. Furthermore, Fremerey et al. [Fre+20] presented lightweight metadata-based and hybrid models for the quality assessment of 360° videos. The hybrid model calculates spatial and temporal information (SI, TI, cf. [ITU99]) as input features, in addition to metadata such as bitrate, framerate, and resolution. Both presented models show performance comparable with the SoA models such as VMAF, ADM2, WS-SSIM, and VIF.

Besides the aforementioned models, extensions to existing bitstream models were proposed to accommodate 360° video-specific transmission aspects such as tile-based streaming. In this regard, Koike et al. [Koi+21] introduced a tile-based extension of the recently standardized ITU-T Rec. P.1204.3 (i.e., the proposed *AVQBits|M3* model), and report good performance in comparison with subjective test results.

6.2.2 360 Streaming Video Quality Dataset

The 360 Streaming Video Quality Dataset [Fre+20] consists of a total of three different subjective tests. The playback, subjective score and head-rotation data collection was automated using the publicly available AVTrack [Fre+18] software¹. The participants were instructed that they could freely explore the 360° videos. A criterion based on PCC with a threshold of 0.7 was used to detect outliers in all three tests.

test_1 and test_2 had a joint objective of comparing the effect of different HMDs on the perceived video quality. Hence, both tests include the same SRCs and HRCs. Eight SRCs with a resolution of 3840×1920 pixels, framerate of 30 *fps*, and duration of 20 *s* were used in these tests. The bitrate and resolutions chosen in the two tests are detailed in Table 6.5. H.265 was used to encode the videos. A 2-pass encoding approach with the preset of *slow* was chosen. The eight SRCs were encoded with the

¹<https://github.com/Telecommunication-Telemedia-Assessment/AVTrack360>

defined HRCs and resulted in a total of 64 PVSs including high-quality audio. In test_1, the videos were presented using an HTC Vive HMD, and in test_2, using an HTC Vive Pro. The total test duration of each test was 90 minutes.

Table 6.5: HRCs for test_1 and test_2.

Resolution	Target Bitrate (Mbps)			
1920×1080	0.5	1	3.5	7
3840×1920	1	2	6	12

In test_1, all the 64 PVSs were rated by a total of 27 participants. 6 outliers were detected following the criterion of $PCC < 0.7$. 27 participants took part in test_2. There were 3 outliers detected in this test.

test_3 focused on the quality assessment of high resolution ($> 3840 \times 1920$) content. For this test, seven SRCs of 7680×3840 pixels were selected. The framerate of the selected SRCs was 30 *fps*, and sequence duration was 20 *s*. There was no overlap with the SRCs from test_1 and test_2. The videos were encoded at three different resolutions, namely, 3840, 5760×2880 and 7680×3840 pixels. Three bitrates each were used for each resolution, the details of which are described in Table 6.6. As in test_1 and test_2, H.265 was used to encode the videos following a 2-pass encoding approach with *slow* preset. In total, 63 PVSs were rated by 27 participants, with 4 outliers detected according to the criterion of $PCC < 0.7$. The PVSs were presented on an HTC Vive Pro HMD.

The overall MOS distribution of all three tests is as illustrated in Figure 6.3.

Table 6.6: HRCs for test_3.

Resolution	Target Bitrate (Mbps)			
3840x1920	0.5	2	6	
5760x2880	1	4.5	13.5	
7680x3840	2	8	24	

6.2.3 Evaluation

As in the case of gaming, no retraining of the proposed models has been performed. In addition to this, as with the gaming use-case, in this thesis, *AVQBits*|M3 / P.1204.3

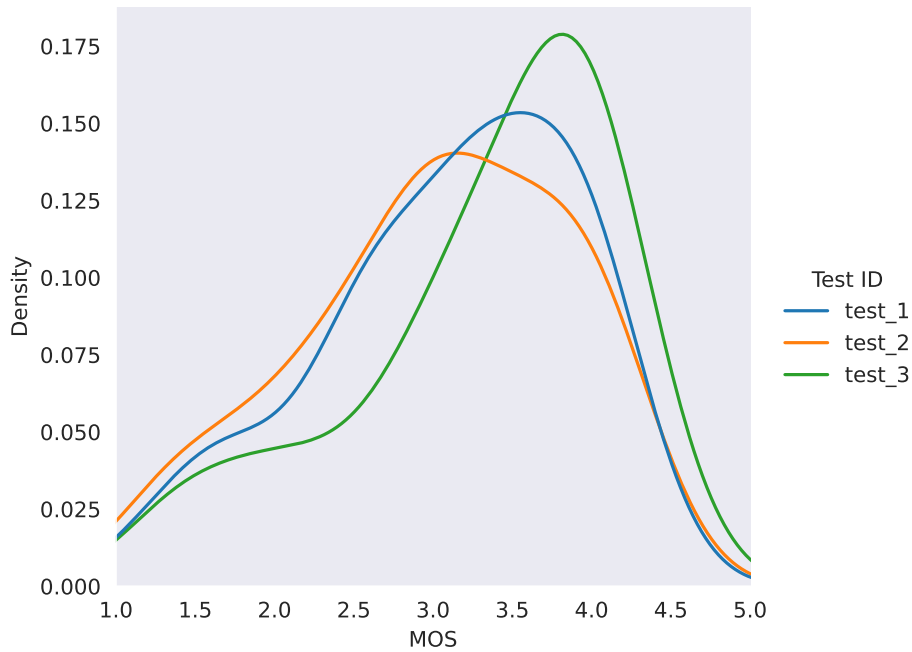


Figure 6.3: MOS distribution of 360 Streaming Video Quality Dataset.

and its extensions are applied directly with the target resolution of 4K/UHD-1 despite the different tests that are considered, have different target resolutions. Table 6.7 provides a detailed view of the performance numbers for all the tests for the proposed bitstream-based models.

Table 6.7: Performance of AVQBits instances using the 360 Streaming Video Dataset.

Test	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
test_1	AVQBits M3 / P.1204.3	0.319	0.917	0.880	0.709	0.841
test_1	AVQBits M1	0.482	0.798	0.785	0.612	0.637
test_1	AVQBits M0	0.558	0.717	0.757	0.578	0.514
test_1	AVQBits H0	0.343	0.903	0.872	0.700	0.816
test_2	AVQBits M3 / P.1204.3	0.314	0.926	0.917	0.757	0.858
test_2	AVQBits M1	0.452	0.841	0.849	0.669	0.707
test_2	AVQBits M0	0.527	0.775	0.834	0.653	0.600
test_2	AVQBits H0	0.314	0.927	0.916	0.753	0.859
test_3	AVQBits M3 / P.1204.3	0.495	0.824	0.707	0.504	0.679
test_3	AVQBits M1	0.780	0.324	0.224	0.134	0.105
test_3	AVQBits M0	0.770	0.382	0.405	0.267	0.146
test_3	AVQBits H0	0.395	0.880	0.796	0.596	0.775

In general, it can be observed from Table 6.7 that $AVQBits|M3 / P.1204.3$ performs well for all the tests. Mode 0 ($AVQBits|M0$) and Mode 1 ($AVQBits|M1$) show satisfactory performance for test_1 and test_2, but perform considerably worse for test_3. The general tendency toward the worse performance of these models can be attributed to the fact that the QP estimation is not optimal for 360° video, as encoders may use different strategies in QP selection for specific bitrates. Hence, a more use-case-specific QP estimation should be considered to enhance the model performance. Especially for these low-complexity bitstream models, a dedicated model could be used, which is usually even how it is handled for existing 2D video streaming applications, due to the sheer amount of different encoding strategies [Rob+22]. The difference in performance for the different bitstream-based models can also be observed in the scatter plots depicted in Figure 6.4. Here, it can be seen that both $AVQBits|M0$ and $AVQBits|M1$ suffer from large prediction errors for certain cases. The difference in performance between the proposed models is most prominent for test_3 which involved a comparison between 4K, 6K, and 8K 360° videos. It should be noted that the proposed models have only been trained and validated on videos up to 4K/UHD-1 resolution. Furthermore, from the results for the Hybrid No-reference Mode 0 model $AVQBits|H0|s$ it can be seen that the model performs well for all three tests, and is on par with the performance of $AVQBits|M3 / P.1204.3$. The $AVQBits|H0|s$ model performs significantly better than $AVQBits|M0$ and $AVQBits|M1$ due to its ability to better estimate the complexity of the content compared to either $AVQBits|M0$ or $AVQBits|M1$, as it can use the entire bitstream information of the QEB. $AVQBits|H0|f$ is not explicitly considered for evaluation because the codec used to encode videos in the test was H.265, which is the default codec for $AVQBits|H0|f$ and hence both $AVQBits|H0|s$ and $AVQBits|H0|f$ are the same models in this case.

In Table 6.8, a comparison of the proposed bitstream models with a number of SoA models is reported. The performance numbers for the SoA models are taken directly from the work by Fremerey et al. [Fre+20]. It can be observed that $AVQBits|M3 / P.1204.3$ performs on par with the best performing FR model, i.e. VMAF. It is further shown that the Mode 0 model proposed in [Fre+20], hereafter referred to as “Mode0F”, performs better than the Mode 0 model $AVQBits|M0$ proposed in this thesis. It should be noted that the Mode0F model was specifically trained for the

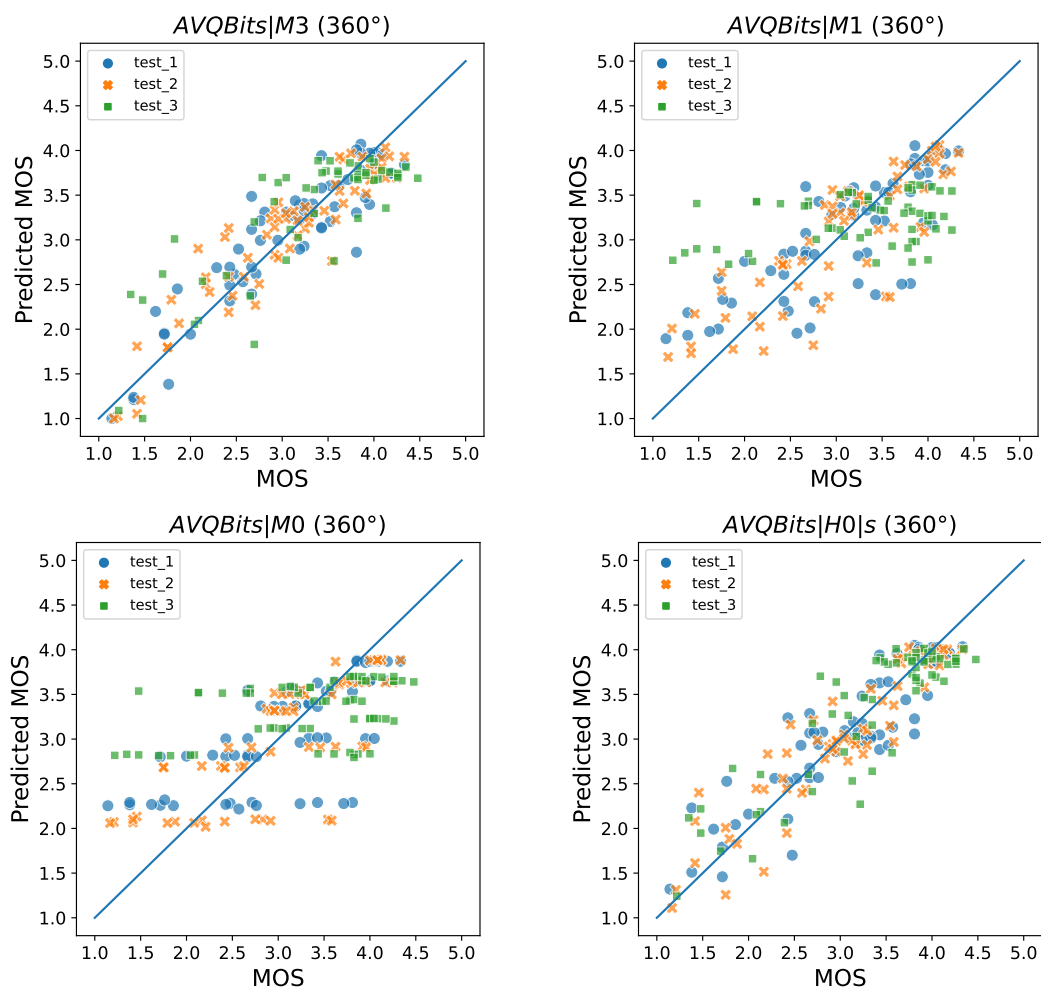


Figure 6.4: Scatter plot of AVQBits instances for 360 Streaming Video Quality Dataset.

360° video use-case and the performance numbers reported in Table 6.8 are based on a 50:50 training-validation strategy. Moreover, the sources in test_1 and test_2 are the same, which leads to an increase in the prediction accuracy of the Mode0F model. Furthermore, it can be seen that the proposed AVQBits|H0|s model outperforms the hybrid model proposed by Fremerey et al. [Fre+20], despite not being specifically trained for 360° videos. This is due to the fact that a more holistic approach is proposed in this thesis with the QEB, using re-encoded bitstream features that are considerably more indicative of content complexity in the Random Forest part of the underlying AVQBits|M3 / P.1204.3 model than the SI and TI information used in [Fre+20].

Table 6.8: Comparison of performance of *AVQBits* instances with SoA models using the 360 Video Streaming Quality Dataset.

Model	RMSE	PCC	SROCC	Kendall
Hybrid [Fre+20]	0.425	0.891	0.890	0.714
Mode 0 [Fre+20]	0.503	0.832	0.865	0.680
VMAF_cc [GKR19]	0.384	0.898	0.872	0.700
VMAF [Net18]	0.431	0.870	0.834	0.664
ADM2 [Li+11]	0.494	0.825	0.819	0.640
WS_SSIM	0.500	0.820	0.864	0.671
VIFP [SB06]	0.554	0.773	0.656	0.502
WS_PSNR	0.598	0.729	0.767	0.582
SSIM [Wan+04]	0.622	0.702	0.730	0.563
PSNR	0.762	0.489	0.627	0.469
<i>AVQBits</i> M3 / P.1204.3	0.377	0.894	0.870	0.679
<i>AVQBits</i> M1	0.581	0.709	0.677	0.497
<i>AVQBits</i> M0	0.627	0.658	0.686	0.401
<i>AVQBits</i> H0	0.356	0.906	0.886	0.695

6.3 High Framerate Video

The third extended application scope is the quality estimation of HFR videos. The LIVE-YT-HFR [Mad+20b] dataset is used to investigate this. A brief overview of the SoA related to HFR video quality assessment along with the details of the dataset and the evaluation process is presented in the following sections.

6.3.1 Related Work for Quality Assessment of HFR Videos

The 4K/UHD-1 and 8K/UHD-2 standards cover higher framerates compared to traditional cinema or TV, which usually has 24 fps or 30 fps. In the following section, the SoA will be briefly analyzed considering the video quality assessment/prediction of videos with a higher framerate of > 60 fps. A study on the impact of framerate on perceived quality was conducted by Mackin, Zhang, and Bull [MZB15] in which videos with framerates varying from 15 Hz to 120 Hz were analyzed. The subjective evaluations conducted using these videos show a significant relationship between framerate and perceived video quality. Further, it was observed that the effect of framerate on perceived video quality is content dependent. The study also reports diminishing improvements in terms of quality as framerates increase.

Furthermore, Mackin, Zhang, and Bull [MZB19] develop a high-framerate video quality database, BVI-HFR, containing videos captured at a framerate of 120 *fps*. Based on their tests they conclude that models such as FRQM [ZMB17] which explicitly account for temporal distortions are more accurate in predicting video quality as compared to traditional metrics such as PSNR.

In addition to this, Madhusudana et al. [Mad+21] conduct a large-scale study on the subjective and objective quality of high framerate video with framerates up to 120 *fps*. For this purpose, a large dataset called the LIVE-YouTube-HFR (LIVE-YT-HFR) with 480 PVSs is created, which are subjectively evaluated by a total of 85 participants. The LIVE-YT-HFR dataset is made publicly available². An evaluation of existing FR and NR models has been performed, and it has been reported that the GSTI [Mad+20a] model outperforms all the SoA models including VMAF. GSTI uses a statistical entropic differencing method based on a Generalized Gaussian Distribution model expressed in both the spatial and temporal band-pass domains to measure the difference in quality between reference and distorted videos.

Furthermore, Lee et al. [Lee+21] conducted a subjective and objective assessment of the video quality of space-time subsampled videos. The ETRI-LIVE Space-Time Subsampled Video Quality (ETRI-LIVE STSVQ) database was created for this purpose and contains a total of 437 PVSs with framerates varying between 30 *fps*, and 120 *fps*. The evaluation shows that the VSTR model proposed by Lee et al. [Lee+20], which is specifically developed to take into account the joint perceptual effects of spatio-temporal subsampling and compression, outperforms all the considered SoA models including VMAF.

6.3.2 LIVE-YT-HFR Dataset

The LIVE-YT-HFR [Mad+20b] dataset was designed to analyse the impact of framerate on perceived video quality, like test_4 of the AVT-VQDB-UHD-1 dataset. For this purpose, 16 SRCs captured at a framerate of 120 *fps* were used. Eleven out of the 16 SRCs are from the BVI-HFR dataset [MZB15]. Although these 11 SRCs were captured at 3840×2160 pixels resolution, the publicly available version of the dataset consists

²https://live.ece.utexas.edu/research/LIVE_YT_HFR/LIVE_YT_HFR/index.html

of SRCs downsampled to 1920×1080 (FHD) resolution, with a sequence duration of 10 s. The remaining five SRCs are of 3840×2160 (UHD) resolution, framerate of 120 *fps*, with a duration of 6 – 8 s. They mainly consist of sports content with high motion. Six different framerates were included in the study, namely 24, 30, 60, 82, 98, and 120 *fps*. All the SRCs were encoded with VP9 at five different CRF values for each framerate, thus resulting in 30 PVSs for each source and a total of 480 PVSs. The dataset is divided into four subsets, with each subset containing 120 PVSs. 85 participants took part in the subjective test, with each subject rating two out of the four subsets, thus rating a total of 240 PVSs in two sessions. The test sequences, both FHD and UHD, were presented on a 27" UHD-1 screen. Each PVS was rated by a minimum of 12 participants. The quality distribution of the PVSs in terms of MOS is as shown in Figure 6.5.

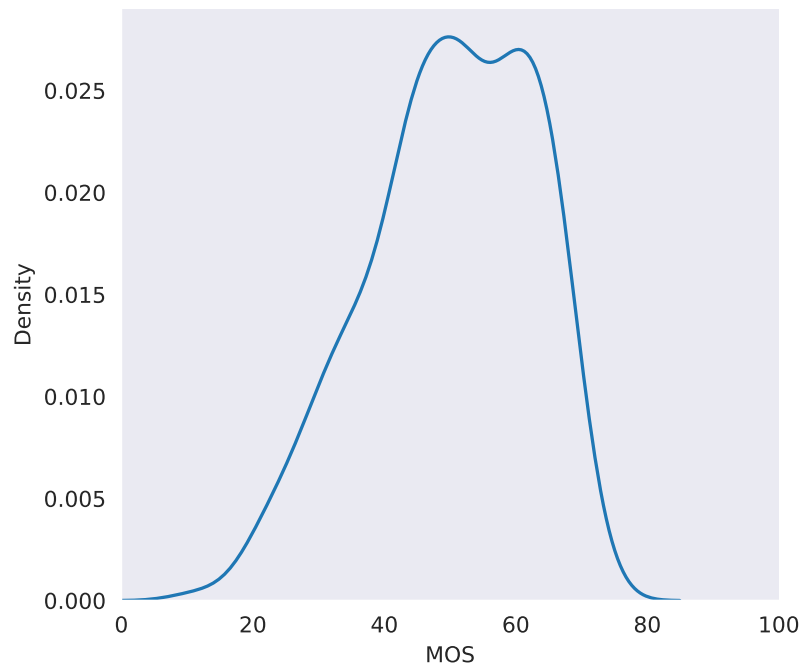


Figure 6.5: MOS distribution of LIVE-YT-HFR dataset.

6.3.3 Evaluation

Although this use-case falls into the broad category of traditional 2D videos, the HFR use-case is still considered an extended application scope as the proposed models

have been trained and validated only for video of framerate up to 60 *fps*. As was the case with gaming and 360° video, no retraining was performed on the proposed AVQBits model instances for the specific use case.

Table 6.9 compares the performance of the proposed AVQBits models with SoA models for each framerate. The performance numbers for the SoA models are taken directly from the work by Madhusudana et al. [Mad+20b]. In general, it can be observed that AVQBits|M3 / P.1204.3 model performs on par with VMAF for all framerates. The performance is similarly good for the hybrid models AVQBits|H0|s and AVQBits|H0|f, although AVQBits|H0|f with its fixed encoder shows a slightly worse performance. The results for this model variant could be enhanced by a dedicated retraining of the a_{cmap} and b_{cmap} for HFR-specific content. The Mode 0 (AVQBits|M0) and Mode 1 (AVQBits|M1) models show similar performance to that of SSIM, MS-SSIM, ST-RRED, and FRQM. It can also be seen that prediction accuracy in terms of both PCC and SROCC is significantly worse for lower framerates than for higher framerates for all the proposed models. This is due to the fact that the *temporal degradation* component of the “Core Model” considers 60 *fps* as the maximum framerate as that was the framerate of the used display for subjective testing for both AVT-PNATS-UHD-1 and AVT-VQDB-UHD-1. The temporal degradation associated with the perceived video quality is then estimated relatively to 60 *fps* thereby underestimating the impact of lower framerates on perceived video quality when viewed on a display with a higher framerate such as 120 *fps*. The models show significantly better performance at higher framerates (≥ 60 *fps*), as the effect of temporal degradation on perceived video quality decreases at higher framerates. This is consistent with findings presented in [MZB15]. Figure 6.6 illustrates the scatter plots for the different AVQBits variants on the LIVE-YT-HFR dataset.

6.4 Live Streaming Sports

One of the biggest challenges in a live streaming scenario is minimizing latency as it has a significant impact on the QoE of the end-user. Latency can be due to any of the following factors in a live streaming set-up: video encoding pipeline, ingest and packaging, network propagation, CDN, and player policies [AWS19]. As the video

Table 6.9: Comparison of performance of AVQBits instances with SoA models using the LIVE-YT-HFR dataset.

Model	24fps			30fps			60fps			82fps			98fps			120fps			Overall		
	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	
PSNR	0.4101	0.3647	0.4414	0.4179	0.6202	0.5719	0.6878	0.6431	0.7171	0.6489	0.6019	0.5937	0.6950	0.6685							
SSIM [Wan+04]	0.1277	0.0949	0.1108	0.0816	0.2123	0.1845	0.2079	0.2430	0.3876	0.3964	0.7485	0.6726	0.4494	0.4526							
MS-SSIM [WSB03]	0.2221	0.1500	0.1929	0.1112	0.2516	0.1900	0.2906	0.2549	0.4237	0.4007	0.6165	0.5843	0.4898	0.4673							
FSIM [Zha+11]	0.3670	0.3038	0.3208	0.2638	0.2472	0.2615	0.3225	0.3055	0.3861	0.2646	0.3056	0.1178	0.4469	0.4435							
ST-RRED [SB13]	0.1541	0.0369	0.1188	0.0307	0.5062	0.4457	0.3394	0.3271	0.4962	0.4556	0.6745	0.5906	0.5531	0.5107							
SpEED [Bam+17]	0.2591	0.1237	0.2278	0.0896	0.1824	0.1110	0.2955	0.2425	0.4118	0.3295	0.6827	0.6097	0.4861	0.4449							
FRQM [ZMB17]	0.1556	0.2089	0.0983	0.0854	0.0947	0.0309	0.0137	0.0035	0.0317	0.0100	-	-	0.4216	0.4520							
VMAF [Net18]	0.1743	0.2669	0.2855	0.3740	0.5408	0.6015	0.6820	0.7390	0.8214	0.8128	0.7943	0.7844	0.7303	0.7071							
deepVQA [Kim+18]	0.1144	0.0495	0.1353	0.1059	0.2527	0.1652	0.1803	0.1515	0.2816	0.2654	0.6865	0.6209	0.3463	0.3329							
GSTII [Mad+20a]	0.4554	0.5827	0.5079	0.6664	0.6853	0.7507	0.7584	0.8194	0.7886	0.7953	0.7508	0.7258	0.7983	0.7917							
AVQBits M3 / P.1204.3	0.5395	0.7806	0.6244	0.8619	0.7722	0.8921	0.8325	0.9145	0.8548	0.9125	0.8752	0.9184	0.7118	0.7805							
AVQBits M1	0.4990	0.6870	0.5433	0.7223	0.7086	0.7638	0.7129	0.7326	0.7417	0.7477	0.7476	0.7218	0.4809	0.5528							
AVQBits M0	0.5053	0.6643	0.5453	0.6953	0.6961	0.7320	0.6788	0.6893	0.7261	0.7061	0.7273	0.6832	0.4947	0.5538							
AVQBits H0 s	0.5535	0.7784	0.6459	0.8326	0.7794	0.8633	0.8133	0.8814	0.8519	0.8843	0.8849	0.8855	0.7324	0.7887							
AVQBits H0 f	0.5680	0.7806	0.6362	0.8303	0.7789	0.8591	0.7832	0.8572	0.8219	0.8424	0.8421	0.8498	0.6740	0.7242							

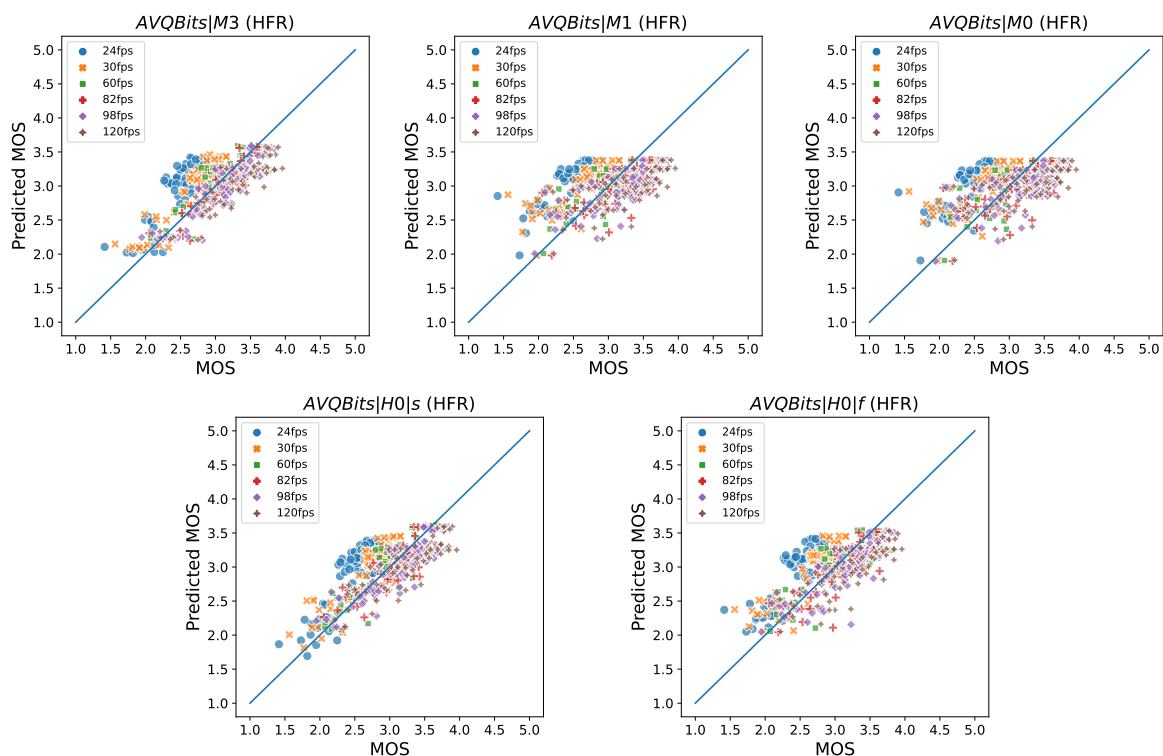


Figure 6.6: Scatter plot of AVQBits instances for LIVE-YT-HFR dataset.

encoding pipeline is one of the factors contributing to the overall latency, dedicated encoding settings are used and can vary in comparison to traditional VoD services. Hence, models developed mainly using particular encoder implementations and for a defined set of encoding settings targeted mainly towards VoD scenarios have to be tested for their efficiency to be applicable to live streaming content. With this objective, the AVQBits instances will be investigated for their applicability for live streaming video quality evaluation. For this, the LIVE-Amazon Prime Video (LIVE-APV) dataset consisting videos of subjected to distortions salient to live streaming is used. Before discussing the results of this evaluation, a brief overview of the existing SoA quality evaluation in live streaming scenarios.

6.4.1 Related work

One notable publicly available video quality assessment database for live streaming is the LIMP video quality database [Tor+16]. This database consists of nine videos

from the Live Quality Video Database [Ses+10]. These videos were then subjected to 12 levels of randomized packet loss to create PVSs. Vega et al. [Veg+17] use this database to evaluate their proposed unsupervised deep-learning model. However, this database suffers from the drawback that the considered videos are of low resolution (768×432) and has only packet-loss degradations.

To overcome this drawback, Shang et al. [Sha+22] develop a database focused on distortions normally encountered in live streaming scenarios. This database is developed using high-resolution high-motion sports content. The details of this database are provided in the subsequent section. In addition to conducting subjective tests using this database, a holistic evaluation of various FR and NR models is performed for this specific use case.

In summary, the increasing amount of live streamed content using HAS-based mechanisms necessitates the need for a quality model to be able to evaluate such different encodings, and hence the proposed models are tested for their applicability for this use case.

6.4.2 LIVE-APV Dataset

The LIVE-APV dataset consists of 315 PVSs derived from 45 different SRCs of duration ranging between 5-8 s. The 45 different SRCs are extracted from 33 different footages. The considered footages were all pristine videos with a resolution of either 1920×1080 or 3840×2160 pixels and had a framerate of 30 fps. This footage was gathered from different publicly available sources.

Six different distortions were applied to the 45 SRCs. These include H.264 compression, aliasing, judder, flicker, frame drops, and interlacing. In this evaluation, the focus is only on the H.264 compression-related distortions as the models that form *AVQBits* are only capable of handling compression-related distortions. Four different CRF values were considered for compression. The four CRF values used for videos with a resolution of 1920×1080 pixels were 9, 25, 35, and 39 and for 4K/UHD-1 videos they were 9, 27, 39, and 43. One CRF value was chosen for each video as a full-factorial test design by considering all six distortions would result in a large number of PVSs and hence would become infeasible for test conditions.

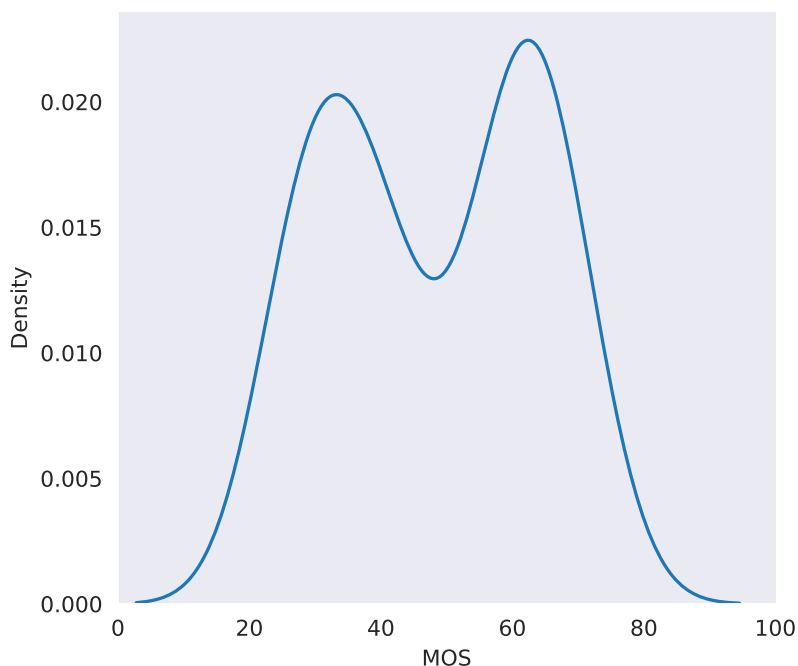


Figure 6.7: MOS distribution of LIVE-APV dataset.

This resulted in 45 PVSs with compression-related distortions and these are used to evaluate the *AVQBits* model instances. The PVSs were rated on a 0-100 scale. The overall MOS distribution of these 45 PVSs is shown in Figure 6.7.

6.4.3 Evaluation

For performance comparison of the models, different SoA NR and FR models were used in addition to the different variants of *AVQBits*. In the case of the proposed models, a simple linear mapping [ITU14a] was applied to the predicted scores to compute the RMSE. Whereas for the other supervised methods, retraining with 1000 random splits with each split having 80% data for training and 20% for testing was performed and within each training set a 5-fold cross-validation approach was used. In doing this retraining, it was ensured that there was no overlap in content between the training and testing set. For unsupervised methods, the predicted scores were re-mapped using a non-linear logistic regression process before computing RMSE and PCC. This additional corresponding retraining and logistic mapping were

6.5 Quality Evaluation of Videos with Pre-Existing Distortions

performed by the original developers of the database, Shang et al. [Sha+22] and the numbers corresponding to the SoA models reported in Table 6.10 are from that paper.

Although just a simple linear mapping was used for all models of *AVQBits*, their performance in terms of RMSE, PCC, and SROCC is comparable to other models that have either been retrained or been re-mapped using a non-linear logistic regression process. This shows that the developed models are well-suited for such content.

Table 6.10: Comparison of performance of *AVQBits* instances with SoA models using the *LIVE-APV* dataset.

Metric	RMSE	PCC	SROCC
NIQE	45.7805	0.2805	0.2775
BRISQUE	9.1434	0.7616	0.6409
CORNIA	8.0173	0.8197	0.7399
HIGRADE	7.7381	0.8395	0.7234
V-BLIINDS	7.7836	0.8313	0.7131
TLVQM	10.0801	0.6991	0.6574
ChipQA	7.7510	0.8408	0.7482
PSNR	4.2304	0.9586	0.8750
SSIM	3.8493	0.9659	0.9171
MS-SSIM	3.6708	0.9690	0.9154
SpEEDQA	4.5223	0.9526	0.8979
ST-RRED	4.7155	0.9483	0.8943
FAST	4.2267	0.9587	0.9283
VMAF	3.7600	0.9675	0.9135
<i>AVQBits</i> M3 / P.1204.3	4.5095	0.9568	0.8996
<i>AVQBits</i> M1	6.6759	0.9026	0.8711
<i>AVQBits</i> M0	7.9932	0.8570	0.8278
<i>AVQBits</i> H0 s	5.6486	0.9313	0.8706
<i>AVQBits</i> H0 f	7.8256	0.8634	0.8755

6.5 Quality Evaluation of Videos with Pre-Existing Distortions

Due to the perpetration of affordable high-quality capture devices, the amount of UGC has increased significantly on the internet, most notably on social media such as TikTok, Facebook, YouTube, etc. These videos are partly characterized by the lack of

pristineness in the source content. Also, these videos have often undergone a series of processes such as editing, compression, etc. before being uploaded to a particular online server. Furthermore, these videos also undergo further compression on the servers where they are uploaded. Here, it is necessary to be able to guide this further compression either without a reference video or by only using the distorted reference instead of a pristine reference video as is the case in VoD services. For this purpose, it becomes very important to have video quality models that can accurately predict the quality of such videos with pre-existing distortions. It should be noted in this evaluation only SoA UGC datasets for which a lab test was conducted by respective authors have been considered.

6.5.1 Related work

Traditionally, the focus of video quality databases has mainly been on simulated distortions related to compression and transmission [Rao+19a; CL18; Ber+15; Mad+21]. However, in recent years, there has been a significant increase in research on the quality assessment of videos with pre-existing distortions. This includes both development of large-scale datasets of videos with pre-existing distortions with subjective annotations and also of models for quality evaluation of such videos. One example of a UGC-relevant dataset is the CVD2014 database [Nuu+16] focused mainly on camera distortions. This database consists of videos captured from 78 different capture devices. Following this, other UGC-related databases have been made public. Of them, the LIVE-Qualcomm mobile in-capture video quality database [Gha+18] comprising 208 videos having six common in-capture distortions, and the KoNViD-1k database [Hos+17] consisting of 1200 videos sampled from a larger YFCC database [Tho+16] are notable ones. The creators of the KoNViD-1k database further extended their in-the-wild dataset and created the KonViD-150k database [Göt+21a; Göt+21b]. This database consists of two parts, namely, KonVid-150k-A and KonVid-150k-B. The KonVid-150k-A part consists of 152,265 videos of 5 s duration with each video having five quality ratings. The KonVid-150k-B part has 1577 videos with a minimum of 89 ratings for each video.

Recently, YouTube developed a large-scale YouTube-UGC dataset [WIA19] consisting of 1500 videos of 20 s duration sampled from millions of content on YouTube and

6.5 Quality Evaluation of Videos with Pre-Existing Distortions

covers different genres such as gaming, sports, etc., and also aspects such as HDR. Also, the dataset is evaluated with three no-reference metrics, namely, noise, banding, and SLEEQ. Furthermore, Yim et al. [Yim+20] conducted a large-scale crowdsourcing study using the YouTube-UGC dataset and collected subjective ratings for all videos with more than 100 ratings per video. In addition, subjective ratings for three overlapping 10 s chunks are collected with the objective of finding the relationship between full video quality and chunk quality.

In parallel to the creation of these databases, there have been models that have been developed for quality prediction of such videos with pre-existing distortions. These models have been predominantly based on an NR approach as there is no availability of pristine reference videos in such a scenario. An example of such a model is the VIDEO quality EVALuator (VIDEVAL) [WIA19] developed by the authors of the YouTube-UGC dataset. The results show that the considered approach shows similar performance to the existing DNN-based SoA models such as FRIQUEE [GB16b] and TVLVQM [Kor19] using a lesser number of features compared to both these models.

Furthermore, Wang et al. [Wan+21] introduced a framework to analyze aspects such as the importance of content, technical quality, and compression level in perceptual quality for UGC videos with the results showing comparable performance to the best performing SoA model TVLVQM. In addition to achieving higher prediction accuracy, the focus of model development has also been on achieving improved performance with considerably lower computational complexity [Tu+21]. Also, Yu et al. [Yu+21] address the problem of predicting the quality of compressed videos whose reference video were distorted UGC. For this purpose, a large database consisting of UGC videos and their compressed versions of them, called “LIVE Wild Compressed Video Quality Database” is created by Yu et al. [Yu+21]. Using this database, the authors develop a framework called 1stepVQA for video quality assessment. This database is used to assess the applicability of the *AVQBits* model instances for quality prediction of distorted videos using videos with pre-existing distortions as source content, using the available data in this thesis.

To summarize, it is evident that there has been an increasing focus on the quality evaluation of videos with pre-existing distortions and hence any video quality model that is developed should be applicable to such videos. Hence, the models proposed in this thesis are investigated for this purpose in the following section.

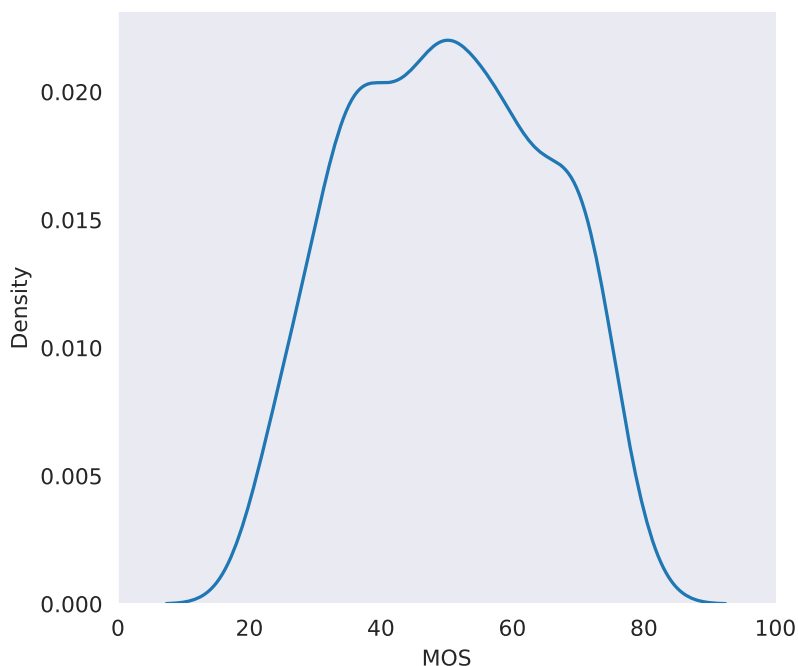


Figure 6.8: MOS distribution of *LIVE Wild Compressed Video Quality Database* dataset.

6.5.2 LIVE Wild Compressed Video Quality Database

Firstly, a set of 55 different randomly selected reference videos from 110 videos of a resolution of 1920×1080 pixels from the LIVE VQC database [SB19] was used. These videos are of 10 s duration each and have been captured with a wide range of mobile cameras. In [Yu+21], each of the 55 videos was subjected to H.264 compression at 17 different settings using 17 different CRF values, such as CRF: 1, 4, 7, 10,, 49 at four different resolutions of 1080p, 720p, 540p, and 360p. This resulted in a total of 3740 ($55 \text{ SRCs} \times 4 \text{ resolutions} \times 17 \text{ crf values}$) PVSs. Using a realistic VMAF-Guided perceptual rate-distortion optimization (RDO) criterion, 220 PVSs spanning VMAF scores between 20 and 90 were selected. These 220 PVSs in addition to the 55 original videos form the LIVE Wild Compressed Video Quality Database. As VMAF was used to sample the PVSs, it is not included in the performance analysis. The PVSs were rated by subjects on a scale of 0-100 and the overall MOS distribution is illustrated in Figure 6.8.

6.5 Quality Evaluation of Videos with Pre-Existing Distortions

Table 6.11: Comparison of performance of *AVQBits* instances with SoA models using the *LIVE Wild Compressed Video Quality Database*.

Metric	RMSE	PCC	SROCC
NIQE	9.1012	0.7149	0.7150
BRISQUE	8.0702	0.7887	0.7877
V-BLIINDS	7.8482	0.8228	0.8276
TLVQM	7.5766	0.8303	0.8381
VSFA	7.0889	0.8339	0.8519
PSNR	11.2782	0.5074	0.5084
MS-SSIM	8.3797	0.7744	0.7856
FSIM	6.3419	0.8776	0.8778
ST-MAD	7.7239	0.8141	0.8197
VSI	8.1939	0.7806	0.7813
2stepQA	7.1600	0.8455	0.8493
1stepVQA	6.0275	0.8902	0.8918
1stepVQA-R	5.1551	0.9224	0.9236
<i>AVQBits</i> / P.1204.3	7.0796	0.8732	0.8728
<i>AVQBits</i> M1	6.6951	0.8875	0.9017
<i>AVQBits</i> M0	7.1317	0.8712	0.8925
<i>AVQBits</i> H0 s	7.0464	0.8745	0.8822
<i>AVQBits</i> H0 f	6.7218	0.8865	0.8911

6.5.3 Evaluation

In the following thesis contribution, the *AVQBits* models are tested for their applicability to a particular use case of predicting the quality of compressed videos with pre-existing distortions. In addition to this test, the performance of the *AVQBits* model instances is compared with the SoA FR and NR models. The developers of the database, Yu et al. [Yu+21] perform a similar re-training and mapping as explained in Section 6.4.3 before computing the performance metrics such as RMSE, PCC and SROCC. The performance numbers corresponding to these SoA models outlined in Table 6.11 have been taken from this work. Like with all other previous evaluations, the predictions from the *AVQBits* models are linearly mapped before computing the RMSE values.

From Table 6.11, it can be seen that the proposed models perform better than the SoA NR models with a simple linear mapping. This shows that the *AVQBits* models can take into account the effect of pre-existing distortions in the source video and thereby predict the quality of the compressed versions of such source content well

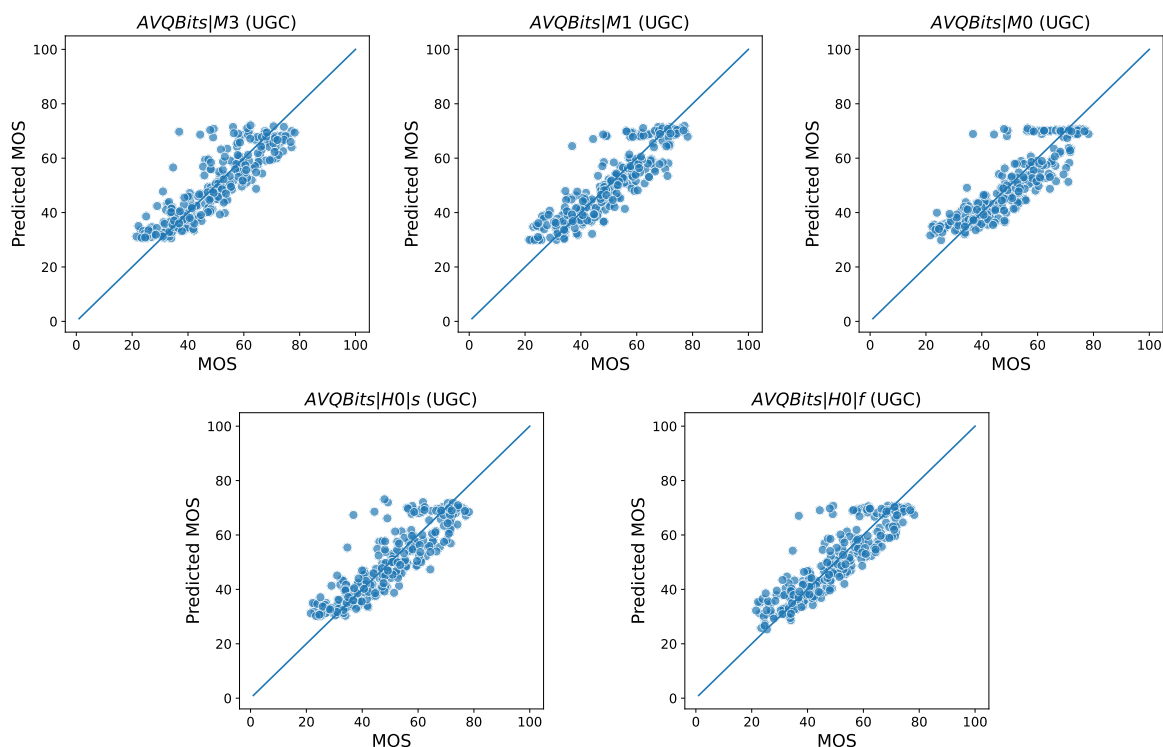


Figure 6.9: Scatter plot of AVQBits instances for LIVE Wild Compressed Video Quality Database.

too. Furthermore, Figure 6.9 illustrates the scatter plot of all the three bitstream models.

6.6 Image Quality Evaluation

JPEG, GIF, and PNG have traditionally been the most popular image compression formats in recent years. However, there has been a steady increase in the development of new formats such as WebP³, BPG⁴, HEIF [Lai+16]⁵ and AVIF⁶. These new formats share the commonality of being based on video codecs with WebP based on VP8, BPG and HEIF on HEVC and AVIF using AV1. Such a development results in a question of whether the video quality models developed for quality

³<https://developers.google.com/speed/webp/>

⁴<https://bellard.org/bpg/>

⁵<https://nokiotech.github.io/heif/>

⁶<https://aomediacodec.github.io/av1-avif/>

evaluation of videos encoded with the underlying video codecs can be adapted for quality prediction of images using the corresponding image formats. To answer this question, the $AVQBits|M3$ and $AVQBits|M0$ variants of $AVQBits$ are used for quality evaluation of images encoded using video codecs such as H.264, H.265, and VP9. For this purpose, the IC_{test} [GRR23] is used.

6.6.1 Related work

A comparison of WebP with other image compression methods such as JPEG, JPEG-XR, and JPEG-2000 was conducted by Pintus et al. [Pin+11]. For comparison, PSNR and SSIM were used. The results indicate that JPEG is better than the WebP format. This low performance of WebP can be attributed to the fact that it is based on VP8 and hence may not have the advantages in terms of compression efficiency that its successor VP9 has. Furthermore, Lainema et al. [Lai+16] compare HEIF which is based on H.265 with JPEG on high-resolution images. The highest resolution considered in this study is 4064×4064 pixels. It is shown that HEIF results in lower file sizes and still produces the same quality in comparison with JPEG. In addition, other studies have also indicated the suitability of HEVC for image compression with good results in comparison with traditional image compression algorithms [NM15; AG17].

Along similar lines, Göring and Raake [GR19] use H.264, H.265, VP9, and AV1 for image compression and compare the results with JPEG using PSNR, SSIM, VMAF, and VIF. Video codec-based image compression outperforms JPEG compression with AV1 performing best amongst all the considered video codecs. This is followed by a subjective evaluation, both lab- and crowd-based, of H.265 encoded images by Göring, Rao, and Raake [GRR23]. The dataset used in this study is used to evaluate the applicability of the $AVQBits|M3$ and $AVQBits|M0$ variants of $AVQBits$ for image quality evaluation. Also, Barman and Martini [BM20] conduct a comparative study of AVIF, the image codec based on AV1, with SoA image codecs with the results showing AVIF having highest bitrate savings across all the considered objective models, namely, VMAF, SSIM, MS-SSIM, VIF, and PSNR.

In essence, there is an increasing scope of using video codecs with intra-frame coding and also image codecs based on video codecs for image compression. Hence, the

adaptation of video quality models to predict image quality needs similar attention for better and faster model development.

6.6.2 Dataset

For quality assessment of high-resolution images, Göring, Rao, and Raake [GRR23] use 4K/UHD-1 frames extracted from several different 4K/UHD-1 videos. In total, 39 such 4K/UHD-1 frames are extracted, and following this, the frames were cropped to have a height and width of 2160 pixels. The frames were extracted from videos covering different genres such as animated content, movies, documentaries, etc. These frames were then encoded with H.265 for several different resolutions from 144×144 pixels to 2160×2160 pixels with a step size of 16 pixels (along both width and height) using CRF values in the range $[0,1,2,\dots,51]$ with a step size of 1 resulting in a total of 246,126 compressed images. A CRF-based one-pass scheme was used to encode the images. Following this, VMAF was computed for all the encoded images. Using VMAF as the criterion and the approach described in [GRR23], 371 images are selected from this larger set for conducting a lab-based subjective test. In the test, these images were presented on a 55" 4K LG OLED55C7D screen with a viewing distance of approximately 1.6 times the height of the screen as recommended in ITU-R BT.500-13 [ITU14b]. The participants rated the images on a 5-point ACR scale. A total of 21 participants took part in the study. The overall distribution of the resulting MOS of the images is shown in Figure 6.10.

6.6.3 Evaluation

For evaluating the applicability of video quality models for quality prediction of images compressed using video codecs, the $AVQBits|M0$ and $AVQBits|M3$ instances of $AVQBits$ are considered. The rationale behind omitting $AVQBits|M1$ in this analysis is that $AVQBits|M1$ essentially becomes a $AVQBits|M0$ model for an image as there is only a single frame and the additional features related to I- and Non-I-frame sizes are redundant. Furthermore, as the models have been originally developed for quality evaluation of videos, the models may have to be adapted for image quality prediction in terms of the available input data. Hence, before describing the details of

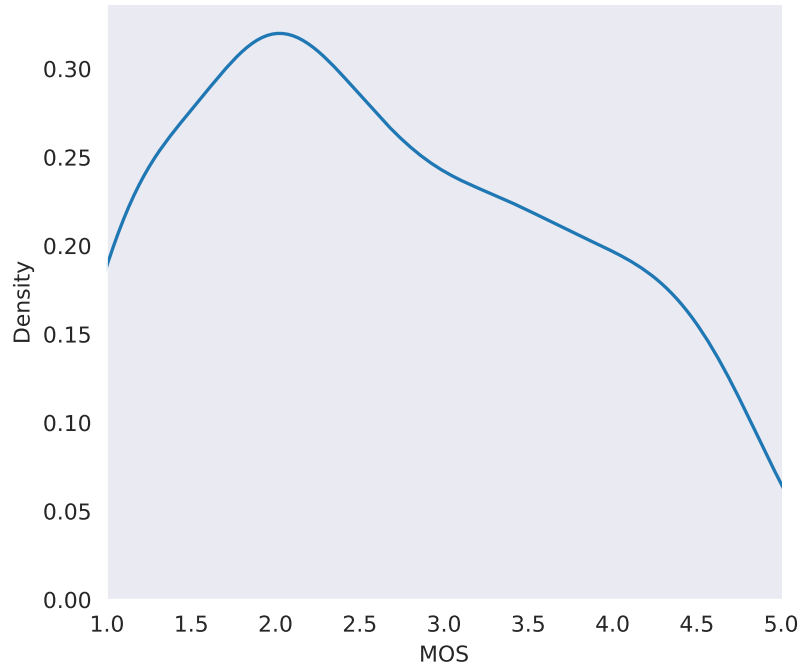


Figure 6.10: MOS distribution of the considered dataset.

the evaluation of the models on the considered dataset, the modifications necessary for the considered models are presented.

6.6.3.1 Mode 3 Modifications for Image Quality Evaluation

The *AVQBits|M3* model has two main components, namely, the traditional curve-fitting part referred to as the “Core Model” and the Random Forest part. The “Core Model” comprises three different degradations namely, quantization degradation, upscaling degradation, and temporal degradation. The quantization degradation is a function of the QP values of the non-I-frames. As an encoded image consists only of one I-frame, the quantization degradation part has to be modified to reflect this. The “new quantization degradation” is now just a function of the corresponding QP value of the compressed image and is given by Equation (6.5). The values of QP_{max} are as described in Section 4.2.1.1 of Chapter 4.

$$quant = \frac{QP}{QP_{max}} \quad (6.5)$$

In addition to this, the temporal degradation component which originally was targeted to model the degradation associated with upsampling of the encoded video to the corresponding display framerate is not applicable for images as there is only one single static frame that is displayed to the user for rating. Furthermore, the Random Forest which has features related to QP, frame sizes, and motion vectors is also not considered as the frame sizes and motion vectors related features are irrelevant for an image. Hence, the modified model consists of only the “Core Model” with the quantization degradation (D_q) and upscaling degradation (D_u) parts as shown in Equation (6.6) with D_u as defined in Equation (4.13).

$$M_{p_{[0,100]}} = 100 - (D_q + D_u) \quad (6.6)$$

$$M_{p_{[1,4.5]}} = \text{MOS}_{\text{fromR}}(M_{p_{[0,100]}}) \quad (6.7)$$

$$M_{\text{par}} = \text{scalet05}(M_{p_{[1,4.5]}}) \quad (6.8)$$

The corresponding coefficients of both the quantization and upscaling degradations remain unchanged and no retraining was done to modify the coefficients.

6.6.3.2 Mode 0 Modifications for Image Quality Evaluation

Similar to the Mode 3 modification, the quantization degradation part is modified to include only the resolution and bitrate components and not the framerate component. This new QP_{pred} calculation is given by Equation (6.6.3.2).

$$\begin{aligned} QP_{\text{pred}} = & a_{qp_m0} + b_{qp_m0} \cdot \log(\text{Bitrate}) \\ & + c_{qp_m0} \cdot \log(\text{Resolution}) \end{aligned} \quad (6.9)$$

The final model consists of only the quantization and upscaling degradations as was the case for AVQBits|M3 and the final MOS-prediction model is as given in Equation (6.6).

6.6.3.3 Results

With the aforementioned modifications, the modified $AVQBits|M0$ and $AVQBits|M3$ variants are evaluated for prediction accuracy along with different SoA FR models. For computing the RMSE, a linear mapping as proposed in ITU-T Rec. P.1401 [ITU14a] is applied to the scores predicted by all the models. The results of the evaluation in terms of RMSE, PCC, SROCC, Kendall correlation, and R^2 Score are detailed in Table 6.12. It can be observed that both the $AVQBits|M3$ and $AVQBits|M0$ models outperform all other models. From the results, it can be concluded that QP is a defining factor for the quality prediction of images encoded with video codecs. Based on the performance of the simple $AVQBits|M0$ model, it can also be stated that bitrate and resolution are good features and hence simple models with this information will result in high prediction accuracy.

Table 6.12: Comparison of performance of $AVQBits|M3$ and $AVQBits|M0$ with SoA models.

Metric	RMSE	PCC	SROCC	Kendall	R^2 Score
PSNR	0.799	0.698	0.719	0.524	0.487
SSIM	0.839	0.658	0.948	0.802	0.434
MS_SSIM	0.796	0.701	0.851	0.658	0.491
VIF_scale0	0.876	0.619	0.643	0.472	0.384
VIF_scale1	0.594	0.846	0.859	0.674	0.716
VIF_scale2	0.566	0.861	0.911	0.740	0.742
VIF_scale3	0.583	0.852	0.941	0.786	0.726
ADM2	0.554	0.868	0.901	0.722	0.754
VMAF	0.440	0.919	0.925	0.757	0.845
$AVQBits M3$ / P.1204.3	0.319	0.958	0.967	0.846	0.918
$AVQBits M0$	0.377	0.942	0.951	0.808	0.886

Figure 6.11 shows the scatter plot for the $AVQBits|M3$ and $AVQBits|M0$ models respectively. It can be seen that for the $AVQBits|M0$ model, there are cases with prediction errors that are significantly larger than the average RMSE and also, the lower bound saturates around a MOS of 1.5. It should be noted that the coefficients have been taken from the video quality counterpart and dedicated retraining for images may somewhat improve on these problems.

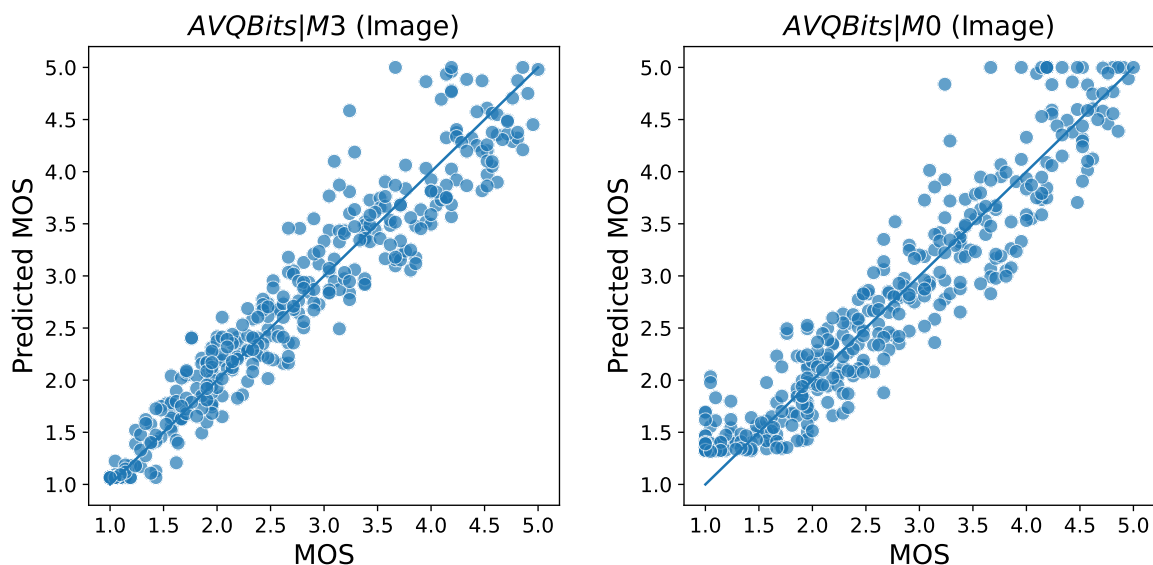


Figure 6.11: Scatter plot of AVQBits instances for the considered dataset.

6.7 Summary

The models proposed in Chapter 4 were mainly developed for the use case of VoD and hence the training and validation data consisted of encoding settings that are typical to this scenario. With the aim of testing the applicability of the proposed AVQBits model variants for use cases other than VoD, the models were tested for a total of six use cases, namely, gaming video, 360° videos, HFR content, videos from a live streaming context, compressed videos created using videos with pre-existing distortions as sources and images encoded with video codecs. For this, as a first step, publicly available databases for each of these use cases were gathered. Following that, a brief description and analysis in terms of the overall MOS distribution were presented. Finally, the performance of the proposed models was analyzed and further compared with SoA models for each scenario. For the calculation of performance metrics for the developed models, only a simple linear mapping as proposed in ITU-T Rec. P.1401 [ITU14a] was used. In turn, for the SoA models, in several cases, the performance numbers calculated using different retraining and non-linear logistic mapping from the original works have been used. This comparison showed that the AVQBits model variants perform either on par with or better than the SoA models.

It was also observed that the *AVQBits*|M3 / P.1204.3 model, the most complex bitstream models, along with *AVQBits*|H0|s were the best performing models as compared to the other models included in the SoA. The other two less complex bitstream models, namely, *AVQBits*|M0 and *AVQBits*|M1 also perform well and in most cases outperform the considered SoA NR models. With this extensive evaluation, it can be concluded that all five variants of *AVQBits* can be applied to other use cases.

This demonstration of the applicability of *AVQBits* for other application scopes along with its usage for the prediction of the overall quality of a HAS session as shown in Chapter 5 shows the versatility of the *AVQBits* model instances, and thus addresses the objectives outlined in research question 3. A conclusion of this thesis along with outlook for future work will be presented in the next chapter.

Conclusion and Future Work

In recent years, the video landscape on the internet has seen an enormous change along multiple dimensions. These dimensions include the capture and upload of high-resolution content and the ability to stream these high-resolution contents (4K/UHD-1 and above) owing to the increase in both devices that can display such content and average internet bandwidth and the corresponding streaming technologies, with HTTP-based adaptive streaming being the most dominant. This warrants investigation of the overall QoE of the end-user in the context of HTTP-based adaptive streaming, especially for high-resolution content. For such a quality assessment, two methods can be used. These include subjective quality assessment and instrumental methods using video quality models. This thesis addresses the problem of quality assessment using both aforementioned methods. For this, five different research questions were formulated with four of them focusing on instrumental methods of assessment and one on the subjective method. The four research questions for instrumental quality assessment focused on developing video quality models using bitstream information. The primary target of the developed models is traditional VoD 2D content for different application scenarios, depending on the available input information. Furthermore, additional use cases other than VoD 2D content for the developed models are investigated. The subjective assessment-related research question focused on analysing how quality assessment of high-resolution videos can be conducted in an out-of-the-lab setting.

Subjective testing is considered the gold standard for multimedia quality assessment. Hence, as a first method to assess the quality of high-resolution videos and also the overall quality of a HAS session, subjective testing was considered. The goal to use subjective testing was multi-fold. This included creating ground truth for

developing quality models and also comparing the developed models and existing state-of-the-art video quality models for short-term video quality prediction in the context of 4K/UHD-1.

Firstly, a lab-based approach was used to conduct subjective tests. For these tests, the ACR test paradigm and lab environment following the guidelines provided by ITU-R Rec. BT.500-13 were used. The subjective assessment studies were divided into two parts, with the first part focusing on the quality assessment of short-duration videos and the other on the overall integral quality assessment of a HAS session. The short-duration video quality assessment tests used videos of 4K/UHD-1 resolution of 7-10 s in duration with framerates up to 60 fps. Two parallel tracks were followed for conducting subjective tests for short-duration video quality. In the first track, four databases were created using 17 different source contents resulting in the publicly available AVT-VQDB-UHD-1 dataset [Rao+19a]. A parallel second track was part of a larger competition known as *P.NATS Phase 2* at ITU-T Study Group 12/Question 14. This involved the creation of databases for developing short-term video quality models as a successor to the ITU-T P.1203.1 video quality models. The result of this was a total of 26 different subjective tests conducted in collaboration with nine other proponent labs in the competition. Out of these 26 tests, four were conducted as part of the work presented in this thesis. The test conditions for all these tests were based on varying different encoding parameters within a defined range, which are used in a typical VoD scenario. Overall, eight different short-duration video quality tests were conducted, which also doubled up as ground truth for model development for automated video quality evaluation.

Following this, the focus was on assessing the overall quality of a HAS session in a typical lab environment. For this, six different overall integral assessment studies were designed as part of the same *P.NATS Phase 2* competition. These tests included videos of 1 *min* to 5 *min* duration with these videos subjected to typical HAS-specific degradations such as initial loading delay, stalling events, and quality switching. Out of these six tests, one was conducted by the author as part of this work. The overall integral assessment tests followed an “immersive” paradigm, in which the participants never view the same source stimulus more than once. All the data created using lab tests were then used as ground truth for model development.

Furthermore, as an alternative to lab-based testing, out-of-the-lab test methodology was also investigated for both short-term video quality assessment and overall quality of a HAS session. This was considered to facilitate subjective testing when lab-based testing is not feasible, e.g.: as it was the case in earlier phases of the COVID-19 pandemic. For this purpose, an approach using a pre-defined crop cut out from the centre of the video was developed and used for the quality assessment of high-resolution videos. This method is referred to as the “centre-crop” approach. Using the centre-crop approach, a test to assess short-duration video quality using the PVSs from one of the lab tests was conducted and the results showed good agreement of the out-of-the-lab test in comparison with the corresponding lab test, in terms of Pearson and Spearman correlations and the Standard deviation of Opinion Scores (SOS) factor [HSE11]. Similarly, an overall integral assessment study with the same centre-crop-based approach has been conducted using the PVSs from a corresponding lab test. The results indicate that the crowd test had good agreement as compared to the corresponding test. In addition, the AVT-VQDB-UHD-1 dataset along with the corresponding online test data have been made publicly available.

The second major focus of the thesis was the development of video quality models for quality evaluation of videos up to a resolution of 4K/UHD-1. The main model development activity of this doctoral work was conducted as part of the *P.NATS Phase 2* competition. As a result, three bitstream-based models were developed, namely Mode 3, Mode 1, and Mode 0, corresponding to three different modes of operation. All three models were submitted to the *P.NATS Phase 2* competition and were evaluated against other competing models. Following this comparison, it was determined that all three models developed by the author were either winning models or part of the winning group. The Mode 3 model developed as part of this thesis was a winning candidate in the *P.NATS Phase 2* competition and was subsequently standardized as ITU-T Rec. P.1204.3. The Mode 0 and Mode 1 models were part of the winning group, but there were no corresponding recommendations due to a lack of agreement between winning proponents to merge the winning candidates into one single model, as required by the rules of the competition. Furthermore, the Mode 3 model was the best among the 35 models spanning 10 different categories and was significantly better than all these models.

Chapter 7 Conclusion and Future Work

The proposed Mode 3 model consists of a traditional curve-fitting part and a Random Forest part. The final model is the ensemble of these two parts. Using this model as the basis, four other models, namely, Mode 0, Mode 1, and two variants of HYN0 models were developed. Both variants of the HYN0 model involve creation of a quality-equivalent bitstream (QEB) in the first step followed by the application of the Mode 3 model on the QEB for quality prediction. All these models together form the *AVQBits* model family. An extensive large-scale evaluation of these models was conducted and the results showed that these models outperform the SoA NR models for the VoD use case. Furthermore, from the results, it was observed that the *AVQBits|M3 / P.1204.3* and *AVQBits|H0* variants of *AVQBits* outperformed the SoA FR models, too.

The different variants of *AVQBits* were developed to be applied to different scopes, based on the available input information. Furthermore, all the developed models are made publicly available to the research community.

In addition to the development of the short-duration video quality models, a modification for the ITU-T P.1203.3 long-term integration model for overall integral quality evaluation of a HAS session was proposed as part of this thesis. This modification is added as an appendix to ITU-T P.1204.3. Using the different variants of *AVQBits* as the video quality prediction module, the proposed long-term model was evaluated using five different databases created as part of the *P.NATS Phase 2* competition, containing audiovisual sequences ranging from 1 *min* to 5 *min*. The results show that all the five *AVQBits* model variants can be used for this purpose with prediction accuracy dependent on the used model variant.

Following this, the *AVQBits* model variants were tested for their applicability to other use cases. For this, six different use cases, namely, gaming videos, 360° videos, HFR content, videos from a live streaming context, compressed videos created using videos with pre-existing distortions as sources, and images encoded with video codecs were considered. Using publicly available datasets for each of these use cases, the models developed as part of this work were evaluated. The results illustrated that the *AVQBits* model variants perform either on par with or better than the considered SoA models, without further modifications.

To summarize, all five research questions outlined in Chapter 1 have been addressed with the development of *AVQBits* model instances and their extensive evaluation and the development of the “centre-crop” approach for subjective testing in an out-of-the-lab setting, .

Although both subjective and instrumental quality assessments of high-resolution videos have been considered as part of the thesis, there are a lot of open problems still existing, especially with out-of-the-lab subjective testing methods, newer formats and newer streaming mechanisms. For out-of-the-lab subjective testing, the pre-defined centre crop approach can be extended by using more sophisticated methods to determine best-suited patches over the different video frames of a video, for example using saliency-based region-of-interest estimation or formal derivation with according eye-tracking data, if available. Furthermore, it can be investigated if such an approach can be extended to quality assessment of HDR videos.

In the model development context, a first simple extension of the work conducted as part of this thesis could be to extend the proposed models to 8K/UHD-2 videos and newer codecs such as AV1 or VVC. A corresponding bitstream parser for these newer codecs have to be developed. For model development, newer features that can better estimate the content complexity both in terms of spatial and temporal complexities can be considered. In addition to the existing motion based feature, macro-block-based features could be used for such content complexity estimation. In this regard, there is an ongoing activity at ITU-T SG12/Q14 to extend the ITU-T Rec. P.1204.3 to be applicable for AV1 encoded videos, with active contribution from the author. Within the domain of traditional 2D video, for better handling of the framerate variation, existing Mode 3 bitstream features can be adapted to be more framerate specific. One possible approach for this would be to scale the motion vectors in a more precise way to the actual speed of motion in pixels per time. This is expected to improve the specificity of the motion complexity information utilized in the Random Forest part of the Mode 3 model.

The proposed models can be optimized to different 2D streaming use cases such as live streaming. This would potentially involve adaptation of the existing Mode 3 features and addition of new features to take into account live-streaming-specific distortions such as frame drops etc. Furthermore, investigation on the quality assessment of high dynamic range (HDR) videos, both subjective and instrumental

can be considered. Also, the models developed as part of this thesis can be further used for perceptual encoding optimization of codecs.

In the domain of gaming video quality assessment, the focus can be on a more holistic assessment of QoE in an interactive gaming session and the corresponding model development. The corresponding model development can follow a modular approach as done on ITU-T Rec. P.1203 [ITU16b] and lead to the development of an integration model that takes into account the impact of factors such as delay on overall game play in addition to video quality. In addition to this, dedicated assessment of more immersive media such as 360° videos, light field images, etc. can be addressed. For 360° videos, modifications to motion-related features as well as other features that better handle the specific projection geometry can be considered for better performance of the models.

Another important area of focus for future work is the extension of the quality models for machine-learning and DNN-based codecs, and also enhancement codecs, e.g. LCEVC [MPE20]. This becomes important, as the degradations introduced by these codecs may differ from the degradations caused by traditional video codecs. Hence, both subjective and instrumental quality assessment is needed for these use cases. Also, different streaming mechanisms especially for newer formats are being developed, tested, and deployed, e.g. tiled-streaming for 360° videos, with P.1204.3 already being adapted for this use case [Koi+21]. Therefore, the models developed as part of this thesis can be extended and adapted for such newer streaming technologies.

Furthermore, different approaches to assessing the overall integral quality assessment of an HAS session can be devised and tested. This is important, since the preferences of the users both in terms of content and sensitivity to certain distortions change over time and hence dedicated methodologies that can take these preferences into account have to be developed. A corresponding activity referred to as *P.NATS Phase 3* is ongoing in ITU-T SG12/Q14, with the active participation of the author. Following the development of such new methodologies, models that can not only predict quality but also other aspects, e.g. “quitting probability” can be developed. Here, a user quitting a particular video is due to an annoyance related either to the encountered degradations or content itself.

In essence, there is ample scope to extend the work conducted as part of this doctoral thesis to different domains and application scopes, both in terms of subjective and instrumental quality assessment.

In the end, it's all about perception!!!

Subjective Test

The details of the protocol followed, instructions provided and the used post-test questionnaires are described in this appendix.

For short tests, no post-test questionnaire was used whereas for the long test, a post-test questionnaire was used to gather information on whether the participants were able to judge quality differences easily when longer sequences ($> 1 \text{ min}$) were used.

The protocol form and the instructions were provided in both English and German as the participants were mostly from the university comprising both German and international students.

Appendix A Subjective Test

Protocol- Project Name

AVT (date)

Testinformationen

Verantwortlich für den Test: Versuchsbetreuer:
Datum (JJJJ-MM-TT) Uhrzeit (hh:mm):
Testnummer: Probanden ID:

Personenbezogene Daten

Vorname: Name:
Alter: Geschlecht:
Sehhilfe benötigt?: nein ja
Falls ja, Art und Stärke:

Sehtest: Snellen-Index

Testaufbau: Sehprobentafel Snellen-Index (optimale Sehschärfe bei 8: 20/20 = 1 Winkelminute)
Korrekt erkannte Zeilen: 1 2 3 4 5 6 7 8 9 10 11

Ziel der Studie

Ziel dieser Studie ist es die subjektiv wahrnehmbare Qualität von Video zu testen. Die Ergebnisse helfen zu verstehen, wie bestimmte Videocodecs in Kombination mit verschiedenen Auflösungen und Bitraten die Qualitätswahrnehmung von UHD-Videostreaming beeinflussen.

Risiken

Dieser Versuch birgt keine speziellen gesundheitlichen Risiken. Allerdings können bei manchen Menschen sogenannte „photosensitive epileptische Anfälle“ auftreten, wenn sie bestimmten visuellen Reizen ausgesetzt sind. Falls während des Versuchs gesundheitliche Beschwerden oder Unwohlsein auftreten sollten (Schwindelgefühl, veränderte Wahrnehmung, Augen- oder Muskelzucken, Zittern an Armen oder Beinen, Desorientierung, Verwirrung), informieren Sie bitte unmittelbar den Versuchsleiter. Personen mit bekannter Epilepsie sollten nicht an diesem Test teilnehmen.

Datenschutz

Im Rahmen dieser Studie werden persönliche Daten erhoben. Zusätzlich werden die Antworten der mündlichen Fragen mit einem Smartphone aufgenommen. Diese Daten werden anonymisiert gespeichert und ausgewertet. Alle im Rahmen der Studie erhobenen Daten und Aufzeichnungen werden strikt vertraulich und gemäß dem Datenschutz behandelt. Bei einer Publikation oder Präsentation der Studienergebnisse werden nur anonymisierte Daten verwendet, sodass kein Rückschluss auf Ihre Person möglich ist.

Freiwillige Teilnahme

Die Teilnahme an dieser Studie ist rein freiwillig. Das Experiment kann durch Sie zu jeder Zeit ohne die Angabe von Gründen abgebrochen werden. Daraus entstehen Ihnen keinerlei Nachteile. Sie müssen keinerlei Fragen beantworten, die Sie nicht beantworten wollen.

Interne Teilnehmer Datenbank

Für folgende Tests, unabhängig vom aktuellen, würden wir Sie gern in unsere interne Teilnehmer-Datenbank aufnehmen. Eine solche Aufnahme beinhaltet, dass wir Sie im Falle weiterer Tests direkt kontaktieren können. Die Aufnahme ist vollkommen freiwillig und umfasst nur eine Kontakt-Email Adresse.

Aufnahme: ja nein
Kontakt-Email:

Einverständniserklärung

Der/die oben genannte Teilnehmer(-in) erklärt hiermit, dass er/sie die rechtlichen Hinweise zur Versehrtheit und zum Umgang mit seinen/ihren personenbezogenen Daten gelesen und verstanden hat und dass er/sie den gemachten Angaben zustimmt.

Vorname: Name:
Ort und Datum: Unterschrift:

Test information

Responsible for the test: Test supervisor:
Date (YYYY-MM-DD): Time (hh:mm):
Test number: Subject ID:

Personal data

First name: Name:
Age: Sex:
Visual aids used?: no yes
If yes, kind and strength:

Visual test: Snellen-Index

Test setup: Visual test panel Snellen-Index (optimal visual acuity at 8: 20/20 = 1 angular minutes)
Correct perceived lines: 1 2 3 4 5 6 7 8 9 10 11

Study objective

The objective of this study is to understand the perceived subjective video quality. The results will allow to understand how certain video codecs in combination with different resolutions and bit rates will influence the quality perception of UHD video streaming.

Risks

This test will not have an influence on your physical health. However, for some people it can lead to an epileptic seizure if they are confronted with certain visual stimulus. If during the test under any circumstances, if symptoms like: dizziness, odd perception, eye or muscle twitches, shivering arms or legs, disorientation or confusion etc. appear, please inform your supervisor immediately. People with known epileptic seizure attacks are not allowed to take part in this test.

Privacy

In the scope of this study we will record personal data. In addition, the responses to the oral questions will be recorded via a smartphone. This data will be anonymized, saved and evaluated. All recorded personal data in this study will be treated confidentially (according to German data privacy law). In case of publication or presentation of the results we will only use anonymized data, so that conclusion to any participant is impossible.

Voluntary participant

This is a voluntary study. The test can be aborted at anytime without reasons. There will not be any disadvantage on your side. If you feel uncomfortable answering certain questions you do not have to answer.

Internal Database for Participants

For future tests, independent of the current test, we would like to include you in our internal database of participants. Based on this inclusion we would contact you in the case of new tests. It is completely voluntary and we will only collect your contact email address.

Inclusion: yes no
Contact-email:

Declaration of agreement

The named participant agrees to the above stated points and has read them carefully. He/She is informed and has understood how his/her personal data will be treated and used. A signature states an agreement to all mentioned points.

First name: Name:
Place and date: Signature:

Appendix A Subjective Test

Erläuterungen zum Testablauf

Vielen Dank für Ihre Teilnahme am Versuch!

Sehtest und Fragebogen

Bitte füllen Sie den vorliegenden Fragebogen und das Formular für die Aufwandsentschädigung aus. Lesen Sie dabei sehr sorgfältig die Erklärungen.

Anschließend wird ein Sehtest durchgeführt, um Ihre Sehfähigkeit zu überprüfen. Dabei können Sie eventuelle Sehhilfen (Brille oder Kontaktlinsen) tragen, sofern Sie sie dann auch im anschließenden Test tragen werden.

Test

Im Folgenden werden einige Erläuterungen zu dem Testablauf umrissen. Dabei ist der Test im Allgemeinen in zwei Phasen (Training- und Testphase) geteilt, die jeweils vom Ablauf her identisch sind.

Zunächst wird ein Video abgespielt. Anschließend werden Sie als TeilnehmerIn aufgefordert, eine Bewertung über die Videoqualität durchzuführen.

Trainingsphase

In der Trainingsphase werden erst 6 Videosequenzen mit unterschiedlichen Qualitäten und Inhalten gezeigt. Ziel dieser Phase ist es, dass Sie als TeilnehmerIn einen ersten Eindruck über die verschiedenen Inhalte und Qualitäten erhalten. Die Bewertungen aus dieser Phase werden daher nicht in die Gesamtbewertungen aufgenommen.

Testphase

In der Testphase werden Ihnen 187 Videosequenzen gezeigt, dabei variieren der Inhalt und die Videoqualität. Während des Tests bitten wir Sie, zwei Pausen (maximal jeweils 5 Minuten) durchzuführen.

Bewertung der Videoqualität

Die Bewertung erfolgt auf einer Skala von 1 bis 5, wobei 1 der schlechtesten Qualität ("Bad") und 5 der besten wahrgenommenen Qualität ("Excellent") entspricht.

What is your opinion of the video quality? User ID: 1

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

© 2017 - Maxime Schaefer, Steve Orlitzky

Abbildung 1: Bewertung

Bitte bewerten Sie dabei intuitiv, ohne viel nachzudenken. Bitte beachten Sie auch, dass es hierbei um die Qualität des Videos geht und nicht darum, wie sehr Ihnen der Inhalt gefällt.

Viel Spaß

Illustration of test procedure

Thank you for participating in this test.

Eyesight test and questionnaire

Please fill out the handed questionnaire completely. Please also fill out the form of the „Aufwandsentschädigung“ (allowance). Please read the questions and explanations carefully so that the tasks are clear. If not, do not hesitate to ask questions to the supervisor.

An Eye sight test will then be conducted to check your vision.

Test

A description of the test procedure follows below. The test is divided into 2 parts (training- and test-phase) whereas the procedure for both are same. A video will be played and the participant will be asked to rate the quality of the video.

Training Phase

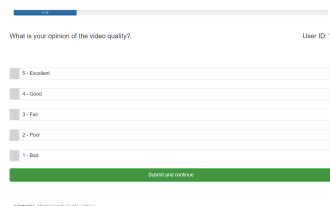
In this part, 6 different video sequences with changing qualities and content will be shown. The goal is to enable the participant to understand what differences in qualities and content exist in the test. The ratings of this part will not be included in the final results.

Test Phase

In this part, 187 different videos will be shown with changing quality and content according to different video codecs. During the test you are allowed to take 2 breaks with max 5 min each.

Rating of video quality

The rating scale is from 1 to 5, where 1 is the worst (“Bad”) and 5 the best quality (“Excellent”) you perceive.



What is your opinion of the video quality? User ID: 1

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

Submit and continue

© 2017, Matthias Schall, RWTH Aachen

Figure 1: Rating Screen

Please do your rating intuitively, without thinking a lot. Please consider also, that your rating is based on the video quality and not how much you liked the content.

Have fun!

Appendix A Subjective Test

Erläuterungen zum Testablauf

Vielen Dank für Ihre Teilnahme am Versuch!

Sehtest und Fragebogen

Bitte füllen Sie den vorliegenden Fragebogen und das Formular für die Aufwandsentschädigung aus. Lesen Sie dabei sehr sorgfältig die Erklärungen.

Anschließend wird ein Sehtest durchgeführt, um Ihre Sehfähigkeit zu überprüfen. Dabei können Sie eventuelle Sehhilfen (Brille oder Kontaktlinsen) tragen, sofern Sie sie dann auch im anschließenden Test tragen werden.

Test

Im Folgenden werden einige Erläuterungen zu dem Testablauf umrissen. Dabei ist der Test im Allgemeinen in zwei Phasen (Training- und Testphase) geteilt, die jeweils vom Ablauf her identisch sind.

Zunächst wird ein Video abgespielt. Anschließend werden Sie als TeilnehmerIn aufgefordert, eine Bewertung über die Videoqualität durchzuführen.

Trainingsphase

In der Trainingsphase werden erst 2 Videosequenzen mit unterschiedlichen Qualitäten und Inhalten gezeigt. Ziel dieser Phase ist es, dass Sie als TeilnehmerIn einen ersten Eindruck über die verschiedenen Inhalte und Qualitäten erhalten. Die Bewertungen aus dieser Phase werden daher nicht in die Gesamtbewertungen aufgenommen.

Testphase

In der Testphase werden Ihnen 30 Videosequenzen gezeigt, dabei variieren der Inhalt und die Videoqualität.

Es gibt eine 5-10 minütige Pause nach jeweils 25 Minuten während des Tests.

Bewertung der Videoqualität

Die Bewertung erfolgt auf einer Skala von 1 bis 5, wobei 1 der schlechtesten Qualität ("Bad") und 5 der besten wahrgenommenen Qualität ("Excellent") entspricht.

What is your opinion of the video quality? User ID: 1

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

Submit and save data

© 2017 - Maxime Schaefer, Steve Orlitzky

Abbildung 1: Bewertung

Bitte bewerten Sie dabei intuitiv, ohne viel nachzudenken. Bitte beachten Sie auch, dass es hierbei um die Qualität des Videos geht und nicht darum, wie sehr Ihnen der Inhalt gefällt.

Viel Spaß

Post-test questionnaire

Thank you for participating in this test.

Eyesight test and questionnaire

Please fill out the handed questionnaire completely. Please also fill out the form of the „Aufwandsentschädigung“ (allowance). Please read the questions and explanations carefully so that the tasks are clear. If not, do not hesitate to ask questions to the supervisor.

An Eye sight test will then be conducted to check your vision.

Test

A description of the test procedure follows below. The test is divided into 2 parts (training- and test-phase) whereas the procedure for both are same. A video will be played and the participant will be asked to rate the quality of the video.

Training Phase

In this part, 2 different video sequences with changing qualities and content will be shown. The goal is to enable the participant to understand what differences in qualities and content exist in the test. The ratings of this part will not be included in the final results.

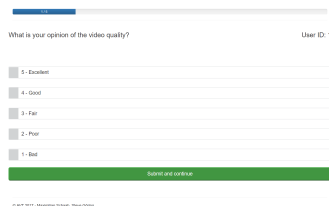
Test Phase

In this part, 30 different videos will be shown with changing quality and content according to different video codecs.

There will be a 5-10 break after 25 minutes during the test.

Rating of video quality

The rating scale is from 1 to 5, where 1 is the worst (“Bad”) and 5 the best quality (“Excellent”) you perceive.



What is your opinion of the video quality? User ID: 1

5 - Excellent

4 - Good

3 - Fair

2 - Poor

1 - Bad

Submit and continue

© 2017 Matthias Schall, RWTH Aachen

Figure 1: Rating Screen

Please do your rating intuitively, without thinking a lot. Please consider also, that your rating is based on the video quality and not how much you liked the content.

Have fun!

Post-test Questionnaire

Name : _____

Questions:

- 1) What is your subject background (technical/non-technical)? Please specify the subject area if you are a student.

- 2) Were the breaks helpful? Please mention the reason for your answer.

- 3) Was the test too long?

- 4) Were you able to judge the quality difference easily? If no, please let us know why you were not able to judge the quality differences easily.

- 5) Was the test room environment comfortable (too hot/too cold/too stuffy etc)?

- 6) Any other suggestions to improve the test.

Date : _____

Signature : _____

P.NATS Phase 2 Test Plan

The details of the test design of the 26 databases created as part of the *P.NATS Phase 2* competition is provided in this appendix. Four out of the 26 databases, namely, P2STR09, P2STR10, P2SVL01, and P2SVL09 form the AVT-PNATS-UHD-1 dataset described in Chapter 3.

The following abbreviations are used in the test plans:

- ▷ HRC: Hypothetical Reference Circuit, this refers to the test condition
- ▷ Passes: number of encoding passes used
- ▷ Preset: in the case of H.264 and H.265 it refers to the used “preset” parameter value for encoding and in the case of VP9, it refers to the “speed” parameter value
- ▷ Bitrate: target video bitrate in kbps
- ▷ Height: encoding video height
- ▷ Width: encoding video width
- ▷ CRF: Constant Rate Factor
- ▷ iFI: iFrameInterval (I-Frame interval in seconds)
- ▷ MFR: Maximum Factor Bitrate (specifies the maximum bitrate w.r.t the target bitrate to be used for encoding)
- ▷ Framerate is expressed in fps

Appendix B P.NATS Phase 2 Test Plan

Table B.1: Test Plan for P2STR01.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0011	H.265	2	slower	240.0	426.0	95/190	None	15	2	2.0
HRC0012	VP9	2	0	240.0	426.0	95/190	None	15	2	2.0
HRC0021	VP9	1	3	240.0	426.0	300/375	None	15	2	2.0
HRC0029	H.265	1	veryfast	360.0	640.0	225/375	None	15	2	2.0
HRC0087	VP9	2	2	240.0	426.0	75/150	None	24/25/30	5	2.0
HRC0090	VP9	1	3	240.0	426.0	190/300	None	24/25/30	2	2.0
HRC0099	VP9	2	0	240.0	426.0	190/300	None	24/25/30	2	2.0
HRC0109	H.264	1	fast	240.0	426.0	650/800	None	24/25/30	2	2.0
HRC0110	H.265	1	ultrafast	240.0	426.0	490/600	None	24/25/30	2	2.0
HRC0111	VP9	1	3	240.0	426.0	490/600	None	24/25/30	2	2.0
HRC0112	H.264	2	medium	240.0	426.0	650/800	None	24/25/30	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0120	VP9	1	3	360.0	640.0	300/450	None	24/25/30	2	2.0
HRC0124	H.264	2	medium	360.0	640.0	400/600	None	24/25/30	5	2.0
HRC0169	H.264	2	medium	480.0	854.0	1100/1400	None	24/25/30	5	2.0
HRC0180	VP9	2	2	480.0	854.0	1050/1500	None	24/25/30	5	2.0
HRC0193	H.264	2	medium	720.0	1280.0	500/750	None	24/25/30	5	2.0
HRC0200	H.265	2	veryslow	720.0	1280.0	375/565	None	24/25/30	2	2.0
HRC0221	H.265	2	medium	720.0	1280.0	1500/2250	None	24/25/30	5	1.1
HRC0230	H.265	2	medium	720.0	1280.0	2400/3000	None	24/25/30	2	2.0
HRC0245	H.265	2	medium	1080.0	1920.0	750/1500	None	24/25/30	5	2.0
HRC0249	VP9	2	2	1080.0	1920.0	750/1500	None	24/25/30	5	1.1
HRC0252	VP9	2	2	1080.0	1920.0	2250/4500	None	24/25/30	5	2.0
HRC0265	H.264	2	medium	1080.0	1920.0	7000/9500	None	24/25/30	2	1.1
HRC0272	H.265	2	medium	1080.0	1920.0	5250/7125	None	24/25/30	5	1.1
HRC0273	VP9	2	2	1080.0	1920.0	375/565	None	24/25/30	5	1.1
HRC0312	VP9	2	2	1440.0	2560.0	3000/6000	None	24/25/30	5	2.0
HRC0337	H.264	2	medium	1440.0	2560.0	15000/20000	None	24/25/30	2	2.0
HRC0386	H.265	1	veryfast	720.0	1280.0	375/565	None	50/60	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0402	VP9	2	2	720.0	1280.0	825/1200	None	50/60	5	2.0
HRC0423	VP9	2	0	720.0	1280.0	2400/3000	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0437	H.265	2	medium	1080.0	1920.0	2625/5250	None	50/60	2	2.0
HRC0438	VP9	2	2	1080.0	1920.0	2625/5250	None	50/60	2	2.0
HRC0450	VP9	1	3	1080.0	1920.0	5250/7125	None	50/60	2	2.0
HRC0452	H.265	2	medium	1080.0	1920.0	5250/7125	None	50/60	2	2.0
HRC0458	H.265	2	medium	1080.0	1920.0	5250/7125	None	50/60	5	2.0
HRC0477	VP9	2	2	1440.0	2560.0	750/1500	None	50/60	5	1.1
HRC0704	H.264	2	medium	1440.0	2560.0	1000/1750	None	50/60	2	2.0
HRC0491	H.265	2	medium	1440.0	2560.0	750/1500	None	50/60	5	2.0
HRC0498	VP9	1	4	1440.0	2560.0	7500/11250	None	50/60	2	2.0
HRC0502	H.264	2	medium	1440.0	2560.0	10000/15000	None	50/60	2	1.1
HRC0521	H.265	2	medium	1440.0	2560.0	11250/15000	None	50/60	2	1.1
HRC0522	VP9	2	2	1440.0	2560.0	11250/15000	None	50/60	2	1.1

Table B.2: Test Plan for P2STR02.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0002	H.265	2	medium	240.0	426.0	60/115	None	15	5	2.0
HRC0006	VP9	1	3	240.0	426.0	95/190	None	15	2	2.0
HRC0007	H.264	2	medium	240.0	426.0	125/250	None	15	2	2.0
HRC0014	H.265	1	fast	360.0	640.0	190/265	None	24/25/30	2	2.0
HRC0015	VP9	1	3	360.0	640.0	190/265	None	24/25/30	2	2.0
HRC0018	VP9	2	2	360.0	640.0	190/265	None	24/25/30	2	2.0
HRC0022	H.264	2	medium	240.0	426.0	400/500	None	15	2	2.0
HRC0031	H.264	2	medium	360.0	640.0	100/100	None	15	2	2.0
HRC0048	VP9	2	2	360.0	640.0	600/750	None	15	2	2.0
HRC0053	H.265	2	medium	480.0	854.0	190/375	None	24/25/30	5	1.1
HRC0064	H.264	2	veryslow	480.0	854.0	500/800	None	24/25/30	2	2.0
HRC0070	H.264	2	medium	480.0	854.0	900/1200	None	24/25/30	2	2.0
HRC0077	H.265	1	fast	480.0	854.0	900/1050	None	24/25/30	2	2.0
HRC0085	H.264	2	medium	240.0	426.0	100/200	None	24/25/30	5	2.0
HRC0089	H.265	1	ultrafast	240.0	426.0	190/300	None	24/25/30	2	2.0
HRC0102	VP9	1	4	240.0	426.0	300/450	None	24/25/30	2	2.0
HRC0103	H.264	2	medium	240.0	426.0	400/600	None	24/25/30	2	2.0
HRC0106	H.264	2	medium	240.0	426.0	400/600	None	24/25/30	5	2.0
HRC0107	H.265	2	medium	240.0	426.0	300/450	None	24/25/30	5	2.0
HRC0108	VP9	2	2	240.0	426.0	95/150	None	24/25/30	5	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0123	VP9	2	2	360.0	640.0	300/450	None	24/25/30	2	2.0
HRC0148	H.264	2	medium	480.0	854.0	200/250	None	24/25/30	5	1.1
HRC0149	H.265	2	medium	480.0	854.0	190/200	None	24/25/30	5	1.1
HRC0154	H.264	2	medium	480.0	854.0	700/1000	None	24/25/30	5	2.0
HRC0167	H.265	2	medium	480.0	854.0	825/1050	None	24/25/30	2	1.1
HRC0197	H.265	2	medium	720.0	1280.0	375/565	None	24/25/30	5	1.1
HRC0212	H.265	2	medium	720.0	1280.0	1500/2250	None	24/25/30	2	2.0
HRC0218	H.265	2	medium	720.0	1280.0	1500/2250	None	24/25/30	5	2.0
HRC0232	H.264	2	slow	720.0	1280.0	3200/4000	None	24/25/30	2	2.0
HRC0258	VP9	2	0	1080.0	1920.0	2250/4500	None	24/25/30	2	2.0
HRC0263	H.265	2	medium	1080.0	1920.0	5250/7125	None	24/25/30	2	2.0
HRC0292	H.264	1	ultrafast	1440.0	2560.0	1500/3000	None	24/25/30	2	2.0
HRC0297	VP9	2	2	1440.0	2560.0	1125/2250	None	24/25/30	2	2.0
HRC0309	VP9	2	0	1440.0	2560.0	1125/2250	None	24/25/30	2	2.0
HRC0336	VP9	1	3	1440.0	2560.0	11250/15000	None	24/25/30	2	2.0
HRC0385	H.264	1	ultrafast	720.0	1280.0	500/750	None	50/60	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0420	VP9	2	2	720.0	1280.0	2400/3000	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0446	H.265	2	medium	1080.0	1920.0	2625/5250	None	50/60	5	1.1
HRC0480	VP9	2	1	1440.0	2560.0	300/500	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0488	H.265	2	medium	1440.0	2560.0	500/500	None	50/60	2	1.1
HRC0492	VP9	2	2	1440.0	2560.0	3000/5250	None	50/60	5	2.0
HRC0497	H.265	1	ultrafast	1440.0	2560.0	7500/11250	None	50/60	2	2.0
HRC0500	H.265	2	medium	1440.0	2560.0	7500/11250	None	50/60	2	2.0
HRC0509	H.265	2	medium	1440.0	2560.0	7500/11250	None	50/60	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.3: Test Plan for P2STR03.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0003	VP9	2	2	240.0	426.0	60/115	None	15	5	2.0
HRC0005	H.265	1	veryfast	240.0	426.0	95/190	None	15	2	2.0
HRC0016	H.264	2	medium	240.0	426.0	250/350	None	15	2	2.0
HRC0019	H.264	1	veryfast	240.0	426.0	400/500	None	15	2	2.0
HRC0020	H.265	1	veryfast	240.0	426.0	300/375	None	15	2	2.0
HRC0023	H.265	2	medium	240.0	426.0	300/375	None	15	2	2.0
HRC0033	VP9	2	2	360.0	640.0	225/375	None	15	2	2.0
HRC0037	H.264	1	ultrafast	360.0	640.0	600/800	None	15	2	2.0
HRC0039	VP9	1	3	360.0	640.0	450/600	None	15	2	2.0
HRC0086	H.265	2	medium	240.0	426.0	75/150	None	24/25/30	5	2.0
HRC0088	H.264	1	ultrafast	240.0	426.0	250/400	None	24/25/30	2	2.0
HRC0096	VP9	2	2	240.0	426.0	190/300	None	24/25/30	5	2.0
HRC0098	H.265	2	slower	240.0	426.0	190/300	None	24/25/30	2	2.0
HRC0100	H.264	1	ultrafast	240.0	426.0	400/600	None	24/25/30	2	2.0
HRC0101	H.265	1	ultrafast	240.0	426.0	300/450	None	24/25/30	2	2.0
HRC0104	H.265	2	medium	240.0	426.0	300/450	None	24/25/30	2	2.0
HRC0105	VP9	2	2	240.0	426.0	300/450	None	24/25/30	2	2.0
HRC0114	VP9	2	2	240.0	426.0	490/600	None	24/25/30	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0125	H.265	2	medium	360.0	640.0	300/450	None	24/25/30	5	2.0
HRC0127	H.264	2	slow	360.0	640.0	400/600	None	24/25/30	2	2.0
HRC0136	H.264	2	medium	360.0	640.0	800/1000	None	24/25/30	5	2.0
HRC0156	VP9	2	2	480.0	854.0	525/750	None	24/25/30	5	2.0
HRC0183	VP9	2	2	480.0	854.0	1050/1500	None	24/25/30	5	1.1
HRC0189	VP9	2	2	720.0	1280.0	375/565	None	24/25/30	2	2.0
HRC0194	H.265	2	medium	720.0	1280.0	375/565	None	24/25/30	5	2.0
HRC0217	H.264	2	medium	720.0	1280.0	2000/3000	None	24/25/30	5	2.0
HRC0222	VP9	2	2	720.0	1280.0	1500/2250	None	24/25/30	5	1.1
HRC0236	H.265	1	veryfast	1080.0	1920.0	750/1500	None	24/25/30	2	2.0
HRC0276	VP9	1	3	1080.0	1920.0	7125/9000	None	24/25/30	2	2.0
HRC0305	H.265	2	medium	1440.0	2560.0	1125/2250	None	24/25/30	5	1.1
HRC0316	H.264	1	fast	1440.0	2560.0	10000/15000	None	24/25/30	2	2.0
HRC0318	VP9	1	3	1440.0	2560.0	7500/11250	None	24/25/30	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0390	VP9	2	2	720.0	1280.0	375/565	None	50/60	2	2.0
HRC0426	VP9	2	2	1080.0	1920.0	825/1200	None	50/60	5	2.0
HRC0427	H.264	2	medium	1080.0	1920.0	1100/1600	None	50/60	5	1.1
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0448	H.264	1	veryfast	1080.0	1920.0	7000/9500	None	50/60	2	2.0
HRC0460	H.264	2	medium	1080.0	1920.0	7000/9500	None	50/60	5	1.1
HRC0463	H.264	2	slower	1080.0	1920.0	7000/9500	None	50/60	2	2.0
HRC0470	H.265	2	medium	1080.0	1920.0	7125/9000	None	50/60	5	1.1
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0496	H.264	1	ultrafast	1440.0	2560.0	10000/15000	None	50/60	2	2.0
HRC0505	H.264	2	medium	1440.0	2560.0	10000/15000	None	50/60	5	2.0
HRC0528	VP9	2	2	1440.0	2560.0	11250/15000	None	50/60	5	1.1

Table B.4: Test Plan for P2STR04.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0004	H.264	1	ultrafast	240.0	426.0	125/250	None	15	2	2.0
HRC0008	H.265	2	medium	240.0	426.0	95/190	None	15	2	2.0
HRC0009	VP9	2	2	240.0	426.0	95/190	None	15	2	2.0
HRC0010	H.264	2	veryslow	240.0	426.0	125/250	None	15	2	2.0
HRC0013	H.264	1	fast	360.0	640.0	250/350	None	24/25/30	2	2.0
HRC0017	H.265	2	medium	240.0	426.0	190/265	None	15	2	2.0
HRC0024	VP9	2	2	360.0	640.0	300/375	None	24/25/30	2	2.0
HRC0060	VP9	2	2	480.0	854.0	375/600	None	24/25/30	2	2.0
HRC0065	H.265	2	slower	480.0	854.0	375/600	None	15	2	2.0
HRC0068	H.265	1	veryfast	480.0	854.0	675/900	None	24/25/30	2	2.0
HRC0076	H.264	1	veryfast	480.0	854.0	1200/1400	None	15	2	2.0
HRC0091	H.264	2	medium	240.0	426.0	250/400	None	24/25/30	2	2.0
HRC0092	H.265	2	medium	240.0	426.0	190/300	None	24/25/30	2	2.0
HRC0093	VP9	2	2	240.0	426.0	190/300	None	24/25/30	2	2.0
HRC0094	H.264	2	medium	240.0	426.0	250/400	None	24/25/30	5	2.0
HRC0095	H.265	2	medium	240.0	426.0	190/300	None	24/25/30	5	2.0
HRC0097	H.264	2	slow	240.0	426.0	250/400	None	24/25/30	2	2.0
HRC0113	H.265	2	medium	240.0	426.0	490/600	None	24/25/30	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0143	H.265	2	medium	360.0	640.0	750/900	None	24/25/30	2	2.0
HRC0144	VP9	2	2	360.0	640.0	750/900	None	24/25/30	2	2.0
HRC0147	VP9	2	2	480.0	854.0	265/525	None	24/25/30	5	2.0
HRC0157	H.264	2	medium	480.0	854.0	700/1000	None	24/25/30	5	1.1
HRC0166	H.264	2	medium	480.0	854.0	1100/1400	None	24/25/30	2	1.1
HRC0196	H.264	2	medium	720.0	1280.0	500/750	None	24/25/30	5	1.1
HRC0198	VP9	2	2	720.0	1280.0	375/565	None	24/25/30	5	1.1
HRC0250	H.264	2	medium	1080.0	1920.0	3000/6000	None	24/25/30	5	2.0
HRC0262	H.264	2	medium	1080.0	1920.0	7000/9500	None	24/25/30	2	2.0
HRC0266	H.265	2	medium	1080.0	1920.0	5250/7125	None	24/25/30	2	1.1
HRC0269	H.265	2	medium	1080.0	1920.0	5250/7125	None	24/25/30	5	2.0
HRC0287	H.265	2	medium	1080.0	1920.0	7125/9000	None	24/25/30	5	1.1
HRC0290	H.265	2	slower	1080.0	1920.0	7125/9000	None	24/25/30	2	2.0
HRC0295	H.264	2	medium	1440.0	2560.0	1500/3000	None	24/25/30	2	2.0
HRC0300	VP9	2	2	1440.0	2560.0	1125/2250	None	24/25/30	2	1.1
HRC0319	H.264	2	medium	1440.0	2560.0	10000/15000	None	24/25/30	2	2.0
HRC0321	VP9	2	2	1440.0	2560.0	7500/11250	None	24/25/30	2	2.0
HRC0330	VP9	2	2	1440.0	2560.0	7500/11250	None	24/25/30	5	1.1
HRC0334	H.264	1	ultrafast	1440.0	2560.0	15000/20000	None	24/25/30	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0401	H.265	2	medium	720.0	1280.0	825/1200	None	50/60	5	2.0
HRC0407	H.265	2	veryslow	720.0	1280.0	825/1200	None	50/60	2	2.0
HRC0410	H.265	2	medium	720.0	1280.0	1800/2400	None	50/60	5	2.0
HRC0412	H.264	2	medium	720.0	1280.0	2400/3200	None	50/60	5	1.1
HRC0431	H.265	2	veryslow	1080.0	1920.0	825/1200	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0439	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	1.1
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0489	VP9	2	2	1440.0	2560.0	3000/5250	None	50/60	2	1.1
HRC0494	H.265	2	medium	1440.0	2560.0	3000/5250	None	50/60	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.5: Test Plan for P2STR05.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0028	H.264	1	veryfast	360.0	640.0	300/500	None	15	2	2.0
HRC0032	H.265	2	medium	360.0	640.0	225/375	None	15	2	2.0
HRC0044	H.265	1	ultrafast	360.0	640.0	600/750	None	15	2	2.0
HRC0083	H.265	2	slow	480.0	854.0	900/1050	None	15	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0134	H.265	2	medium	360.0	640.0	600/750	None	24/25/30	2	2.0
HRC0135	VP9	2	2	360.0	640.0	600/750	None	24/25/30	2	2.0
HRC0138	VP9	2	2	360.0	640.0	600/750	None	24/25/30	5	2.0
HRC0163	H.264	2	medium	480.0	854.0	1100/1400	None	24/25/30	2	2.0
HRC0175	H.264	2	slower	480.0	854.0	1100/1400	None	24/25/30	2	2.0
HRC0178	H.264	2	medium	480.0	854.0	1400/2000	None	24/25/30	5	2.0
HRC0184	H.264	1	ultrafast	720.0	1280.0	500/750	None	24/25/30	2	2.0
HRC0187	H.264	2	medium	720.0	1280.0	500/750	None	24/25/30	2	2.0
HRC0195	VP9	2	2	720.0	1280.0	375/565	None	24/25/30	5	2.0
HRC0206	H.265	2	medium	720.0	1280.0	750/1125	None	24/25/30	5	1.1
HRC0229	H.264	2	medium	720.0	1280.0	500/750	None	24/25/30	2	2.0
HRC0235	H.264	1	fast	1080.0	1920.0	500/750	None	24/25/30	2	2.0
HRC0256	H.264	2	slow	1080.0	1920.0	3000/6000	None	24/25/30	2	2.0
HRC0278	H.265	2	medium	1080.0	1920.0	1000/2000	None	24/25/30	2	2.0
HRC0304	H.264	2	medium	1440.0	2560.0	1500/3000	None	24/25/30	5	1.1
HRC0310	H.264	2	medium	1440.0	2560.0	3000/6000	None	24/25/30	5	2.0
HRC0326	H.265	2	medium	1440.0	2560.0	3000/6000	None	24/25/30	5	2.0
HRC0349	H.264	2	medium	2160.0	3840.0	6000/12000	None	24/25/30	5	2.0
HRC0353	H.265	2	medium	2160.0	3840.0	4500/9000	None	24/25/30	5	1.1
HRC0360	VP9	2	2	2160.0	3840.0	4500/9000	None	24/25/30	5	1.1
HRC0370	H.264	2	medium	2160.0	3840.0	11250/16500	None	24/25/30	2	1.1
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0400	H.264	2	medium	720.0	1280.0	1100/1600	None	50/60	5	2.0
HRC0404	H.265	2	medium	720.0	1280.0	825/1200	None	50/60	5	1.1
HRC0413	H.265	2	medium	720.0	1280.0	1800/2400	None	50/60	5	1.1
HRC0422	H.265	2	slower	720.0	1280.0	825/1200	None	50/60	2	2.0
HRC0702	H.264	2	medium	1080.0	1920.0	1800/2400	None	50/60	2	2.0
HRC0456	VP9	2	2	1080.0	1920.0	3500/7000	None	50/60	2	1.1
HRC0462	VP9	2	2	1080.0	1920.0	3500/7000	None	50/60	5	1.1
HRC0466	H.264	2	medium	1080.0	1920.0	9500/12000	None	50/60	5	2.0
HRC0473	H.265	2	medium	1440.0	2560.0	1315/2625	None	50/60	5	2.0
HRC0483	VP9	1	4	1440.0	2560.0	3000/5250	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0541	H.264	2	medium	2160.0	3840.0	6000/10000	None	50/60	2	1.1
HRC0544	H.264	2	medium	2160.0	3840.0	6000/10000	None	50/60	5	2.0
HRC0557	H.265	2	medium	2160.0	3840.0	6000/10000	None	50/60	2	2.0
HRC0563	H.265	2	medium	2160.0	3840.0	16500/22500	None	50/60	5	2.0
HRC0564	VP9	2	2	2160.0	3840.0	16500/22500	None	50/60	5	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0

Table B.6: Test Plan for P2STR06.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0034	H.264	2	veryslow	360.0	640.0	100/200	None	15	2	2.0
HRC0045	VP9	1	4	360.0	640.0	100/200	None	15	2	2.0
HRC0049	H.264	2	medium	480.0	854.0	150/350	None	15	5	2.0
HRC0050	H.265	2	medium	480.0	854.0	140/325	None	15	5	2.0
HRC0054	VP9	2	2	480.0	854.0	140/325	None	15	5	1.1
HRC0055	H.264	1	veryfast	480.0	854.0	300/650	None	15	2	2.0
HRC0084	VP9	2	0	480.0	854.0	650/800	None	15	2	2.0
HRC0700	H.264	2	medium	360.0	640.0	100/200	None	24/25/30	5	2.0
HRC0121	H.264	2	medium	360.0	640.0	250/450	None	24/25/30	2	2.0
HRC0132	VP9	1	4	360.0	640.0	350/550	None	24/25/30	2	2.0
HRC0141	VP9	1	3	360.0	640.0	350/550	None	24/25/30	2	2.0
HRC0160	H.264	1	fast	480.0	854.0	1100/1400	None	24/25/30	2	2.0
HRC0161	H.265	1	veryfast	480.0	854.0	225/350	None	24/25/30	2	2.0
HRC0171	VP9	2	2	480.0	854.0	225/350	None	24/25/30	5	2.0
HRC0186	VP9	1	4	720.0	1280.0	225/415	None	24/25/30	2	2.0
HRC0191	H.265	2	medium	720.0	1280.0	225/415	None	24/25/30	2	1.1
HRC0213	VP9	2	2	720.0	1280.0	1500/2250	None	24/25/30	2	2.0
HRC0233	H.265	2	slower	720.0	1280.0	2400/3000	None	24/25/30	2	2.0
HRC0261	VP9	1	4	1080.0	1920.0	5250/7125	None	24/25/30	2	2.0
HRC0281	H.265	2	medium	1080.0	1920.0	7125/9000	None	24/25/30	2	1.1
HRC0285	VP9	2	2	1080.0	1920.0	7125/9000	None	24/25/30	5	2.0
HRC0294	VP9	1	4	1440.0	2560.0	1125/2250	None	24/25/30	2	2.0
HRC0298	H.264	2	medium	1440.0	2560.0	1500/3000	None	24/25/30	2	1.1
HRC0350	H.265	2	medium	2160.0	3840.0	4500/9000	None	24/25/30	5	2.0
HRC0355	H.264	2	medium	2160.0	3840.0	15000/22000	None	24/25/30	5	2.0
HRC0365	H.265	1	ultrafast	2160.0	3840.0	16500/22500	None	24/25/30	2	2.0
HRC0368	H.265	2	medium	2160.0	3840.0	16500/22500	None	24/25/30	2	2.0
HRC0372	VP9	2	2	2160.0	3840.0	16500/22500	None	24/25/30	2	1.1
HRC0373	H.264	2	medium	2160.0	3840.0	22000/30000	None	24/25/30	5	2.0
HRC0701	H.264	2	medium	720.0	1280.0	400/900	None	50/60	2	2.0
HRC0418	H.264	2	medium	720.0	1280.0	3200/4000	None	50/60	2	2.0
HRC0429	VP9	2	2	1080.0	1920.0	400/800	None	50/60	5	1.1
HRC0430	H.264	2	slower	1080.0	1920.0	1100/1600	None	50/60	2	2.0
HRC0434	H.265	1	fast	1080.0	1920.0	2625/5250	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0457	H.264	2	medium	1080.0	1920.0	7000/9500	None	50/60	5	2.0
HRC0459	VP9	2	2	1080.0	1920.0	5250/7125	None	50/60	5	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0490	H.264	2	medium	1440.0	2560.0	4000/7000	None	50/60	5	2.0
HRC0510	VP9	2	2	1440.0	2560.0	7500/11250	None	50/60	5	1.1
HRC0525	VP9	2	2	1440.0	2560.0	11250/15000	None	50/60	5	2.0
HRC0532	H.264	2	medium	2160.0	3840.0	6000/12000	None	50/60	5	1.1
HRC0555	VP9	1	3	2160.0	3840.0	16500/22500	None	50/60	2	2.0
HRC0558	VP9	2	2	2160.0	3840.0	16500/22500	None	50/60	2	2.0
HRC0561	VP9	2	2	2160.0	3840.0	16500/22500	None	50/60	2	1.1
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0581	H.265	2	medium	2160.0	3840.0	22500/33750	None	50/60	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.7: Test Plan for P2STR08.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0041	H.265	2	medium	360.0	640.0	450/600	None	15	2	2.0
HRC0058	H.264	2	medium	480.0	854.0	500/800	None	15	2	2.0
HRC0061	H.264	2	medium	480.0	854.0	500/800	None	15	2	1.1
HRC0069	VP9	1	3	480.0	854.0	675/900	None	15	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0165	VP9	2	2	480.0	854.0	825/1050	None	24/25/30	2	2.0
HRC0181	H.264	2	medium	480.0	854.0	1400/2000	None	24/25/30	5	1.1
HRC0185	H.265	1	ultrafast	720.0	1280.0	375/565	None	24/25/30	2	2.0
HRC0203	H.265	2	medium	720.0	1280.0	750/1125	None	24/25/30	5	2.0
HRC0204	VP9	2	2	720.0	1280.0	750/1125	None	24/25/30	5	2.0
HRC0210	VP9	1	3	720.0	1280.0	1500/2250	None	24/25/30	2	2.0
HRC0214	H.264	2	medium	720.0	1280.0	2000/3000	None	24/25/30	2	1.1
HRC0231	VP9	2	2	720.0	1280.0	2400/3000	None	24/25/30	2	2.0
HRC0242	H.265	2	medium	1080.0	1920.0	750/1500	None	24/25/30	2	1.1
HRC0243	VP9	2	2	1080.0	1920.0	750/1500	None	24/25/30	2	1.1
HRC0260	H.265	1	fast	1080.0	1920.0	5250/7125	None	24/25/30	2	2.0
HRC0268	H.264	2	medium	1080.0	1920.0	7000/9500	None	24/25/30	5	2.0
HRC0314	H.265	2	medium	1440.0	2560.0	1050/2000	None	24/25/30	5	1.1
HRC0322	H.264	2	medium	1440.0	2560.0	4500/7000	None	24/25/30	2	1.1
HRC0328	H.264	2	medium	1440.0	2560.0	10000/15000	None	24/25/30	5	1.1
HRC0329	H.265	2	medium	1440.0	2560.0	7500/11250	None	24/25/30	5	1.1
HRC0335	H.265	1	veryfast	480.0	854.0	1100/1500	None	24/25/30	2	2.0
HRC0358	H.264	2	medium	2160.0	3840.0	15000/22000	None	24/25/30	5	1.1
HRC0366	VP9	1	4	360.0	640.0	550/800	None	24/25/30	2	2.0
HRC0374	H.265	2	medium	1080.0	1920.0	600/1200	None	24/25/30	5	2.0
HRC0380	H.265	1	fast	1080.0	1920.0	750/1500	None	24/25/30	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0391	H.264	2	medium	720.0	1280.0	500/750	None	50/60	2	1.1
HRC0393	VP9	2	2	720.0	1280.0	375/565	None	50/60	2	1.1
HRC0398	H.265	2	medium	720.0	1280.0	375/565	None	50/60	5	1.1
HRC0409	H.264	2	medium	720.0	1280.0	2400/3200	None	50/60	5	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0444	VP9	2	2	1080.0	1920.0	2625/5250	None	50/60	5	2.0
HRC0476	H.265	2	medium	1440.0	2560.0	1315/2625	None	50/60	5	1.1
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0520	H.264	2	medium	1440.0	2560.0	15000/20000	None	50/60	2	1.1
HRC0527	H.265	2	medium	1440.0	2560.0	11250/15000	None	50/60	5	1.1
HRC0534	VP9	2	2	2160.0	3840.0	4500/9000	None	50/60	5	1.1
HRC0548	H.265	2	medium	2160.0	3840.0	11250/16500	None	50/60	5	1.1
HRC0562	H.264	2	medium	2160.0	3840.0	22000/30000	None	50/60	5	2.0
HRC0570	VP9	1	3	2160.0	3840.0	22500/33750	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0574	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	1.1

Table B.8: Test Plan for P2STR09.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0038	H.265	1	fast	360.0	640.0	450/600	None	15	2	2.0
HRC0059	H.265	2	medium	480.0	854.0	375/600	None	15	2	2.0
HRC0067	H.264	1	veryfast	480.0	854.0	900/1200	None	15	2	2.0
HRC0074	H.265	2	medium	480.0	854.0	675/900	None	15	2	1.1
HRC0079	H.264	2	medium	480.0	854.0	1200/1400	None	15	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0118	H.264	1	ultrafast	360.0	640.0	400/600	None	24/25/30	2	2.0
HRC0130	H.264	1	ultrafast	360.0	640.0	800/1000	None	24/25/30	2	2.0
HRC0137	H.265	2	medium	360.0	640.0	600/750	None	24/25/30	5	2.0
HRC0142	H.264	2	medium	360.0	640.0	1000/1200	None	24/25/30	2	2.0
HRC0151	H.264	2	slow	480.0	854.0	350/700	None	24/25/30	2	2.0
HRC0190	H.264	2	medium	720.0	1280.0	500/750	None	24/25/30	2	1.1
HRC0211	H.264	2	medium	720.0	1280.0	2000/3000	None	24/25/30	2	2.0
HRC0216	VP9	2	2	720.0	1280.0	1500/2250	None	24/25/30	2	1.1
HRC0223	H.264	2	slower	720.0	1280.0	2000/3000	None	24/25/30	2	2.0
HRC0225	VP9	2	1	720.0	1280.0	1500/2250	None	24/25/30	2	2.0
HRC0246	VP9	2	2	1080.0	1920.0	750/1500	None	24/25/30	5	2.0
HRC0254	H.265	2	medium	1080.0	1920.0	2250/4500	None	24/25/30	5	1.1
HRC0282	VP9	2	2	1080.0	1920.0	7125/9000	None	24/25/30	2	1.1
HRC0299	H.265	2	medium	1440.0	2560.0	1125/2250	None	24/25/30	2	1.1
HRC0308	H.265	2	slower	1440.0	2560.0	1125/2250	None	24/25/30	2	2.0
HRC0317	H.265	1	ultrafast	1440.0	2560.0	7500/11250	None	24/25/30	2	2.0
HRC0339	VP9	2	2	1440.0	2560.0	11250/15000	None	24/25/30	2	2.0
HRC0348	VP9	2	2	2160.0	3840.0	4500/9000	None	24/25/30	2	1.1
HRC0377	H.265	2	medium	2160.0	3840.0	16500/22500	None	24/25/30	5	1.1
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0395	H.265	2	medium	720.0	1280.0	375/565	None	50/60	5	2.0
HRC0397	H.264	2	medium	720.0	1280.0	500/750	None	50/60	5	1.1
HRC0417	VP9	1	4	720.0	1280.0	2400/3000	None	50/60	2	2.0
HRC0432	VP9	2	1	1080.0	1920.0	825/1200	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0464	H.265	2	slower	1080.0	1920.0	5250/7125	None	50/60	2	2.0
HRC0472	H.264	2	medium	1440.0	2560.0	1750/3500	None	50/60	5	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0504	VP9	2	2	1440.0	2560.0	7500/11250	None	50/60	2	1.1
HRC0512	H.265	2	slow	1440.0	2560.0	7500/11250	None	50/60	2	2.0
HRC0513	VP9	2	1	1440.0	2560.0	7500/11250	None	50/60	2	2.0
HRC0515	H.265	1	ultrafast	1440.0	2560.0	11250/15000	None	50/60	2	2.0
HRC0535	H.264	1	ultrafast	2160.0	3840.0	15000/22000	None	50/60	2	2.0
HRC0537	VP9	1	3	2160.0	3840.0	11250/16500	None	50/60	2	2.0
HRC0538	H.264	2	medium	2160.0	3840.0	15000/22000	None	50/60	2	2.0
HRC0539	H.265	2	medium	2160.0	3840.0	11250/16500	None	50/60	2	2.0
HRC0553	H.264	1	ultrafast	2160.0	3840.0	22000/30000	None	50/60	2	2.0
HRC0554	H.265	1	ultrafast	2160.0	3840.0	16500/22500	None	50/60	2	2.0
HRC0559	H.264	2	medium	2160.0	3840.0	22000/30000	None	50/60	2	1.1
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0572	H.265	2	medium	2160.0	3840.0	22500/33750	None	50/60	2	2.0
HRC0575	H.265	2	medium	2160.0	3840.0	22500/33750	None	50/60	2	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.9: Test Plan for P2STR10.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0026	H.265	2	medium	360.0	640.0	95/190	None	15	5	2.0
HRC0043	H.264	1	fast	360.0	640.0	800/1000	None	15	2	2.0
HRC0046	H.264	2	medium	360.0	640.0	800/1000	None	15	2	2.0
HRC0071	H.265	2	medium	480.0	854.0	675/900	None	15	2	2.0
HRC0072	VP9	2	2	480.0	854.0	675/900	None	15	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0126	VP9	2	2	360.0	640.0	300/450	None	24/25/30	5	2.0
HRC0152	H.265	2	slower	480.0	854.0	265/525	None	24/25/30	2	2.0
HRC0168	VP9	2	2	480.0	854.0	825/1050	None	24/25/30	2	1.1
HRC0174	VP9	2	2	480.0	854.0	825/1050	None	24/25/30	5	1.1
HRC0188	H.265	2	medium	720.0	1280.0	375/565	None	24/25/30	2	2.0
HRC0234	VP9	2	0	720.0	1280.0	2400/3000	None	24/25/30	2	2.0
HRC0237	VP9	1	3	1080.0	1920.0	750/1500	None	24/25/30	2	2.0
HRC0251	H.265	2	medium	1080.0	1920.0	2250/4500	None	24/25/30	5	2.0
HRC0259	H.264	1	fast	1080.0	1920.0	150/250	None	24/25/30	2	2.0
HRC0277	H.264	2	medium	1080.0	1920.0	9500/12000	None	24/25/30	2	2.0
HRC0286	H.264	2	medium	1080.0	1920.0	9500/12000	None	24/25/30	5	1.1
HRC0291	VP9	2	0	1080.0	1920.0	7125/9000	None	24/25/30	2	2.0
HRC0313	H.264	2	medium	1440.0	2560.0	4000/8000	None	24/25/30	5	1.1
HRC0345	VP9	2	2	2160.0	3840.0	300/450	None	24/25/30	2	2.0
HRC0346	H.264	2	medium	2160.0	3840.0	6000/12000	None	24/25/30	2	1.1
HRC0347	H.265	2	medium	2160.0	3840.0	4500/9000	None	24/25/30	2	1.1
HRC0382	H.264	2	medium	2160.0	3840.0	30000/45000	None	24/25/30	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0389	H.265	2	medium	720.0	1280.0	375/565	None	50/60	2	2.0
HRC0403	H.264	2	medium	720.0	1280.0	1100/1600	None	50/60	5	1.1
HRC0405	VP9	2	2	720.0	1280.0	825/1200	None	50/60	5	1.1
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0440	H.265	2	medium	1080.0	1920.0	2625/5250	None	50/60	2	1.1
HRC0447	VP9	2	2	1080.0	1920.0	2625/5250	None	50/60	5	1.1
HRC0449	H.265	1	ultrafast	1080.0	1920.0	600/750	None	50/60	2	2.0
HRC0453	VP9	2	2	1080.0	1920.0	5250/7125	None	50/60	2	2.0
HRC0468	VP9	2	2	1080.0	1920.0	7125/9000	None	50/60	5	2.0
HRC0475	H.264	2	medium	1440.0	2560.0	1750/3500	None	50/60	5	1.1
HRC0481	H.264	1	veryfast	1440.0	2560.0	4000/7000	None	50/60	2	2.0
HRC0482	H.265	1	fast	1440.0	2560.0	3000/5250	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0485	H.265	2	medium	1440.0	2560.0	3000/5250	None	50/60	2	2.0
HRC0499	H.264	2	medium	1440.0	2560.0	10000/15000	None	50/60	2	2.0
HRC0501	VP9	2	2	1440.0	2560.0	7500/11250	None	50/60	2	2.0
HRC0518	H.265	2	medium	1440.0	2560.0	600/750	None	50/60	2	2.0
HRC0536	H.265	1	ultrafast	2160.0	3840.0	11250/16500	None	50/60	2	2.0
HRC0540	VP9	2	2	2160.0	3840.0	11250/16500	None	50/60	2	2.0
HRC0543	VP9	2	2	2160.0	3840.0	11250/16500	None	50/60	2	1.1
HRC0566	H.265	2	medium	2160.0	3840.0	16500/22500	None	50/60	5	1.1
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0577	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	5	2.0

Table B.10: Test Plan for P2STR11.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0025	H.264	2	medium	360.0	640.0	125/250	None	15	5	2.0
HRC0036	VP9	2	0	360.0	640.0	225/375	None	15	2	2.0
HRC0040	H.264	2	medium	360.0	640.0	600/800	None	15	2	2.0
HRC0042	VP9	2	2	360.0	640.0	150/200	None	15	2	2.0
HRC0047	H.265	2	medium	360.0	640.0	600/750	None	15	2	2.0
HRC0062	H.265	2	medium	480.0	854.0	375/600	None	15	2	1.1
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0116	H.265	2	medium	360.0	640.0	115/190	None	24/25/30	5	2.0
HRC0122	H.265	2	medium	360.0	640.0	300/450	None	24/25/30	2	2.0
HRC0128	H.265	2	slower	360.0	640.0	300/450	None	24/25/30	2	2.0
HRC0146	H.265	2	medium	480.0	854.0	265/525	None	24/25/30	5	2.0
HRC0155	H.265	2	medium	480.0	854.0	250/375	None	24/25/30	5	2.0
HRC0159	VP9	2	2	480.0	854.0	250/375	None	24/25/30	5	1.1
HRC0162	VP9	1	4	480.0	854.0	825/1050	None	24/25/30	2	2.0
HRC0172	H.264	2	medium	480.0	854.0	1100/1400	None	24/25/30	5	1.1
HRC0192	VP9	2	2	720.0	1280.0	375/565	None	24/25/30	2	1.1
HRC0205	H.264	2	medium	720.0	1280.0	1000/1500	None	24/25/30	5	1.1
HRC0219	VP9	2	2	720.0	1280.0	1500/2250	None	24/25/30	5	2.0
HRC0239	H.265	2	medium	1080.0	1920.0	300/600	None	24/25/30	2	2.0
HRC0241	H.264	2	medium	1080.0	1920.0	1000/2000	None	24/25/30	2	1.1
HRC0253	H.264	2	medium	1080.0	1920.0	1500/3000	None	24/25/30	5	1.1
HRC0257	H.265	2	slower	1080.0	1920.0	2250/4500	None	24/25/30	2	2.0
HRC0283	H.264	2	medium	1080.0	1920.0	9500/12000	None	24/25/30	5	2.0
HRC0288	VP9	2	2	1080.0	1920.0	7125/9000	None	24/25/30	5	1.1
HRC0296	H.265	2	medium	1440.0	2560.0	500/1000	None	24/25/30	2	2.0
HRC0341	H.265	1	veryfast	2160.0	3840.0	1500/3000	None	24/25/30	2	2.0
HRC0342	VP9	1	4	2160.0	3840.0	1500/3000	None	24/25/30	2	2.0
HRC0375	VP9	2	2	2160.0	3840.0	6000/8000	None	24/25/30	5	2.0
HRC0379	H.264	1	veryfast	2160.0	3840.0	30000/45000	None	24/25/30	2	2.0
HRC0387	VP9	1	3	720.0	1280.0	375/565	None	50/60	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0394	H.264	2	medium	720.0	1280.0	500/750	None	50/60	5	2.0
HRC0408	VP9	2	1	720.0	1280.0	825/1200	None	50/60	2	2.0
HRC0419	H.265	2	medium	720.0	1280.0	2400/3000	None	50/60	2	2.0
HRC0421	H.264	2	slower	720.0	1280.0	3200/4000	None	50/60	2	2.0
HRC0703	H.264	2	medium	1080.0	1920.0	1500/3000	None	50/60	2	2.0
HRC0443	H.265	2	medium	1080.0	1920.0	2625/5250	None	50/60	5	2.0
HRC0445	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	5	1.1
HRC0478	H.264	2	veryslow	1440.0	2560.0	1750/3500	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0531	VP9	2	2	2160.0	3840.0	1500/3000	None	50/60	5	2.0
HRC0533	H.265	2	medium	2160.0	3840.0	1500/3000	None	50/60	5	1.1
HRC0549	VP9	2	2	2160.0	3840.0	11250/16500	None	50/60	5	1.1
HRC0556	H.264	2	medium	2160.0	3840.0	7000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0

Appendix B P.NATS Phase 2 Test Plan

Table B.11: Test Plan for P2STR12.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0027	VP9	2	2	360.0	640.0	95/190	None	15	5	2.0
HRC0063	VP9	2	2	480.0	854.0	375/600	None	15	2	1.1
HRC0075	VP9	2	2	480.0	854.0	675/900	None	15	2	1.1
HRC0082	H.264	2	slower	480.0	854.0	1200/1400	None	15	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0133	H.264	2	medium	360.0	640.0	800/1000	None	24/25/30	2	2.0
HRC0145	H.264	2	medium	480.0	854.0	350/700	None	24/25/30	5	2.0
HRC0158	H.265	2	medium	480.0	854.0	525/750	None	24/25/30	5	1.1
HRC0176	H.265	2	veryslow	480.0	854.0	825/1050	None	24/25/30	2	2.0
HRC0207	VP9	2	2	720.0	1280.0	750/1125	None	24/25/30	5	1.1
HRC0215	H.265	2	medium	720.0	1280.0	1500/2250	None	24/25/30	2	1.1
HRC0226	H.264	1	ultrafast	720.0	1280.0	3200/4000	None	24/25/30	2	2.0
HRC0227	H.265	1	ultrafast	720.0	1280.0	2400/3000	None	24/25/30	2	2.0
HRC0248	H.265	2	medium	1080.0	1920.0	750/1500	None	24/25/30	5	1.1
HRC0271	H.264	2	medium	1080.0	1920.0	7000/9500	None	24/25/30	5	1.1
HRC0274	H.264	1	fast	1080.0	1920.0	9500/12000	None	24/25/30	2	2.0
HRC0301	H.264	2	medium	1440.0	2560.0	1500/3000	None	24/25/30	5	2.0
HRC0344	H.265	2	medium	2160.0	3840.0	4500/9000	None	24/25/30	2	2.0
HRC0351	VP9	2	2	2160.0	3840.0	4500/9000	None	24/25/30	5	2.0
HRC0352	H.264	2	medium	2160.0	3840.0	6000/12000	None	24/25/30	5	1.1
HRC0357	VP9	2	2	2160.0	3840.0	11250/16500	None	24/25/30	5	2.0
HRC0376	H.264	2	medium	2160.0	3840.0	22000/30000	None	24/25/30	5	1.1
HRC0383	H.265	2	medium	2160.0	3840.0	500/800	None	24/25/30	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0406	H.264	2	slower	720.0	1280.0	1100/1600	None	50/60	2	2.0
HRC0428	H.265	2	medium	1080.0	1920.0	825/1200	None	50/60	5	1.1
HRC0433	H.264	1	ultrafast	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0435	VP9	1	3	1080.0	1920.0	2625/5250	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0442	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	5	2.0
HRC0451	H.264	2	medium	1080.0	1920.0	500/800	None	50/60	2	2.0
HRC0461	H.265	2	medium	1080.0	1920.0	5250/7125	None	50/60	5	1.1
HRC0471	VP9	2	2	1080.0	1920.0	7125/9000	None	50/60	5	1.1
HRC0474	VP9	2	2	1440.0	2560.0	1315/2625	None	50/60	5	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0487	H.264	2	medium	1440.0	2560.0	4000/7000	None	50/60	2	1.1
HRC0506	H.265	2	medium	1440.0	2560.0	7500/11250	None	50/60	5	2.0
HRC0508	H.264	2	medium	1440.0	2560.0	10000/15000	None	50/60	5	1.1
HRC0514	H.264	1	fast	1440.0	2560.0	15000/20000	None	50/60	2	2.0
HRC0516	VP9	1	4	1440.0	2560.0	11250/15000	None	50/60	2	2.0
HRC0517	H.264	2	medium	1440.0	2560.0	15000/20000	None	50/60	2	2.0
HRC0519	VP9	2	2	1440.0	2560.0	500/800	None	50/60	2	2.0
HRC0545	H.265	2	medium	2160.0	3840.0	11250/16500	None	50/60	5	2.0
HRC0550	H.264	2	veryslow	2160.0	3840.0	15000/22000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0573	VP9	2	2	2160.0	3840.0	22500/33750	None	50/60	2	2.0
HRC0580	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	5	1.1

Table B.12: Test Plan for P2STR13.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0052	H.264	2	medium	480.0	854.0	250/500	None	15	5	1.1
HRC0080	H.265	2	medium	480.0	854.0	900/1050	None	15	2	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0119	H.265	1	fast	360.0	640.0	300/450	None	24/25/30	2	2.0
HRC0129	VP9	2	0	360.0	640.0	300/450	None	24/25/30	2	2.0
HRC0140	H.265	1	ultrafast	360.0	640.0	750/900	None	24/25/30	2	2.0
HRC0150	VP9	2	2	480.0	854.0	75/150	None	24/25/30	5	1.1
HRC0164	H.265	2	medium	480.0	854.0	400/500	None	24/25/30	2	2.0
HRC0179	H.265	2	medium	480.0	854.0	1050/1500	None	24/25/30	5	2.0
HRC0209	H.265	1	fast	720.0	1280.0	1500/2250	None	24/25/30	2	2.0
HRC0220	H.264	2	medium	720.0	1280.0	2000/3000	None	24/25/30	5	1.1
HRC0228	VP9	1	4	720.0	1280.0	2400/3000	None	24/25/30	2	2.0
HRC0238	H.264	2	medium	1080.0	1920.0	1000/2000	None	24/25/30	2	2.0
HRC0247	H.264	2	medium	1080.0	1920.0	1000/2000	None	24/25/30	5	1.1
HRC0264	VP9	2	2	1080.0	1920.0	5250/7125	None	24/25/30	2	2.0
HRC0303	VP9	2	2	1440.0	2560.0	400/800	None	24/25/30	5	2.0
HRC0307	H.264	2	veryslow	1440.0	2560.0	1500/3000	None	24/25/30	2	2.0
HRC0320	H.265	2	medium	1440.0	2560.0	2500/3750	None	24/25/30	2	2.0
HRC0323	H.265	2	medium	1440.0	2560.0	2500/3750	None	24/25/30	2	1.1
HRC0325	H.264	2	medium	1440.0	2560.0	10000/15000	None	24/25/30	5	2.0
HRC0332	H.265	2	veryslow	1440.0	2560.0	2500/3750	None	24/25/30	2	2.0
HRC0333	VP9	2	1	1440.0	2560.0	2500/3750	None	24/25/30	2	2.0
HRC0338	H.265	2	medium	1440.0	2560.0	11250/15000	None	24/25/30	2	2.0
HRC0361	H.264	2	slow	2160.0	3840.0	5000/7500	None	24/25/30	2	2.0
HRC0363	VP9	2	0	2160.0	3840.0	11250/16500	None	24/25/30	2	2.0
HRC0371	H.265	2	medium	2160.0	3840.0	6000/8000	None	24/25/30	2	1.1
HRC0378	VP9	2	2	2160.0	3840.0	6000/8000	None	24/25/30	5	1.1
HRC0384	VP9	2	2	2160.0	3840.0	7500/10000	None	24/25/30	2	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0396	VP9	2	2	720.0	1280.0	375/565	None	50/60	5	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0441	VP9	2	2	1080.0	1920.0	2625/5250	None	50/60	2	1.1
HRC0465	VP9	2	1	1080.0	1920.0	5250/7125	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0503	H.265	2	medium	1440.0	2560.0	7500/11250	None	50/60	2	1.1
HRC0523	H.264	2	medium	1440.0	2560.0	15000/20000	None	50/60	5	2.0
HRC0551	H.265	2	veryslow	2160.0	3840.0	11250/16500	None	50/60	2	2.0
HRC0560	H.265	2	medium	2160.0	3840.0	6000/8000	None	50/60	2	1.1
HRC0565	H.264	2	medium	2160.0	3840.0	22000/30000	None	50/60	5	1.1
HRC0569	H.265	1	veryfast	2160.0	3840.0	7500/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0576	VP9	2	2	2160.0	3840.0	7500/10000	None	50/60	2	1.1
HRC0582	VP9	2	2	2160.0	3840.0	7500/10000	None	50/60	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.13: Test Plan for P2STR14.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0035	H.265	2	slower	360.0	640.0	225/375	None	15	2	2.0
HRC0066	VP9	2	0	480.0	854.0	375/600	None	15	2	2.0
HRC0073	H.264	2	medium	480.0	854.0	900/1200	None	15	2	1.1
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0131	H.265	1	ultrafast	360.0	640.0	600/750	None	24/25/30	2	2.0
HRC0153	VP9	2	0	480.0	854.0	265/525	None	24/25/30	2	2.0
HRC0170	H.265	2	medium	480.0	854.0	825/1050	None	24/25/30	5	2.0
HRC0173	H.265	2	medium	480.0	854.0	825/1050	None	24/25/30	5	1.1
HRC0177	VP9	2	0	480.0	854.0	825/1050	None	24/25/30	2	2.0
HRC0202	H.264	2	medium	720.0	1280.0	1000/1500	None	24/25/30	5	2.0
HRC0208	H.264	1	fast	720.0	1280.0	2000/3000	None	24/25/30	2	2.0
HRC0224	H.265	2	slower	720.0	1280.0	1500/2250	None	24/25/30	2	2.0
HRC0255	VP9	2	2	1080.0	1920.0	2250/4500	None	24/25/30	5	1.1
HRC0275	H.265	1	veryfast	1080.0	1920.0	7125/9000	None	24/25/30	2	2.0
HRC0280	H.264	2	medium	1080.0	1920.0	9500/12000	None	24/25/30	2	1.1
HRC0284	H.265	2	medium	1080.0	1920.0	7125/9000	None	24/25/30	5	2.0
HRC0289	H.264	2	veryslow	1080.0	1920.0	9500/12000	None	24/25/30	2	2.0
HRC0302	H.265	2	medium	1440.0	2560.0	1125/2250	None	24/25/30	5	2.0
HRC0306	VP9	2	2	1440.0	2560.0	1125/2250	None	24/25/30	5	1.1
HRC0315	VP9	2	2	1440.0	2560.0	3000/6000	None	24/25/30	5	1.1
HRC0324	VP9	2	2	1440.0	2560.0	7500/11250	None	24/25/30	2	1.1
HRC0327	VP9	2	2	1440.0	2560.0	7500/11250	None	24/25/30	5	2.0
HRC0340	H.264	1	ultrafast	2160.0	3840.0	6000/12000	None	24/25/30	2	2.0
HRC0359	H.265	2	medium	2160.0	3840.0	11250/16500	None	24/25/30	5	1.1
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0392	H.265	2	medium	720.0	1280.0	375/565	None	50/60	2	1.1
HRC0411	VP9	2	2	720.0	1280.0	1800/2400	None	50/60	5	2.0
HRC0414	VP9	2	2	720.0	1280.0	1800/2400	None	50/60	5	1.1
HRC0415	H.264	1	fast	720.0	1280.0	3200/4000	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0454	H.264	2	medium	1080.0	1920.0	7000/9500	None	50/60	2	1.1
HRC0467	H.265	2	medium	1080.0	1920.0	7125/9000	None	50/60	5	2.0
HRC0479	H.265	2	slower	1440.0	2560.0	1315/2625	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0493	H.264	2	medium	1440.0	2560.0	4000/7000	None	50/60	5	1.1
HRC0495	VP9	2	2	1440.0	2560.0	3000/5250	None	50/60	5	1.1
HRC0511	H.264	2	slow	1440.0	2560.0	10000/15000	None	50/60	2	2.0
HRC0524	H.265	2	medium	1440.0	2560.0	11250/15000	None	50/60	5	2.0
HRC0529	H.264	2	medium	2160.0	3840.0	6000/12000	None	50/60	5	2.0
HRC0530	H.265	2	medium	2160.0	3840.0	4500/9000	None	50/60	5	2.0
HRC0546	VP9	2	2	2160.0	3840.0	11250/16500	None	50/60	5	2.0
HRC0552	VP9	2	1	2160.0	3840.0	11250/16500	None	50/60	2	2.0
HRC0567	VP9	2	2	2160.0	3840.0	16500/22500	None	50/60	5	1.1
HRC0568	H.264	1	fast	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0578	H.265	2	medium	2160.0	3840.0	22500/33750	None	50/60	5	2.0
HRC0579	VP9	2	2	2160.0	3840.0	22500/33750	None	50/60	5	2.0

Table B.14: Test Plan for P2SVL01.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0618	VP9	2	3	720.0	1280.0	2206/2942	None	24/25/30	2	1.1
HRC0619	H.264	2	slow	1080.0	1920.0	300/450	None	24/25/30	2	1.1
HRC0620	H.264	1	slow	360.0	640.0	984/1313	None	24/25/30	2	1.1
HRC0621	H.264	2	medium	1080.0	1920.0	1326/1769	None	50/60	2	1.1
HRC0622	VP9	2	4	480.0	854.0	710/947	None	24/25/30	2	1.1
HRC0624	H.265	2	medium	2160.0	3840.0	8750/11667	None	24/25/30	5	1.1
HRC0625	H.264	2	medium	2160.0	3840.0	24507/32676	None	50/60	2	1.1
HRC0626	H.265	2	medium	480.0	854.0	895/1194	None	24/25/30	5	1.1
HRC0627	H.264	2	medium	2160.0	3840.0	5538/7384	None	50/60	2	1.1
HRC0628	H.265	2	medium	2160.0	3840.0	3720/4960	None	24/25/30	2	1.1
HRC0629	VP9	2	2	1080.0	1920.0	606/809	None	24/25/30	2	1.1
HRC0630	VP9	1	2	1440.0	2560.0	2824/3766	None	24/25/30	5	1.1
HRC0631	VP9	2	2	2160.0	3840.0	3501/4669	None	50/60	2	1.1
HRC0632	H.264	2	medium	2160.0	3840.0	5343/7124	None	50/60	2	1.1
HRC0633	H.265	1	fast	1440.0	2560.0	7943/10591	None	24/25/30	5	1.1
HRC0634	H.265	2	medium	1440.0	2560.0	9627/12837	None	50/60	5	1.1
HRC0635	H.265	2	medium	2160.0	3840.0	9554/12739	None	24/25/30	5	1.1
HRC0636	VP9	2	3	2160.0	3840.0	4493/5991	None	50/60	5	1.1
HRC0638	H.264	2	medium	1080.0	1920.0	1201/1602	None	50/60	2	1.1
HRC0639	H.265	1	medium	720.0	1280.0	4232/5643	None	24/25/30	5	1.1
HRC0640	H.264	2	medium	480.0	854.0	1381/1842	None	24/25/30	5	1.1
HRC0641	VP9	1	2	2160.0	3840.0	400/550	None	24/25/30	5	1.1
HRC0642	VP9	2	2	1080.0	1920.0	3155/4207	None	24/25/30	2	1.1
HRC0643	VP9	2	2	2160.0	3840.0	450/600	None	24/25/30	2	1.1
HRC0644	H.265	2	medium	360.0	640.0	306/408	None	24/25/30	2	1.1
HRC0645	VP9	1	0	540.0	960.0	988/1318	None	24/25/30	2	1.1
HRC0646	H.265	1	medium	2160.0	3840.0	4939/6586	None	24/25/30	5	1.1
HRC0647	H.265	2	ultrafast	1440.0	2560.0	500/650	None	24/25/30	2	1.1
HRC0648	H.265	2	ultrafast	1440.0	2560.0	10499/13999	None	24/25/30	2	1.1
HRC0650	H.265	2	medium	540.0	960.0	300/450	None	24/25/30	2	1.1
HRC0651	H.265	2	medium	480.0	854.0	663/884	None	24/25/30	2	1.1
HRC0652	VP9	2	2	720.0	1280.0	4735/6314	None	24/25/30	5	1.1
HRC0653	VP9	2	1	2160.0	3840.0	15123/20165	None	24/25/30	2	1.1
HRC0654	H.264	2	medium	1440.0	2560.0	1461/1949	None	24/25/30	5	1.1
HRC0655	H.264	2	medium	720.0	1280.0	913/1218	None	24/25/30	2	1.1
HRC0656	H.264	1	medium	540.0	960.0	1461/1949	None	24/25/30	5	1.1
HRC0657	H.265	2	medium	1440.0	2560.0	11190/14921	None	24/25/30	5	1.1
HRC0658	H.265	1	fast	720.0	1280.0	1413/1884	None	24/25/30	5	1.1
HRC0659	H.264	2	medium	1080.0	1920.0	744/993	None	24/25/30	2	1.1
HRC0660	H.264	1	medium	2160.0	3840.0	3759/5013	None	24/25/30	2	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.15: Test Plan for P2SVL02.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0663	H.265	2	medium	1080.0	1920.0	4344/5793	None	24/25/30	5	1.1
HRC0664	H.265	2	medium	480.0	854.0	374/499	None	24/25/30	2	1.1
HRC0668	H.264	2	medium	1440.0	2560.0	3396/4529	None	24/25/30	2	1.1
HRC0669	H.264	2	veryfast	1080.0	1920.0	2453/3271	None	24/25/30	2	1.1
HRC0670	VP9	2	2	1440.0	2560.0	7869/10492	None	24/25/30	5	1.1
HRC0671	VP9	2	2	1440.0	2560.0	1337/1783	None	24/25/30	2	1.1
HRC0672	VP9	1	2	540.0	960.0	1402/1870	None	50/60	2	1.1
HRC0673	VP9	2	0	720.0	1280.0	1394/1859	None	24/25/30	2	1.1
HRC0675	H.264	2	medium	1080.0	1920.0	3341/4455	None	50/60	2	1.1
HRC0676	H.265	1	medium	480.0	854.0	1351/1802	None	24/25/30	5	1.1
HRC0677	H.265	2	medium	720.0	1280.0	762/1017	None	24/25/30	2	1.1
HRC0678	VP9	1	2	1080.0	1920.0	7416/9888	None	24/25/30	2	1.1
HRC0679	H.264	1	medium	240.0	426.0	474/632	None	24/25/30	2	1.1
HRC0680	H.265	1	veryfast	360.0	640.0	154/206	None	24/25/30	2	1.1
HRC0681	H.265	2	medium	720.0	1280.0	506/675	None	24/25/30	2	1.1
HRC0683	H.265	1	medium	540.0	960.0	469/626	None	24/25/30	2	1.1
HRC0684	VP9	1	2	480.0	854.0	1630/2174	None	24/25/30	2	1.1
HRC0685	H.265	1	slower	360.0	640.0	768/1025	None	24/25/30	5	1.1
HRC0687	H.265	1	fast	720.0	1280.0	3996/5328	None	24/25/30	5	1.1
HRC0689	H.265	2	medium	360.0	640.0	667/890	None	24/25/30	2	1.1
HRC0690	H.265	1	fast	1080.0	1920.0	1093/1458	None	24/25/30	2	1.1
HRC0692	VP9	2	2	720.0	1280.0	1245/1661	None	24/25/30	2	1.1
HRC0694	VP9	2	2	240.0	426.0	588/785	None	24/25/30	2	1.1
HRC0695	H.264	2	fast	480.0	854.0	1488/1985	None	50/60	5	1.1
HRC0696	VP9	2	2	720.0	1280.0	516/689	None	24/25/30	5	1.1
HRC0697	VP9	2	2	1440.0	2560.0	6120/8161	None	50/60	2	1.1
HRC0699	H.264	2	veryfast	360.0	640.0	440/587	None	24/25/30	5	1.1
HRC0700	VP9	2	2	480.0	854.0	2471/3295	None	15	2	1.1
HRC0701	H.265	2	fast	540.0	960.0	456/609	None	50/60	5	1.1
HRC0702	H.264	1	ultrafast	1440.0	2560.0	1194/1592	None	24/25/30	5	1.1
HRC0703	H.264	2	medium	540.0	960.0	1697/2263	None	24/25/30	5	1.1
HRC0705	VP9	1	2	240.0	426.0	120/161	None	24/25/30	5	1.1
HRC0706	H.265	2	ultrafast	1440.0	2560.0	5737/7650	None	24/25/30	5	1.1
HRC0707	H.264	2	medium	540.0	960.0	762/1016	None	50/60	5	1.1

Table B.16: Test Plan for P2SVL03.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0708	H.264	1	medium	540.0	960.0	249/332	None	24/25/30	2	1.1
HRC0709	H.264	2	medium	720.0	1280.0	2582/3443	None	24/25/30	2	1.1
HRC0710	H.264	2	medium	1440.0	2560.0	5699/7599	None	24/25/30	5	1.1
HRC0711	VP9	2	0	720.0	1280.0	716/955	None	24/25/30	5	1.1
HRC0712	VP9	2	3	540.0	960.0	365/487	None	24/25/30	2	1.1
HRC0713	VP9	2	3	720.0	1280.0	646/862	None	24/25/30	2	1.1
HRC0715	H.264	2	medium	720.0	1280.0	1512/2017	None	24/25/30	5	1.1
HRC0717	VP9	2	2	1080.0	1920.0	431/575	None	24/25/30	2	1.1
HRC0718	VP9	2	2	720.0	1280.0	2265/3020	None	24/25/30	2	1.1
HRC0719	VP9	1	3	1080.0	1920.0	401/535	None	50/60	2	1.1
HRC0720	H.265	1	medium	240.0	426.0	746/995	None	24/25/30	2	1.1
HRC0721	H.264	2	medium	1080.0	1920.0	10088/13451	None	24/25/30	2	1.1
HRC0722	VP9	2	2	240.0	426.0	393/524	None	24/25/30	5	1.1
HRC0723	H.265	1	ultrafast	240.0	426.0	697/930	None	24/25/30	5	1.1
HRC0724	H.264	2	medium	540.0	960.0	1186/1582	None	15	2	1.1
HRC0725	VP9	1	2	480.0	854.0	2238/2984	None	50/60	2	1.1
HRC0726	VP9	2	2	480.0	854.0	1184/1579	None	50/60	2	1.1
HRC0727	H.264	1	medium	1080.0	1920.0	6894/9193	None	50/60	5	1.1
HRC0728	VP9	2	2	1440.0	2560.0	3778/5038	None	24/25/30	5	1.1
HRC0729	H.264	2	medium	1080.0	1920.0	720/961	None	50/60	2	1.1
HRC0730	H.265	2	medium	540.0	960.0	1956/2609	None	24/25/30	2	1.1
HRC0731	H.265	2	medium	360.0	640.0	718/958	None	24/25/30	2	1.1
HRC0732	H.264	1	medium	480.0	854.0	1194/1592	None	24/25/30	5	1.1
HRC0733	H.264	2	ultrafast	480.0	854.0	438/584	None	24/25/30	5	1.1
HRC0735	H.264	2	medium	480.0	854.0	512/683	None	24/25/30	2	1.1
HRC0736	H.264	2	medium	720.0	1280.0	1406/1875	None	50/60	5	1.1
HRC0737	H.264	2	medium	1440.0	2560.0	4655/6207	None	24/25/30	2	1.1
HRC0739	H.265	1	medium	240.0	426.0	704/939	None	24/25/30	5	1.1
HRC0740	H.264	2	medium	1080.0	1920.0	8316/11088	None	24/25/30	2	1.1
HRC0742	VP9	2	0	720.0	1280.0	5604/7473	None	24/25/30	2	1.1
HRC0743	H.265	2	medium	240.0	426.0	428/571	None	24/25/30	5	1.1
HRC0744	H.264	2	slower	240.0	426.0	535/714	None	24/25/30	5	1.1
HRC0745	H.264	2	medium	1440.0	2560.0	1440/1921	None	24/25/30	2	1.1
HRC0746	VP9	2	2	720.0	1280.0	3625/4834	None	24/25/30	2	1.1
HRC0747	VP9	1	2	1080.0	1920.0	7482/9977	None	24/25/30	2	1.1
HRC0748	H.265	2	slow	720.0	1280.0	2478/3304	None	24/25/30	2	1.1
HRC0749	VP9	2	2	240.0	426.0	561/748	None	24/25/30	5	1.1
HRC0750	H.264	2	medium	540.0	960.0	883/1178	None	50/60	5	1.1
HRC0752	H.264	2	medium	1440.0	2560.0	3826/5102	None	50/60	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.17: Test Plan for P2SVL04.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0753	H.265	2	medium	240.0	426.0	703/938	None	24/25/30	2	1.1
HRC0754	H.264	2	medium	1440.0	2560.0	5956/7942	None	24/25/30	5	1.1
HRC0755	H.264	2	ultrafast	720.0	1280.0	1827/2436	None	24/25/30	5	1.1
HRC0756	H.265	2	medium	540.0	960.0	1122/1496	None	24/25/30	2	1.1
HRC0758	VP9	1	2	720.0	1280.0	3369/4493	None	24/25/30	2	1.1
HRC0759	VP9	2	2	1080.0	1920.0	7593/10125	None	24/25/30	2	1.1
HRC0760	VP9	2	1	360.0	640.0	774/1032	None	24/25/30	5	1.1
HRC0761	H.265	2	medium	360.0	640.0	237/316	None	24/25/30	5	1.1
HRC0762	H.265	2	ultrafast	1080.0	1920.0	300/400	None	50/60	2	1.1
HRC0763	H.264	2	medium	1080.0	1920.0	717/956	None	24/25/30	2	1.1
HRC0765	H.264	2	medium	240.0	426.0	739/986	None	15	5	1.1
HRC0766	VP9	2	2	360.0	640.0	438/585	None	24/25/30	5	1.1
HRC0768	H.264	1	slower	240.0	426.0	342/457	None	24/25/30	5	1.1
HRC0769	H.264	2	medium	1440.0	2560.0	1227/1637	None	50/60	2	1.1
HRC0770	H.265	2	medium	720.0	1280.0	416/555	None	24/25/30	5	1.1
HRC0774	VP9	2	2	720.0	1280.0	5273/7031	None	24/25/30	5	1.1
HRC0775	H.264	1	medium	1440.0	2560.0	1200/1600	None	24/25/30	5	1.1
HRC0776	H.265	2	ultrafast	240.0	426.0	342/457	None	24/25/30	5	1.1
HRC0777	H.265	1	medium	480.0	854.0	90/120	None	24/25/30	5	1.1
HRC0778	VP9	2	2	1080.0	1920.0	4177/5570	None	24/25/30	2	1.1
HRC0779	VP9	2	4	240.0	426.0	135/180	None	24/25/30	5	1.1
HRC0780	VP9	2	2	720.0	1280.0	3990/5320	None	24/25/30	2	1.1
HRC0781	H.265	1	medium	360.0	640.0	946/1262	None	50/60	2	1.1
HRC0782	H.265	2	veryfast	480.0	854.0	749/999	None	24/25/30	2	1.1
HRC0784	H.264	1	medium	1080.0	1920.0	4481/5975	None	24/25/30	5	1.1
HRC0785	H.265	2	medium	1080.0	1920.0	435/580	None	50/60	2	1.1
HRC0787	VP9	1	2	1440.0	2560.0	1432/1910	None	50/60	2	1.1
HRC0788	H.265	2	fast	720.0	1280.0	1791/2388	None	24/25/30	5	1.1
HRC0789	H.264	1	medium	240.0	426.0	126/168	None	24/25/30	5	1.1
HRC0790	VP9	2	2	1080.0	1920.0	532/710	None	24/25/30	2	1.1
HRC0791	H.264	2	fast	540.0	960.0	705/941	None	24/25/30	2	1.1
HRC0793	VP9	2	2	540.0	960.0	863/1151	None	24/25/30	2	1.1
HRC0794	H.265	1	medium	240.0	426.0	100/140	None	24/25/30	5	1.1
HRC0796	H.265	1	ultrafast	360.0	640.0	502/670	None	24/25/30	5	1.1
HRC0797	H.264	2	ultrafast	480.0	854.0	340/440	None	24/25/30	2	1.1

Table B.18: Test Plan for P2SVL05.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0799	H.264	1	medium	720.0	1280.0	1448/1931	None	24/25/30	5	1.1
HRC0800	H.265	2	slow	2160.0	3840.0	2498/3331	None	24/25/30	2	1.1
HRC0801	H.264	2	medium	360.0	640.0	671/895	None	24/25/30	2	1.1
HRC0802	H.265	2	medium	480.0	854.0	866/1155	None	15	2	1.1
HRC0803	VP9	2	3	1440.0	2560.0	3581/4775	None	24/25/30	2	1.1
HRC0804	H.265	2	veryfast	1080.0	1920.0	550/650	None	24/25/30	2	1.1
HRC0807	VP9	1	3	360.0	640.0	450/650	None	24/25/30	5	1.1
HRC0808	VP9	2	2	360.0	640.0	200/250	None	24/25/30	2	1.1
HRC0809	H.264	2	medium	360.0	640.0	200/300	None	24/25/30	5	1.1
HRC0810	VP9	2	4	1440.0	2560.0	8298/11064	None	24/25/30	5	1.1
HRC0811	VP9	2	4	1080.0	1920.0	579/773	None	24/25/30	2	1.1
HRC0812	H.264	2	medium	1080.0	1920.0	2658/3545	None	50/60	5	1.1
HRC0813	H.265	2	fast	1440.0	2560.0	3336/4449	None	24/25/30	2	1.1
HRC0815	VP9	2	2	480.0	854.0	1024/1366	None	24/25/30	2	1.1
HRC0817	H.264	2	medium	1080.0	1920.0	2043/2724	None	24/25/30	2	1.1
HRC0818	VP9	2	3	2160.0	3840.0	2361/3149	None	24/25/30	5	1.1
HRC0819	H.264	1	fast	1440.0	2560.0	1495/1994	None	50/60	2	1.1
HRC0820	H.264	1	slow	1080.0	1920.0	3763/5018	None	24/25/30	5	1.1
HRC0821	VP9	2	2	480.0	854.0	809/1079	None	24/25/30	2	1.1
HRC0822	VP9	2	2	720.0	1280.0	1515/2021	None	24/25/30	5	1.1
HRC0823	VP9	2	2	360.0	640.0	250/350	None	24/25/30	2	1.1
HRC0824	VP9	2	0	1440.0	2560.0	1011/1349	None	24/25/30	2	1.1
HRC0825	H.265	2	medium	2160.0	3840.0	6550/8734	None	24/25/30	5	1.1
HRC0826	H.264	1	fast	2160.0	3840.0	24474/32633	None	24/25/30	5	1.1
HRC0827	H.265	1	ultrafast	540.0	960.0	487/650	None	24/25/30	2	1.1
HRC0828	H.265	2	medium	360.0	640.0	250/350	None	24/25/30	5	1.1
HRC0829	H.264	2	medium	480.0	854.0	300/400	None	24/25/30	2	1.1
HRC0830	VP9	2	2	720.0	1280.0	734/979	None	24/25/30	5	1.1
HRC0831	H.265	1	medium	2160.0	3840.0	14631/19508	None	24/25/30	2	1.1
HRC0832	VP9	2	2	1080.0	1920.0	691/922	None	24/25/30	2	1.1
HRC0833	H.265	1	medium	2160.0	3840.0	8718/11624	None	50/60	2	1.1
HRC0835	VP9	2	2	1440.0	2560.0	5889/7853	None	24/25/30	2	1.1
HRC0836	H.265	2	slower	480.0	854.0	2059/2746	None	24/25/30	2	1.1
HRC0837	H.265	2	slow	2160.0	3840.0	1602/2136	None	50/60	2	1.1
HRC0838	VP9	2	2	1440.0	2560.0	890/1187	None	24/25/30	2	1.1
HRC0839	VP9	2	4	2160.0	3840.0	9425/12567	None	24/25/30	5	1.1
HRC0840	VP9	2	4	1440.0	2560.0	4650/6200	None	50/60	5	1.1
HRC0841	H.264	2	medium	1440.0	2560.0	5988/7985	None	24/25/30	5	1.1
HRC0842	VP9	2	3	1080.0	1920.0	2786/3715	None	24/25/30	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.19: Test Plan for P2SVL06.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0843	H.264	2	ultrafast	480.0	854.0	1614/2152	None	24/25/30	2	1.1
HRC0846	VP9	2	2	2160.0	3840.0	14809/19746	None	50/60	2	1.1
HRC0847	H.265	2	medium	480.0	854.0	263/351	None	24/25/30	5	1.1
HRC0848	H.264	2	medium	540.0	960.0	1664/2219	None	24/25/30	2	1.1
HRC0849	H.265	2	medium	1080.0	1920.0	633/844	None	24/25/30	2	1.1
HRC0850	H.265	2	medium	540.0	960.0	524/699	None	24/25/30	5	1.1
HRC0851	H.265	1	fast	720.0	1280.0	5445/7260	None	24/25/30	2	1.1
HRC0852	H.265	2	medium	720.0	1280.0	4932/6576	None	50/60	2	1.1
HRC0853	VP9	2	1	2160.0	3840.0	12942/17257	None	50/60	5	1.1
HRC0855	H.265	1	medium	1440.0	2560.0	11111/14815	None	24/25/30	2	1.1
HRC0856	H.265	2	medium	1440.0	2560.0	4852/6470	None	24/25/30	2	1.1
HRC0857	VP9	2	2	480.0	854.0	498/664	None	24/25/30	5	1.1
HRC0858	H.265	2	medium	1440.0	2560.0	4590/6121	None	50/60	5	1.1
HRC0859	H.265	2	medium	1080.0	1920.0	3657/4877	None	50/60	2	1.1
HRC0860	H.265	2	medium	720.0	1280.0	1610/2147	None	50/60	5	1.1
HRC0861	VP9	2	2	1440.0	2560.0	1411/1882	None	24/25/30	5	1.1
HRC0863	H.264	2	medium	2160.0	3840.0	32393/43191	None	24/25/30	5	1.1
HRC0864	H.264	2	veryslow	1440.0	2560.0	9548/12731	None	24/25/30	5	1.1
HRC0865	VP9	2	3	1440.0	2560.0	8475/11300	None	24/25/30	5	1.1
HRC0866	VP9	1	1	1080.0	1920.0	829/1106	None	24/25/30	5	1.1
HRC0867	H.264	2	medium	1440.0	2560.0	1267/1690	None	24/25/30	2	1.1
HRC0868	VP9	2	2	360.0	640.0	1050/1400	None	24/25/30	5	1.1
HRC0869	H.265	2	medium	1440.0	2560.0	3291/4388	None	50/60	2	1.1
HRC0870	H.265	1	medium	720.0	1280.0	502/670	None	24/25/30	2	1.1
HRC0871	H.264	2	veryslow	480.0	854.0	228/304	None	24/25/30	2	1.1
HRC0872	H.264	2	medium	720.0	1280.0	2091/2788	None	24/25/30	2	1.1
HRC0873	VP9	1	2	2160.0	3840.0	10347/13797	None	24/25/30	2	1.1
HRC0874	VP9	1	3	480.0	854.0	2274/3032	None	24/25/30	2	1.1
HRC0875	H.265	2	medium	540.0	960.0	1819/2426	None	24/25/30	5	1.1
HRC0876	VP9	1	3	2160.0	3840.0	1967/2623	None	24/25/30	5	1.1
HRC0877	VP9	2	2	1080.0	1920.0	3408/4544	None	24/25/30	5	1.1
HRC0878	H.264	2	medium	720.0	1280.0	3027/4036	None	24/25/30	5	1.1
HRC0879	VP9	1	4	1440.0	2560.0	3420/4560	None	24/25/30	5	1.1
HRC0880	H.265	2	fast	2160.0	3840.0	1975/2634	None	24/25/30	2	1.1
HRC0881	VP9	2	4	720.0	1280.0	2928/3904	None	24/25/30	2	1.1
HRC0882	H.265	1	medium	2160.0	3840.0	20463/27285	None	50/60	5	1.1
HRC0883	VP9	1	2	2160.0	3840.0	2168/2891	None	24/25/30	2	1.1
HRC0884	H.264	2	medium	2160.0	3840.0	1938/2584	None	50/60	5	1.1
HRC0885	H.265	2	medium	1440.0	2560.0	9250/12334	None	24/25/30	5	1.1
HRC0886	H.265	2	medium	720.0	1280.0	1671/2228	None	24/25/30	5	1.1
HRC0887	H.265	2	medium	540.0	960.0	269/359	None	24/25/30	5	1.1

Table B.20: Test Plan for P2SVL07.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0888	H.264	2	medium	720.0	1280.0	750/1000	None	50/60	2	1.1
HRC0889	H.264	2	medium	360.0	640.0	127/170	None	24/25/30	5	1.1
HRC0890	H.264	2	medium	1440.0	2560.0	1133/1511	None	24/25/30	2	1.1
HRC0891	H.264	2	medium	1440.0	2560.0	1340/1787	None	24/25/30	5	1.1
HRC0892	H.265	2	medium	1440.0	2560.0	1614/2153	None	50/60	2	1.1
HRC0894	H.264	2	medium	1440.0	2560.0	1332/1777	None	24/25/30	2	1.1
HRC0895	H.265	2	medium	360.0	640.0	1000/1250	None	24/25/30	2	1.1
HRC0896	VP9	2	0	1440.0	2560.0	3793/5058	None	24/25/30	2	1.1
HRC0897	VP9	2	2	2160.0	3840.0	10874/14499	None	24/25/30	5	1.1
HRC0898	H.264	2	medium	1440.0	2560.0	6359/8479	None	24/25/30	2	1.1
HRC0899	H.265	2	medium	720.0	1280.0	418/558	None	24/25/30	5	1.1
HRC0900	VP9	1	4	1080.0	1920.0	1871/2495	None	24/25/30	2	1.1
HRC0901	VP9	2	2	1440.0	2560.0	889/1186	None	24/25/30	5	1.1
HRC0902	VP9	2	2	360.0	640.0	800/1000	None	24/25/30	5	1.1
HRC0903	H.265	2	medium	720.0	1280.0	984/1313	None	24/25/30	2	1.1
HRC0904	VP9	2	2	2160.0	3840.0	1572/2096	None	24/25/30	5	1.1
HRC0905	H.264	2	medium	360.0	640.0	750/1000	None	50/60	5	1.1
HRC0906	VP9	1	2	2160.0	3840.0	1917/2557	None	50/60	2	1.1
HRC0907	H.265	2	medium	360.0	640.0	150/201	None	15	2	1.1
HRC0908	H.265	2	medium	720.0	1280.0	639/853	None	24/25/30	5	1.1
HRC0909	H.265	1	medium	480.0	854.0	500/1000	None	24/25/30	5	1.1
HRC0910	VP9	2	2	360.0	640.0	507/676	None	24/25/30	2	1.1
HRC0911	VP9	2	2	480.0	854.0	400/800	None	24/25/30	2	1.1
HRC0912	H.264	1	ultrafast	720.0	1280.0	5673/7564	None	50/60	2	1.1
HRC0913	VP9	2	2	1440.0	2560.0	7851/10468	None	24/25/30	5	1.1
HRC0914	VP9	1	4	360.0	640.0	1057/1410	None	15	5	1.1
HRC0915	H.265	2	medium	720.0	1280.0	1777/2370	None	50/60	2	1.1
HRC0916	H.265	1	fast	2160.0	3840.0	2698/3598	None	24/25/30	5	1.1
HRC0917	H.265	2	medium	1080.0	1920.0	8736/11649	None	50/60	2	1.1
HRC0918	H.265	2	medium	360.0	640.0	474/633	None	24/25/30	2	1.1
HRC0920	H.264	1	medium	360.0	640.0	500/750	None	24/25/30	2	1.1
HRC0922	H.264	2	medium	720.0	1280.0	1288/1718	None	50/60	5	1.1
HRC0923	H.265	1	medium	360.0	640.0	297/397	None	24/25/30	2	1.1
HRC0924	VP9	2	2	1080.0	1920.0	1807/2410	None	24/25/30	2	1.1
HRC0925	H.265	2	medium	1080.0	1920.0	864/1153	None	50/60	2	1.1
HRC0926	H.265	1	medium	540.0	960.0	1087/1450	None	24/25/30	2	1.1
HRC0928	H.265	1	medium	2160.0	3840.0	2234/2979	None	24/25/30	2	1.1
HRC0929	VP9	2	2	360.0	640.0	500/750	None	24/25/30	2	1.1
HRC0930	H.264	1	medium	1080.0	1920.0	800/1200	None	24/25/30	2	1.1
HRC0932	VP9	2	2	1080.0	1920.0	3693/4925	None	24/25/30	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.21: Test Plan for P2SVL08.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0934	H.264	2	medium	1080.0	1920.0	8418/11225	None	50/60	5	1.1
HRC0935	H.264	2	medium	1440.0	2560.0	1185/1580	None	50/60	2	1.1
HRC0936	VP9	2	2	720.0	1280.0	1542/2057	None	24/25/30	5	1.1
HRC0937	VP9	2	1	360.0	640.0	132/177	None	24/25/30	2	1.1
HRC0939	VP9	2	2	2160.0	3840.0	1654/2206	None	24/25/30	5	1.1
HRC0940	H.264	2	medium	540.0	960.0	2423/3231	None	24/25/30	2	1.1
HRC0942	H.265	2	medium	540.0	960.0	1006/1342	None	24/25/30	2	1.1
HRC0943	H.264	2	medium	1440.0	2560.0	2745/3660	None	24/25/30	5	1.1
HRC0944	H.265	2	medium	1440.0	2560.0	6193/8258	None	24/25/30	5	1.1
HRC0945	H.265	2	medium	720.0	1280.0	2052/2736	None	50/60	5	1.1
HRC0946	H.265	2	slower	1080.0	1920.0	2944/3926	None	24/25/30	2	1.1
HRC0947	H.265	1	medium	480.0	854.0	496/662	None	24/25/30	5	1.1
HRC0948	H.264	2	ultrafast	2160.0	3840.0	17045/22727	None	24/25/30	2	1.1
HRC0950	VP9	2	2	1440.0	2560.0	1470/1961	None	24/25/30	2	1.1
HRC0951	VP9	1	2	720.0	1280.0	4389/5853	None	24/25/30	2	1.1
HRC0952	H.264	1	fast	360.0	640.0	712/950	None	24/25/30	2	1.1
HRC0953	H.264	2	medium	2160.0	3840.0	1336/1782	None	50/60	2	1.1
HRC0954	VP9	2	2	540.0	960.0	243/324	None	24/25/30	5	1.1
HRC0955	H.265	1	veryfast	2160.0	3840.0	12372/16497	None	24/25/30	2	1.1
HRC0956	H.264	1	medium	480.0	854.0	1370/1827	None	24/25/30	5	1.1
HRC0958	H.264	2	medium	1080.0	1920.0	10303/13738	None	24/25/30	5	1.1
HRC0959	VP9	2	2	2160.0	3840.0	3373/4498	None	24/25/30	2	1.1
HRC0960	VP9	1	2	1080.0	1920.0	7782/10376	None	24/25/30	2	1.1
HRC0962	VP9	2	2	2160.0	3840.0	12686/16915	None	24/25/30	5	1.1
HRC0964	VP9	2	2	1080.0	1920.0	6566/8755	None	24/25/30	5	1.1
HRC0965	H.264	2	ultrafast	720.0	1280.0	1244/1659	None	24/25/30	2	1.1
HRC0966	H.265	1	ultrafast	1440.0	2560.0	4479/5972	None	24/25/30	2	1.1
HRC0967	H.265	2	medium	1440.0	2560.0	1771/2362	None	24/25/30	5	1.1
HRC0968	H.265	1	medium	1440.0	2560.0	1157/1543	None	24/25/30	5	1.1
HRC0970	VP9	2	0	480.0	854.0	318/424	None	24/25/30	2	1.1
HRC0971	H.265	2	veryfast	480.0	854.0	426/569	None	24/25/30	5	1.1
HRC0972	VP9	1	3	2160.0	3840.0	7815/10420	None	24/25/30	2	1.1
HRC0974	VP9	2	2	1440.0	2560.0	10058/13411	None	24/25/30	5	1.1
HRC0975	H.264	2	medium	360.0	640.0	739/986	None	15	2	1.1
HRC0976	H.265	2	fast	540.0	960.0	1089/1453	None	24/25/30	2	1.1

Table B.22: Test Plan for P2SVL09.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC0978	H.265	1	medium	480.0	854.0	883/1178	None	24/25/30	5	1.1
HRC0979	H.264	2	medium	480.0	854.0	375/500	None	24/25/30	2	1.1
HRC0980	VP9	2	4	720.0	1280.0	924/1232	None	24/25/30	5	1.1
HRC0981	H.265	2	medium	1080.0	1920.0	2746/3662	None	24/25/30	5	1.1
HRC0982	H.265	2	medium	1080.0	1920.0	5974/7966	None	50/60	2	1.1
HRC0983	VP9	2	2	540.0	960.0	1805/2407	None	24/25/30	5	1.1
HRC0984	VP9	1	3	1080.0	1920.0	7984/10646	None	50/60	2	1.1
HRC0985	VP9	2	0	1440.0	2560.0	3626/4835	None	50/60	5	1.1
HRC0986	H.264	2	medium	360.0	640.0	1111/1482	None	24/25/30	2	1.1
HRC0988	H.265	2	fast	360.0	640.0	604/806	None	24/25/30	2	1.1
HRC0989	VP9	2	2	1080.0	1920.0	5619/7492	None	50/60	5	1.1
HRC0990	H.265	2	medium	360.0	640.0	432/577	None	24/25/30	5	1.1
HRC0991	VP9	2	1	480.0	854.0	2394/3193	None	24/25/30	5	1.1
HRC0992	H.265	1	medium	1440.0	2560.0	4305/5741	None	24/25/30	5	1.1
HRC0993	H.264	2	medium	1440.0	2560.0	7705/10274	None	24/25/30	2	1.1
HRC0994	H.264	2	ultrafast	2160.0	3840.0	2112/2816	None	24/25/30	2	1.1
HRC0995	VP9	2	3	480.0	854.0	2455/3274	None	24/25/30	2	1.1
HRC0996	H.264	2	slow	2160.0	3840.0	1915/2554	None	24/25/30	5	1.1
HRC0997	H.264	2	medium	1440.0	2560.0	10200/13600	None	24/25/30	2	1.1
HRC0998	VP9	2	1	540.0	960.0	571/762	None	24/25/30	2	1.1
HRC0999	VP9	2	2	1080.0	1920.0	543/724	None	24/25/30	5	1.1
HRC1000	H.264	2	ultrafast	1080.0	1920.0	8303/11071	None	24/25/30	2	1.1
HRC1003	H.265	2	medium	2160.0	3840.0	2964/3952	None	24/25/30	5	1.1
HRC1004	VP9	2	2	360.0	640.0	440/587	None	24/25/30	2	1.1
HRC1006	H.265	1	veryfast	2160.0	3840.0	7800/10400	None	24/25/30	2	1.1
HRC1007	VP9	2	2	1440.0	2560.0	10332/13776	None	24/25/30	5	1.1
HRC1008	H.264	2	medium	360.0	640.0	747/997	None	24/25/30	5	1.1
HRC1011	VP9	2	2	1440.0	2560.0	3395/4527	None	24/25/30	2	1.1
HRC1012	H.265	1	medium	360.0	640.0	1098/1465	None	24/25/30	5	1.1
HRC1013	H.265	1	medium	540.0	960.0	1008/1344	None	24/25/30	5	1.1
HRC1014	VP9	1	3	360.0	640.0	1077/1437	None	24/25/30	5	1.1
HRC1015	H.264	1	medium	2160.0	3840.0	13608/18145	None	24/25/30	5	1.1
HRC1016	H.265	2	fast	720.0	1280.0	413/551	None	24/25/30	2	1.1
HRC1017	H.265	2	medium	2160.0	3840.0	1262/1683	None	24/25/30	2	1.1
HRC1018	H.264	2	medium	1080.0	1920.0	3297/4397	None	50/60	2	1.1
HRC1019	H.264	2	medium	1440.0	2560.0	8213/10951	None	24/25/30	5	1.1
HRC1020	H.265	2	medium	480.0	854.0	753/1004	None	15	2	1.1
HRC1021	H.265	1	medium	1440.0	2560.0	1723/2298	None	50/60	2	1.1
HRC1022	H.265	1	medium	1080.0	1920.0	3222/4297	None	24/25/30	5	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.23: Test Plan for P2SVL10.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC1023	H.264	2	ultrafast	360.0	640.0	345/461	None	24/25/30	2	1.1
HRC1024	H.265	2	medium	1080.0	1920.0	2542/3390	None	50/60	2	1.1
HRC1025	VP9	2	2	540.0	960.0	1102/1470	None	24/25/30	5	1.1
HRC1026	H.265	2	medium	480.0	854.0	576/768	None	24/25/30	5	1.1
HRC1027	H.264	2	ultrafast	1440.0	2560.0	2128/2838	None	24/25/30	2	1.1
HRC1028	H.264	2	medium	540.0	960.0	543/724	None	24/25/30	5	1.1
HRC1029	H.264	2	medium	1440.0	2560.0	8398/11198	None	50/60	2	1.1
HRC1031	H.264	1	veryfast	1080.0	1920.0	5010/6680	None	50/60	5	1.1
HRC1032	H.265	2	medium	2160.0	3840.0	8085/10780	None	50/60	5	1.1
HRC1033	H.265	1	medium	480.0	854.0	688/918	None	24/25/30	2	1.1
HRC1034	H.265	2	medium	1440.0	2560.0	1397/1863	None	24/25/30	5	1.1
HRC1035	H.265	2	medium	1080.0	1920.0	7932/10577	None	24/25/30	2	1.1
HRC1036	H.265	2	medium	2160.0	3840.0	9644/12859	None	24/25/30	2	1.1
HRC1037	H.264	1	medium	360.0	640.0	759/1013	None	24/25/30	5	1.1
HRC1038	VP9	1	2	720.0	1280.0	1795/2394	None	24/25/30	5	1.1
HRC1040	H.264	2	medium	1080.0	1920.0	510/681	None	24/25/30	5	1.1
HRC1041	H.264	2	medium	1440.0	2560.0	3259/4346	None	24/25/30	2	1.1
HRC1042	H.264	2	slower	720.0	1080.0	1401/1869	None	24/25/30	5	1.1
HRC1043	VP9	2	2	720.0	1280.0	5918/7891	None	24/25/30	2	1.1
HRC1044	H.264	2	medium	2160.0	3840.0	1163/1551	None	50/60	5	1.1
HRC1045	VP9	2	2	1440.0	2560.0	1089/1453	None	50/60	5	1.1
HRC1046	VP9	2	4	1080.0	1920.0	1450/1820	None	24/25/30	2	1.1
HRC1047	H.264	2	ultrafast	2160.0	3840.0	22827/30437	None	24/25/30	2	1.1
HRC1048	H.264	2	medium	1440.0	2560.0	10092/13457	None	24/25/30	5	1.1
HRC1049	H.264	1	medium	1440.0	2560.0	4928/6571	None	24/25/30	2	1.1
HRC1050	H.264	2	medium	540.0	960.0	900/1350	None	24/25/30	5	1.1
HRC1051	VP9	2	2	720.0	1280.0	5448/7265	None	24/25/30	2	1.1
HRC1052	VP9	1	2	2160.0	3840.0	1401/1869	None	50/60	2	1.1
HRC1053	H.264	2	slower	1080.0	1920.0	7976/10635	None	50/60	2	1.1
HRC1054	VP9	2	2	720.0	1280.0	5225/6967	None	24/25/30	5	1.1
HRC1056	VP9	2	1	2160.0	3840.0	1205/1607	None	50/60	2	1.1
HRC1057	VP9	2	2	1080.0	1920.0	8147/10863	None	24/25/30	2	1.1
HRC1058	H.264	2	medium	720.0	1280.0	4370/5827	None	50/60	2	1.1
HRC1059	VP9	2	2	1080.0	1920.0	5648/7531	None	24/25/30	2	1.1
HRC1060	H.264	2	medium	480.0	854.0	2580/3441	None	24/25/30	2	1.1
HRC1061	VP9	2	2	1080.0	1920.0	7151/9535	None	24/25/30	5	1.1
HRC1062	H.265	2	medium	1440.0	2560.0	998/1331	None	50/60	2	1.1
HRC1063	H.265	2	medium	2160.0	3840.0	10007/13343	None	24/25/30	5	1.1
HRC1064	H.264	2	ultrafast	540.0	960.0	1728/2304	None	24/25/30	5	1.1
HRC1065	VP9	2	3	1080.0	1920.0	6036/8048	None	50/60	2	1.1
HRC1066	H.264	2	ultrafast	720.0	1280.0	2947/3930	None	24/25/30	5	1.1
HRC1067	VP9	2	2	1080.0	1920.0	1334/1779	None	24/25/30	5	1.1

Table B.24: Test Plan for P2SVL11.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC1068	VP9	2	2	480.0	854.0	952/1270	None	24/25/30	2	1.1
HRC1069	H.265	2	slow	1080.0	1920.0	509/679	None	24/25/30	5	1.1
HRC1070	VP9	1	2	1080.0	1920.0	6713/8951	None	50/60	2	1.1
HRC1071	H.264	2	medium	2160.0	3840.0	2109/2813	None	24/25/30	2	1.1
HRC1072	VP9	2	2	720.0	1280.0	5160/6880	None	24/25/30	2	1.1
HRC1073	VP9	2	2	1080.0	1920.0	6006/8009	None	24/25/30	5	1.1
HRC1074	H.265	2	ultrafast	2160.0	3840.0	18622/24830	None	24/25/30	2	1.1
HRC1075	H.265	2	medium	1440.0	2560.0	1898/2531	None	24/25/30	5	1.1
HRC1076	H.265	2	medium	360.0	640.0	143/191	None	24/25/30	5	1.1
HRC1077	VP9	2	2	1440.0	2560.0	3255/4340	None	50/60	5	1.1
HRC1078	VP9	2	2	540.0	960.0	585/781	None	24/25/30	5	1.1
HRC1079	VP9	2	2	540.0	960.0	1145/1527	None	24/25/30	5	1.1
HRC1080	H.264	2	medium	540.0	960.0	345/460	None	24/25/30	2	1.1
HRC1081	H.264	2	fast	1440.0	2560.0	1428/1905	None	24/25/30	5	1.1
HRC1082	VP9	2	2	1440.0	2560.0	2928/3904	None	24/25/30	2	1.1
HRC1083	H.265	2	medium	480.0	854.0	691/922	None	24/25/30	2	1.1
HRC1084	H.265	2	medium	1440.0	2560.0	4158/5545	None	24/25/30	2	1.1
HRC1086	H.265	2	medium	1440.0	2560.0	2295/3060	None	24/25/30	2	1.1
HRC1087	VP9	2	2	360.0	640.0	575/767	None	24/25/30	2	1.1
HRC1088	VP9	2	2	1080.0	1920.0	2229/2972	None	24/25/30	5	1.1
HRC1089	H.265	2	fast	540.0	960.0	1473/1965	None	24/25/30	5	1.1
HRC1090	VP9	2	4	2160.0	3840.0	7009/9346	None	24/25/30	2	1.1
HRC1091	H.264	2	medium	1080.0	1920.0	468/624	None	24/25/30	2	1.1
HRC1092	H.264	2	medium	480.0	854.0	1524/2033	None	24/25/30	2	1.1
HRC1093	H.265	2	medium	1440.0	2560.0	4860/6480	None	24/25/30	5	1.1
HRC1095	H.264	2	medium	720.0	1280.0	702/937	None	24/25/30	5	1.1
HRC1096	H.265	1	medium	1440.0	2560.0	2675/3567	None	24/25/30	5	1.1
HRC1097	H.265	2	medium	1440.0	2560.0	1371/1828	None	50/60	2	1.1
HRC1098	H.264	2	medium	480.0	854.0	281/375	None	24/25/30	5	1.1
HRC1099	H.264	2	veryfast	540.0	960.0	1057/1410	None	24/25/30	5	1.1
HRC1100	H.265	2	medium	1080.0	1920.0	3735/4981	None	24/25/30	2	1.1
HRC1101	H.264	2	medium	2160.0	3840.0	19722/26297	None	24/25/30	2	1.1
HRC1104	H.264	2	medium	1440.0	2560.0	11344/15126	None	50/60	5	1.1
HRC1105	H.265	1	medium	2160.0	3840.0	1331/1775	None	24/25/30	2	1.1
HRC1106	H.265	1	medium	2160.0	3840.0	7209/9613	None	24/25/30	2	1.1
HRC1107	VP9	2	2	1080.0	1920.0	5961/7949	None	50/60	5	1.1
HRC1108	H.264	2	medium	1440.0	2560.0	7175/9567	None	24/25/30	2	1.1
HRC1109	H.264	2	slower	1080.0	1920.0	744/993	None	50/60	5	1.1
HRC1110	H.264	2	medium	1440.0	2560.0	6714/8953	None	24/25/30	5	1.1
HRC1111	H.264	1	medium	2160.0	3840.0	12276/16369	None	24/25/30	5	1.1
HRC1112	H.264	1	medium	1080.0	1920.0	3848/5131	None	50/60	2	1.1

Appendix B P.NATS Phase 2 Test Plan

Table B.25: Test Plan for P2SVL12.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0001	H.264	2	medium	240.0	426.0	100/200	None	15	5	2.0
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC1113	H.265	2	slow	1440.0	2560.0	10665/14220	None	24/25/30	5	1.1
HRC1114	H.264	2	medium	720.0	1280.0	6254/8339	None	24/25/30	2	1.1
HRC1115	VP9	2	2	360.0	640.0	773/1031	None	24/25/30	5	1.1
HRC1116	H.264	1	medium	540.0	960.0	578/771	None	24/25/30	5	1.1
HRC1118	H.265	2	medium	480.0	854.0	681/908	None	15	2	1.1
HRC1119	VP9	2	3	360.0	640.0	848/1131	None	24/25/30	5	1.1
HRC1120	H.265	2	medium	360.0	640.0	965/1287	None	24/25/30	5	1.1
HRC1122	VP9	1	2	540.0	960.0	425/567	None	24/25/30	2	1.1
HRC1123	VP9	2	2	360.0	640.0	815/1087	None	24/25/30	2	1.1
HRC1124	H.265	1	medium	1080.0	1920.0	486/648	None	24/25/30	2	1.1
HRC1125	VP9	1	2	1440.0	2560.0	1375/1834	None	50/60	5	1.1
HRC1126	VP9	2	2	1080.0	1920.0	534/712	None	24/25/30	2	1.1
HRC1127	H.264	2	medium	480.0	854.0	1398/1864	None	24/25/30	5	1.1
HRC1128	H.264	2	medium	360.0	640.0	879/1172	None	24/25/30	2	1.1
HRC1129	H.264	2	medium	720.0	1280.0	1869/2492	None	24/25/30	5	1.1
HRC1130	H.265	1	veryslow	480.0	854.0	611/815	None	24/25/30	5	1.1
HRC1132	VP9	1	2	240.0	426.0	630/840	None	24/25/30	2	1.1
HRC1133	H.265	2	medium	240.0	426.0	339/452	None	15	2	1.1
HRC1134	H.264	2	slow	360.0	640.0	759/1012	None	24/25/30	2	1.1
HRC1135	H.264	1	medium	540.0	960.0	1434/1912	None	15	5	1.1
HRC1136	VP9	1	0	540.0	960.0	1540/2054	None	24/25/30	5	1.1
HRC1137	H.265	2	ultrafast	360.0	640.0	582/777	None	24/25/30	2	1.1
HRC1138	VP9	2	2	480.0	854.0	1003/1338	None	24/25/30	2	1.1
HRC1140	H.265	2	medium	1440.0	2560.0	8639/11519	None	24/25/30	5	1.1
HRC1141	H.265	2	medium	720.0	1280.0	4853/6471	None	24/25/30	2	1.1
HRC1142	H.264	2	medium	720.0	1280.0	4485/5980	None	24/25/30	2	1.1
HRC1143	VP9	2	3	720.0	1280.0	1574/2099	None	24/25/30	2	1.1
HRC1144	H.264	2	medium	540.0	960.0	394/526	None	24/25/30	5	1.1
HRC1145	VP9	2	0	1080.0	1920.0	4716/6289	None	50/60	2	1.1
HRC1146	VP9	2	2	720.0	1280.0	618/825	None	50/60	2	1.1
HRC1147	H.264	2	medium	240.0	426.0	341/455	None	24/25/30	5	1.1
HRC1149	H.265	2	medium	240.0	426.0	192/257	None	24/25/30	2	1.1
HRC1150	H.264	2	medium	1440.0	2560.0	14520/19360	None	24/25/30	5	1.1
HRC1152	H.264	2	medium	1440.0	2560.0	1647/2196	None	24/25/30	2	1.1
HRC1153	H.265	2	medium	480.0	854.0	795/1061	None	15	2	1.1
HRC1155	H.264	2	veryfast	360.0	640.0	636/849	None	24/25/30	2	1.1
HRC1156	H.265	2	medium	1440.0	2560.0	4984/6646	None	50/60	5	1.1
HRC1157	H.265	2	medium	1440.0	2560.0	834/1113	None	24/25/30	5	1.1

Table B.26: Test Plan for P2SVL13.

HRC	Encoder	Passes	Preset	Height	Width	Bitrate	CRF	Framerate	iFI	MRF
HRC0115	H.264	2	medium	360.0	640.0	300/500	None	24/25/30	5	2.0
HRC0388	H.264	2	medium	720.0	1280.0	800/1600	None	50/60	2	2.0
HRC0436	H.264	2	medium	1080.0	1920.0	3500/7000	None	50/60	2	2.0
HRC0484	H.264	2	medium	1440.0	2560.0	6000/10000	None	50/60	2	2.0
HRC0571	H.264	2	medium	2160.0	3840.0	30000/45000	None	50/60	2	2.0
HRC1158	VP9	2	2	1440.0	2560.0	4459/5946	None	50/60	5	1.1
HRC1159	VP9	2	2	1080.0	1920.0	1802/2403	None	50/60	5	1.1
HRC1161	VP9	1	2	540.0	960.0	540/720	None	24/25/30	2	1.1
HRC1162	H.265	2	medium	2160.0	3840.0	22498/29998	None	24/25/30	2	1.1
HRC1163	VP9	2	2	1440.0	2560.0	1117/1490	None	24/25/30	5	1.1
HRC1164	H.264	2	medium	1080.0	1920.0	1249/1666	None	24/25/30	2	1.1
HRC1165	H.264	2	medium	360.0	640.0	911/1215	None	24/25/30	5	1.1
HRC1168	H.265	2	medium	480.0	854.0	256/342	None	24/25/30	2	1.1
HRC1169	H.264	2	medium	1440.0	2560.0	5754/7672	None	24/25/30	2	1.1
HRC1170	H.265	2	ultrafast	540.0	960.0	882/1176	None	24/25/30	2	1.1
HRC1171	H.264	1	medium	1080.0	1920.0	4340/5787	None	50/60	2	1.1
HRC1172	VP9	2	2	480.0	854.0	403/538	None	24/25/30	2	1.1
HRC1174	VP9	2	2	1440.0	2560.0	865/1154	None	24/25/30	5	1.1
HRC1175	H.265	2	medium	2160.0	3840.0	6765/9020	None	24/25/30	2	1.1
HRC1176	VP9	2	2	1440.0	2560.0	2638/3518	None	24/25/30	2	1.1
HRC1177	H.264	2	medium	480.0	854.0	2388/3185	None	24/25/30	2	1.1
HRC1178	H.264	2	medium	480.0	854.0	1372/1830	None	24/25/30	5	1.1
HRC1179	H.265	1	ultrafast	1080.0	1920.0	507/677	None	24/25/30	5	1.1
HRC1180	VP9	2	4	480.0	854.0	773/1031	None	24/25/30	2	1.1
HRC1181	H.264	1	medium	1080.0	1920.0	9740/12987	None	24/25/30	2	1.1
HRC1182	H.265	2	medium	1440.0	2560.0	1914/2552	None	24/25/30	2	1.1
HRC1183	H.264	1	medium	360.0	640.0	516/689	None	24/25/30	2	1.1
HRC1184	H.264	2	ultrafast	2160.0	3840.0	2159/2879	None	50/60	5	1.1
HRC1186	H.264	2	medium	2160.0	3840.0	10120/13494	None	24/25/30	2	1.1
HRC1187	H.264	2	ultrafast	2160.0	3840.0	12110/16147	None	24/25/30	2	1.1
HRC1188	H.265	2	medium	1080.0	1920.0	8484/11313	None	24/25/30	2	1.1
HRC1189	VP9	1	2	540.0	960.0	704/939	None	24/25/30	5	1.1
HRC1190	H.264	1	medium	2160.0	3840.0	1545/2060	None	50/60	2	1.1
HRC1191	H.264	1	medium	720.0	1280.0	1934/2579	None	24/25/30	5	1.1
HRC1192	H.265	1	medium	1440.0	2560.0	11016/14689	None	24/25/30	2	1.1
HRC1193	H.264	2	medium	720.0	1280.0	5822/7763	None	50/60	2	1.1
HRC1194	H.264	2	medium	360.0	640.0	927/1236	None	24/25/30	5	1.1
HRC1195	H.264	2	medium	720.0	1280.0	2313/3085	None	24/25/30	2	1.1
HRC1196	VP9	1	2	1440.0	2560.0	871/1162	None	24/25/30	5	1.1
HRC1197	VP9	2	2	2160.0	3840.0	2087/2783	None	24/25/30	2	1.1
HRC1198	H.265	1	medium	1080.0	1920.0	5951/7935	None	24/25/30	2	1.1
HRC1199	H.265	2	medium	1080.0	1920.0	566/755	None	24/25/30	5	1.1
HRC1200	VP9	2	2	540.0	960.0	2437/3250	None	24/25/30	2	1.1
HRC1201	H.264	1	medium	2160.0	3840.0	1733/2311	None	24/25/30	5	1.1
HRC1202	H.264	2	medium	2160.0	3840.0	11521/15362	None	50/60	5	1.1

AVQBits Helper Functions

C.1 RfromMOS (5-Point MOS Scale to 100-Point Scale)

This transformation is based on the E-model [ITU09]. If MOS represents the quality score expressed in a 5-point MOS scale and R the quality score expressed in a 100-point scale, then *RfromMOS* is calculated as follows:

```

procedure RFROMMOS(MOS)
  x = (18566 - 6750 * MOS)
  if MOS > 4.5 then
    MOS = 4.5
  end
  if x < 0 then
    num = 15 × √(-903522 + 1113960 × MOS - 202500 × MOS × MOS)
    den = 6750 × MOS - 18566
    fra =  $\frac{num}{den}$ 
    h =  $\frac{pi - arctan(fra)}{3}$ 
  end
  else
    num = 15 × √(-903522 + 1113960 × MOS - 202500 × MOS × MOS)
    den = 18566 - 6750 × MOS
    fra =  $\frac{num}{den}$ 
    h =  $\frac{arctan(fra)}{3}$ 
  end
  R = 20.0 ×  $\frac{(8 - \sqrt{(226) * \cos(h + pi/3)})}{3}$ 
end procedure

```

Algorithm 1: Algorithm for RfromMOS calculation

C.2 MOSfromR (100-Point Scale to 5-Point MOS Scale)

This section describes the algorithm for transforming the quality scores from a 100-point scale to 5-point MOS scale. Like the *RfromMOS* transformation, this is also based on the E-model [ITU09].

```
procedure MOSFROMR(R)
  MOS_MAX = 4.5
  MOS_MIN = 1.0
  if MOS > 4.5 then
|   MOS = MOS_MAX
  end
  if MOS ≤ 0 then
|   MOS = MOS_MIN
  end
  MOS = MOS_MIN + ((MOS_MAX – MOS_MIN) × R/100) + R × (R –
60) × (100 – R) × 0.000007
end procedure
```

Algorithm 2: Algorithm for MOSfromR calculation

C.3 Scalet5 (4.5-point scale of MOSfromR to 5-point scale)

This section describes the final full 5-point scale range transformation that is done in all *AVQBits* model instances. If $MOS_{[1,4.5]}$ denotes the model prediction on a 4.5-point scale, then the final transformation to the 5-point scale is done as follows:

C.3 Scalet5 (4.5-point scale of MOSfromR to 5-point scale)

```
procedure SCALETO5( $MOS_{[1,4.5]}$ )  
   $input\_start = 1.0$   
   $input\_end = 4.5$   
   $outout\_start = 1.0$   
   $outout\_end = 5$   
  if  $MOS_{[1,4.5]} > 4.5$  then  
|    $MOS = 5$   
  end  
  else  
|    $MOS = output\_start + ((output\_end - output\_start) / (input\_end -$   
|    $input\_start)) \times (x - input\_start)$   
  end  
end procedure
```

Algorithm 3: Algorithm for Scalet5 calculation

Bibliography

- [AG17] Abhilash Antony and Sreelekha G. “HEVC-Based Lossless Intra Coding for Efficient Still Image Compression”. In: *Multimedia Tools Appl.* 76.2 (Jan. 2017), pp. 1639–1658. ISSN: 1380-7501.
- [All75] John Allnatt. “Subjective rating and apparent magnitude”. In: *International Journal of Man-Machine Studies* 7.6 (1975), pp. 801–816.
- [AWS19] AWS. *Video Latency in Live Streaming*. 2019. URL: <https://aws.amazon.com/media/tech/video-latency-in-live-streaming/> (visited on 12/06/2019).
- [Bae+13] S. Bae, J. Kim, M. Kim, S. Cho, and J. S. Choi. “Assessments of Subjective Video Quality on HEVC-Encoded 4K-UHD Video for Beyond-HDTV Broadcasting Services”. In: *IEEE Trans. on Broadcasting* 59.2 (June 2013), pp. 209–222. ISSN: 0018-9316.
- [Bam+17] Christos G Bampis, Praful Gupta, Rajiv Soundararajan, and Alan C Bovik. “SpEED-QA: Spatial efficient entropic differencing for image and video quality”. In: *IEEE signal processing letters* 24.9 (2017), pp. 1333–1337.
- [Bam+18] Christos G. Bampis, Zhi Li, Ioannis Katsavounidis, Te-Yuan Huang, Chaitanya Ekanadham, and Alan C. Bovik. *Towards Perceptually Optimized End-to-end Adaptive Video Streaming*. 2018. arXiv: 1808.03898 [eess.IV].
- [Bar+18] Nabajeet Barman, Saman Zadtootaghaj, Steven Schmidt, Maria G Martini, and Sebastian Möller. “GamingVideoSET: a dataset for gaming video streaming applications”. In: *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*. IEEE. 2018, pp. 1–6.

Bibliography

- [Bar+19] Nabajeet Barman, Emmanuel Jammeh, Seyed Ali Ghorashi, and Maria G Martini. “No-reference video quality estimation based on machine learning for passive gaming video streaming applications”. In: *IEEE Access* 7 (2019), pp. 74511–74527.
- [Ben13] Christopher Benitez. *YouTube, changing the way of delivering videos: Chunking and Adaptive Streaming are In, Progressive Download is Out!* 2013. URL: <https://www.netmanias.com/en/?m=view&id=blog&no=5923&xtag=google-http-adaptive-streaming-iptv-video-streaming-youtube&xref=youtube-changing-the-way-of-delivering-videos-chunking-and-adaptive-streaming-are-in-progressive-download-is-out> (visited on 10/04/2013).
- [Ber+15] K. Berger, Y. Koudota, M. Barkowsky, and P. Le Callet. “Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains”. In: *7th Int. Workshop on Quality of Multimedia Experience (QoMEX)*. May 2015, pp. 1–6.
- [Bit21] Bitmovin. *Bitmovin Video Developer Report, 2021*. 2021. URL: <https://bitmovin.com/video-dev-report/>.
- [Ble] Blender Foundation. *Bick Buck Bunny Distribution*. URL: <http://distribution.bbb3d.renderfarming.net/video/png>.
- [Blo15] Netflix Technology Blog. *Per-Title Encode Optimization*. 2015. URL: <https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2> (visited on 12/14/2015).
- [BM20] Nabajeet Barman and Maria G. Martini. “An Evaluation of the Next-Generation Image Coding Standard AVIF”. In: *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 2020, pp. 1–4.
- [Bos+16] S. Bosse, M. Siekmann, J. Rasch, T. Wiegand, and W. Samek. “Quality assessment of image patches distorted by image compression using crowdsourcing”. In: *2016 IEEE ICME*. 2016, pp. 1–6.

- [Bra+23] Florian Braun, **Rakesh Rao Ramachandra Rao**, Werner Robitza, and Alexander Raake. “Automatic Audiovisual Asynchrony Measurement for Quality Assessment of Videoconferencing”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023.
- [Bro+21] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. “Overview of the Versatile Video Coding (VVC) Standard and its Applications”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.10 (2021), pp. 3736–3764.
- [Bru+13] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Doods, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. “Qualinet white paper on definitions of quality of experience”. In: (2013).
- [Cha21] Stream Charts. *All streaming data in one place*. 2021. URL: <https://streamscharts.com/>.
- [Che+18] Yue Chen, Debargha Murherjee, Jingning Han, Adrian Grange, Yaowu Xu, Zoe Liu, Sarah Parker, Cheng Chen, Hui Su, Urvang Joshi, Ching-Han Chiang, Yunqing Wang, Paul Wilkins, Jim Bankoski, Luc Trudeau, Nathan Egge, Jean-Marc Valin, Thomas Davies, Steinar Midtskogen, Andrey Norkin, and Peter de Rivaz. “An Overview of Core Coding Tools in the AV1 Video Codec”. In: *2018 Picture Coding Symposium (PCS)*. 2018, pp. 41–45.
- [Chi+11] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J. Karam. “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison”. In: *IEEE Transactions on Broadcasting* 57.2 (2011), pp. 165–182.
- [Cis] Cisco. *About the Common Media Application Format with HTTP Live Streaming (HLS)*. URL: https://developer.apple.com/documentation/http_live_streaming/about_the_common_media_application_format_with_http_live_streaming_hls.

Bibliography

- [Cis22] Cisco. *Cisco Visual Networking Index: Forecast and Trends, 2017–2022*. 2022. URL: <https://twiki.cern.ch/twiki/pub/HEPIX/TechwatchNetwork/HtwNetworkDocuments/white-paper-c11-741490.pdf>.
- [CL14] M. Cheon and J. Lee. “Objective Quality Comparison of 4K UHD and Up-Scaled 4K UHD Videos”. In: *IEEE Int. Symp. on Multimedia*. Dec. 2014, pp. 78–81.
- [CL18] M. Cheon and J. Lee. “Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience”. In: *IEEE Trans. on Circuits and Systems for Video Technology* 28.7 (July 2018), pp. 1467–1480. ISSN: 1051-8215.
- [Cro+19] Simone Croci, Cagri Ozcinar, Emin Zerman, Julián Cabrera, and Aljosa Smolic. “Voronoi-based Objective Quality Metrics for Omnidirectional Video”. In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 2019, pp. 1–6.
- [CWH16] Robert C. Streijl, Stefan Winkler, and David Hands. “Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives”. In: *Multimedia Systems* 22 (Mar. 2016), pp. 213–227.
- [Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [DG17] Edip Demirbilek and Jean-Charles Grégoire. “Machine learning based reduced reference bitstream audiovisual quality prediction models for realtime communications”. In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. 2017, pp. 571–576.
- [Dia+23] Chenyao Diao, Luljeta Sinani, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Revisiting Videoconferencing QoE: Impact of Network Delay and Resolution as Factors for Social Cue Perceptibility”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023.

- [DMW17] Zhengfang Duanmu, Kede Ma, and Zhou Wang. “Quality-of-Experience of Adaptive Video Streaming: Exploring the Space of Adaptations”. In: *Proceedings of the 25th ACM International Conference on Multimedia*. MM ’17. Mountain View, California, USA: Association for Computing Machinery, 2017, pp. 1752–1760. ISBN: 9781450349062.
- [DRW18] Zhengfang Duanmu, Abdul Rehman, and Zhou Wang. “A Quality-of-Experience Database for Adaptive Video Streaming”. In: *IEEE Transactions on Broadcasting* 64.2 (2018), pp. 474–487.
- [Dua+17] Zhengfang Duanmu, Kai Zeng, Kede Ma, Abdul Rehman, and Zhou Wang. “A Quality-of-Experience Index for Streaming Video”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.1 (2017), pp. 154–166.
- [FAK13] Ó. Figuerola Salas, V. Adzic, and H. Kalva. “Subjective quality evaluations using crowdsourcing”. In: *2013 PCS*. 2013.
- [Far+11] M. C. Q. Farias, M. M. Carvalho, H. T. M. Kussaba, and B. H. A. Noronha. “A hybrid metric for digital video quality assessment”. In: *2011 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 2011, pp. 1–6.
- [Fre+18] Stephan Fremerey, Ashutosh Singla, Kay Meseberg, and Alexander Raake. “AVtrack360: An Open Dataset and Software Recording People’s Head Rotations Watching 360° Videos on an HMD”. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. MMSys ’18. Amsterdam, Netherlands: Association for Computing Machinery, 2018, pp. 403–408. ISBN: 9781450351928.
- [Fre+20] Stephan Fremerey, Steve Göring, **Rao Rakesh Ramachandra Rao**, Rachel Huang, and Alexander Raake. “Subjective Test Dataset and Meta-data-based Models for 360° Streaming Video Quality”. In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020.
- [Gar+13] M. -. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake. “Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2”. In: *2013 IEEE 15th*

Bibliography

- International Workshop on Multimedia Signal Processing (MMSP)*. Sept. 2013, pp. 482–487.
- [GB16a] D. Ghadiyaram and A. C. Bovik. “Massive Online Crowdsourced Study of Subjective and Objective Picture Quality”. In: *IEEE Transactions on Image Processing* 25.1 (2016).
- [GB16b] Deepti Ghadiyaram and Alan C. Bovik. *Perceptual Quality Prediction on Authentically Distorted Images Using a Bag of Features Approach*. 2016. arXiv: 1609.04757 [cs.CV].
- [Gha+18] Deepti Ghadiyaram, Janice Pan, Alan C. Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. “In-Capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.9 (2018), pp. 2061–2077.
- [GKR19] Steve Göring, Christopher Krämmer, and Alexander Raake. “cencro – Speedup of Video Quality Calculation using Center Cropping”. In: *2019 IEEE ISM*. Dec. 2019, pp. 1–8.
- [Gno21] Sully Gnome. *Twitch statistics and analytics*. 2021. URL: <https://sullygnome.com/>.
- [Gör+19] Steve Göring, Julian Zebelein, Simon Wedel, Dominik Keller, and Alexander Raake. “Analyze And Predict the Perceptibility of UHD Video Contents”. In: *EI, HVEI* (2019).
- [Gör+20] Steve Göring, Robert Steger, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Automated Genre Classification for Gaming Videos”. In: *22nd IEEE International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020.
- [Gör+21a] Steve Göring, **Rakesh Rao Ramachandra Rao**, Bernhard Feiten, and Alexander Raake. “Modular Framework and Instances of Pixel-Based Video Quality Models for UHD-1/4K”. In: *IEEE Access* 9 (2021).
- [Gör+21b] Steve Göring, **Rakesh Rao Ramachandra Rao**, Stephan Fremerey, and Alexander Raake. “AVRate Voyager: An open source online testing platform”. In: *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2021.

- [Gör+23] Steve Göring, **Rakesh Rao Ramachandra Rao**, Rasmus Merten, and Alexander Raake. “Appeal and quality assessment for AI-generated images”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023.
- [Göt+21a] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. “KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild”. In: *IEEE Access* 9. IEEE. 2021, pp. 72139–72160.
- [Göt+21b] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. *The Konstanz 150k in-the-Wild Video Database (KonVid-150k)*. 2021. URL: <http://database.mmsp-kn.de>.
- [GPL00] Anthony Greene, Colin Prepscius, and William Levy. “Primacy Versus Recency in a Quantitative Model: Activity Is the Critical Distinction”. In: *Learning & Memory* 7 (Jan. 2000), pp. 48–57.
- [GR11] Marie-Neige Garcia and Alexander Raake. “Frame-layer packet-based parametric video quality model for encrypted video in IPTV services”. In: *2011 Third International Workshop on Quality of Multimedia Experience*. IEEE. 2011, pp. 102–106.
- [GR19] Steve Göring and Alexander Raake. “Evaluation of Intra-Coding Based Image Compression”. In: *2019 8th European Workshop on Visual Information Processing (EUVIP)*. 2019, pp. 169–174.
- [GRR19] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “nofu - A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content”. In: *Eleventh IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany, June 2019.
- [GRR20] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Prenc – Predict Number Of Video Encoding Passes With Machine Learning”. In: *Twelfth IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020.
- [GRR23] Steve Göring, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Quality Assessment of Higher Resolution Images and Videos with Remote Testing”. In: *Quality and User Experience (QUEx)* 8 (2023).

Bibliography

- [GSR10] MN Garcia, R Schleicher, and A Raake. "Towards a content-based parametric video quality model for IPTV". In: *Proceedings of the 3rd International Workshop on Perceptual Quality of Systems (PQS'10)*. 2010.
- [GSR18] Steve Göring, Janto Skowronek, and Alexander Raake. "DeViQ – A deep no reference video quality model". In: *Electronic Imaging, Human Vision Electronic Imaging* (2018).
- [Har] Harmonic. *Free 4K Demo Footage - Ultra HD Demo Footage*. URL: <https://www.harmonicinc.com/4k-demo-footage-download/> (visited on 10/20/2018).
- [He+18] Tiantian He, Rong Xie, Jia Su, Xin Tang, and Li Song. "A No Reference Bitstream-Based Video Quality Assessment Model for H.265/HEVC and H.264/AVC". In: *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 2018, pp. 1–5.
- [HKE13] P. Hanhart, P. Korshunov, and T. Ebrahimi. "Benchmarking of quality metrics on ultra-high definition video sequences". In: *2013 18th Int. Conference on Digital Signal Processing (DSP)*. July 2013, pp. 1–8.
- [Hos+17] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. "The Konstanz natural video database (KoNViD-1k)". In: *QoMEX*. IEEE. 2017.
- [Hos+20] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. "KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment". In: *IEEE Transactions on Image Processing* 29 (2020).
- [Hoß+11] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. "Quantification of YouTube QoE via Crowdsourcing". In: *2011 IEEE International Symposium on Multimedia*. 2011, pp. 494–499.
- [Hoß+14a] T. Hoßfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer. "Survey of web-based crowdsourcing frameworks for subjective quality assessment". In: *2014 IEEE 16th International Workshop on MMSP*. 2014, pp. 1–6.

- [Hoß+14b] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. “Best Practices for QoE Crowdttesting: QoE Assessment With Crowdsourcing”. In: *IEEE Transactions on Multimedia* 16.2 (2014), pp. 541–558.
- [HSE11] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. “SOS: The MOS is not enough!” In: *2011 third international workshop on quality of multimedia experience*. IEEE. 2011, pp. 131–136.
- [HSF15] Xin Huang, Jacob Sogaard, and Soren Forchhammer. “No-reference video quality assessment by HEVC codec analysis”. In: *2015 Visual Communications and Image Processing (VCIP)*. 2015, pp. 1–4.
- [Inc14] Apple Inc. *HLS Authoring Specification for Apple Devices*. 2014. URL: https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices (visited on 02/28/2014).
- [Iqb21] Mansoor Iqbal. *Netflix Revenue and Usage Statistics (2021)*. 2021. URL: <https://www.businessofapps.com/data/netflix-statistics/> (visited on 09/20/2021).
- [ISO19] Information technology ISO/IEC 23009-1:2019. *Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats*. ISO/IEC 23009-1:2019, Information technology, 2019.
- [ITU04] ITU-T Rec. P.563. *Single-ended method for objective speech quality assessment in narrow-band telephony applications*. Geneva, Switzerland: International Telecommunication Union, 2004.
- [ITU07] ITU-T Rec.G.1070. *“Opinion model for video-telephony applications*. Geneva, Switzerland: International Telecommunication Union, 2007.
- [ITU08] ITU-T Rec. E.800. *E.800 : Definitions of terms related to quality of service*. Geneva, Switzerland: International Telecommunication Union, 2008.
- [ITU09] ITU-T Rec. G.107. *The E-Model, a Computational Model for Use in Transmission Planning*. International Telecommunication Union. CH–Geneva, 2009.

Bibliography

- [ITU12a] ITU-T. *BT.2020 : Parameter values for ultra-high definition television systems for production and international programme exchange*. Tech. rep. Int. Telecomm. Union, 2012.
- [ITU12b] ITU-T Rec. P.1201.2. *Parametric non-intrusive assessment of audiovisual media streaming quality - Lower resolution application area*. Geneva, Switzerland: International Telecommunication Union, 2012.
- [ITU14a] ITU-T. *P.1401 : Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*. Tech. rep. Int. Telecommunication Union, 2014.
- [ITU14b] ITU-T. *RECOMMENDATION ITU-R BT.500-13 – Methodology for the subjective assessment of the quality of television pictures*. Tech. rep. Int. Telecommunication Union, 2014.
- [ITU16a] ITU-T. *ITU-T Rec. G.1022 (07/16)*. Tech. rep. Int. Telecommunication Union, 2016.
- [ITU16b] ITU-T. *Recommendation P.1203 - Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*. Tech. rep. International Telecommunication Union, 2016.
- [ITU16c] ITU-T Rec.G.1071. *Opinion model for network planning of video and audio streaming applications*. Geneva, Switzerland: International Telecommunication Union, 2016.
- [ITU17] ITU-T Rec. P.1203.2. *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Audio quality estimation module*. Geneva, Switzerland: International Telecommunication Union, 2017.
- [ITU19a] ITU-T. *Recommendation P.1204 - Video quality assessment of streaming services over reliable transport for resolutions up to 4K*. Tech. rep. International Telecommunication Union, 2019.
- [ITU19b] ITU-T. *Recommendation P.1204.3 : Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information*. Tech. rep. International Telecommunication Union, 2019.

- [ITU19c] ITU-T. *Recommendation P.1204.4 : Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information*. Tech. rep. International Telecommunication Union, 2019.
- [ITU19d] ITU-T. *Recommendation P.1204.5 : Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information*. Tech. rep. International Telecommunication Union, 2019.
- [ITU19e] ITU-T Rec. P.1203.1. *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Video quality estimation module*. Geneva, Switzerland: International Telecommunication Union, 2019.
- [ITU20] ITU-T Rec. P.1203.3. *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport - Quality integration module*. Geneva, Switzerland: International Telecommunication Union, 2020.
- [ITU21a] ITU-T Rec. H.264. *H.264 : Advanced video coding for generic audiovisual services*. Geneva, Switzerland: International Telecommunication Union, 2021.
- [ITU21b] ITU-T Rec. H.265. *H.265 : High efficiency video coding*. Geneva, Switzerland: International Telecommunication Union, 2021.
- [ITU22] ITU-T Rec. H.266. *H.266 : Versatile video coding*. Geneva, Switzerland: International Telecommunication Union, 2022.
- [ITU99] ITU-T Rec. P.910. *Subjective video quality assessment methods for multimedia applications*. Geneva, Switzerland: International Telecommunication Union, 1999.
- [Izu+14] Kosuke Izumi, Kei Kawamura, Tomonobu Yoshino, and Sei Naito. “No reference video quality assessment based on parametric analysis of HEVC bitstream”. In: *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. 2014, pp. 49–50.

Bibliography

- [Joh97] Nils Olof Johannesson. “The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks”. In: *IEEE Communications Magazine* Jan. (1997), pp. 70–79.
- [Kat18] Ioannis Katsavounidis. *Dynamic optimizer — a perceptual video encoding optimization framework*. 2018. URL: <https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f> (visited on 03/05/2018).
- [Kei+12] C. Keimel, J. Habigt, C. Horch, and K. Diepold. “QualityCrowd — A framework for crowd-based quality evaluation”. In: *2012 Picture Coding Symposium*. 2012, pp. 245–248.
- [Kel+21] Dominik Keller, Markus Vaalgamaa, Erkki Paajanen, **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Groovability: Using Groove as a Novel Measure for Audio QoE with the Example of Smartphones”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 2021.
- [Kel+23] Dominik Keller, Felix von Hagen, Julius Prenzel, Kay Strama, **Rakesh Rao Ramachandra Rao**, and Alexander Raake. “Influence of Viewing Distances on 8K HDR Video Quality Perception”. In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023.
- [Kim+18] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. “Deep Video Quality Assessor: From Spatio-Temporal Visual Sensitivity to a Convolutional Neural Aggregation Network”. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Cham: Springer International Publishing, 2018, pp. 224–241. ISBN: 978-3-030-01246-5.
- [Koi+21] Masanori Koike, Yuichiro Urata, Noritsugu Egi, and Kazuhisa Yamagishi. “Extension of ITU-T P.1204.3 Model to Tile-Based VR Streaming Services”. In: *2021 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR 2021)*. 2021, pp. 1–6.

- [Kor19] Jari Korhonen. "Two-Level Approach for No-Reference Consumer Video Quality Assessment". In: *IEEE Transactions on Image Processing* 28.12 (2019), pp. 5923–5938.
- [Lai+16] Jani Lainema, Miska M. Hannuksela, Vinod K. Malamal Vadakital, and Emre B. Aksu. "HEVC still image coding and high efficiency image file format". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 71–75.
- [Leb+15] Pierre Lebreton, Evangelos Skodras, Toni Mäki, Isabelle Hupont, and Matthias Hirth. "Bridging the Gap Between Eye Tracking and Crowdsourcing". In: vol. 9394. Feb. 2015.
- [Lee+17] C. Lee, S. Woo, S. Baek, J. Han, J. Chae, and J. Rim. "Comparison of objective quality models for adaptive bit-streaming services". In: *8th Int. Conf. on Information, Intelligence, Systems Applications (IISA)*. Aug. 2017, pp. 1–4.
- [Lee+20] Dae Yeol Lee, Hyunsuk Ko, Jongho Kim, and Alan C. Bovik. "Video Quality Model for Space-Time Resolution Adaptation". In: *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*. 2020, pp. 34–39.
- [Lee+21] Dae Yeol Lee, Somdyuti Paul, Christos G. Bampis, Hyunsuk Ko, Jongho Kim, Se Yoon Jeong, Blake Homan, and Alan C. Bovik. *A Subjective and Objective Study of Space-Time Subsampled Video Quality*. 2021. arXiv: 2102.00088 [eess.IV].
- [Li+11] S. Li, F. Zhang, L. Ma, and K. N. Ngan. "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments". In: *IEEE Transactions on Multimedia* 13.5 (2011), pp. 935–949.
- [Li+19a] Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, and Xiaoming Tao. "Viewport Proposal CNN for 360° Video Quality Assessment". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10169–10178.
- [Li+19b] Zhuoran Li, Zhengfang Duanmu, Wentao Liu, and Zhou Wang. "AVC, HEVC, VP9, AVS2 or AV1? - A Comparative Study of State-of-the-Art Video Encoders on 4K Videos". In: *ICIAR*. 2019.

Bibliography

- [Lin+12] X. Lin, H. Ma, L. Luo, and Y. Chen. “No-reference video quality assessment in the compressed domain”. In: *IEEE Transactions on Consumer Electronics* 58.2 (2012), pp. 505–512.
- [LMP+12] Patrick Le Callet, Sebastian Möller, Andrew Perkis, et al. “Qualinet white paper on definitions of quality of experience”. In: *European network on quality of experience in multimedia systems and services (COST Action IC 1003) 3.2012* (2012).
- [LY19] Pierre Lebreton and Kazuhisa Yamagishi. “Transferring Adaptive Bit Rate Streaming Quality Models from H.264/HD to H.265/4K-UHD”. In: *IEICE Transactions on Communications* E102.B.12 (2019), pp. 2226–2242.
- [Mad+20a] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. “Capturing Video Frame Rate Variations via Entropic Differencing”. In: *IEEE Signal Processing Letters* 27 (2020), pp. 1809–1813. ISSN: 1558-2361.
- [Mad+20b] Pavan C. Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. *Subjective and Objective Quality Assessment of High Frame Rate Videos*. 2020. arXiv: 2007.11634 [cs.MM].
- [Mad+21] Pavan C. Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. “Subjective and Objective Quality Assessment of High Frame Rate Videos”. In: *IEEE Access* 9 (2021), pp. 108069–108082. ISSN: 2169-3536.
- [MMB12] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. “No-reference image quality assessment in the spatial domain”. In: *IEEE Transactions on Image Processing* 21.12 (2012), pp. 4695–4708.
- [Moc+15] Decebal Mocanu, Jeevan Pokhrel, Juan Pablo Garella, Janne Seppänen, Eirini Liotou, and Manish Narwaria. “No-reference video quality measurement: Added value of machine learning”. In: *Journal of Electronic Imaging* 24 (Dec. 2015), p. 061208.
- [Möl00] Sebastian Möller. *Assessment and Prediction of Speech Quality in Telecommunications*. Springer Science & Business Media, 2000.

- [MPE20] MPEG. *Low Complexity Enhancement Video Coding*. 2020. URL: <https://www.lcevc.org/>.
- [MSB13] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. “Making a “completely blind” image quality analyzer”. In: *IEEE Signal Processing Letters* 20.3 (2013), pp. 209–212.
- [MVV20] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. “UVG Dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development”. In: *Proceedings of the 11th ACM Multimedia Systems Conference*. MM-Sys ’20. Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 297–302. ISBN: 9781450368452.
- [MZB15] Alex Mackin, Fan Zhang, and David R. Bull. “A study of subjective video quality at various frame rates”. In: *2015 IEEE International Conference on Image Processing (ICIP)*. 2015, pp. 3407–3411.
- [MZB19] Alex Mackin, Fan Zhang, and David R. Bull. “A Study of High Frame Rate Video Formats”. In: *IEEE Transactions on Multimedia* 21.6 (2019), pp. 1499–1512.
- [Nad+20] Babak Naderi, Rafael Jiménez, Matthias Hirth, Sebastian Möller, Florian Metzger, and Tobias Hossfeld. “Towards speech quality assessment using a crowdsourcing approach: evaluation of standardized methods”. In: *Quality and User Experience* 6 (Dec. 2020).
- [Net+96] J. Neter, M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. WCB McGraw-Hill, 1996.
- [Net18] Netflix. *VMAF 4K included*. [Online; 07.09.2018]. 2018. URL: <https://github.com/Netflix/vmaf>.
- [NM15] Tung Nguyen and Detlev Marpe. “Objective Performance Evaluation of the HEVC Main Still Picture Profile”. In: *IEEE Trans. Cir. and Sys. for Video Technol.* 25.5 (May 2015), pp. 790–797. ISSN: 1051-8215.
- [NR10] Stefanie Nowak and Stefan Rürger. “How Reliable Are Annotations via Crowdsourcing: A Study about Inter-Annotator Agreement for Multi-Label Image Annotation”. In: *Proceedings of the International Conference on Multimedia Information Retrieval*. MIR ’10. Philadelphia, Pennsylvania, USA: ACM, 2010. ISBN: 9781605588155.

Bibliography

- [Nuu+16] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. “CVD2014—A Database for Evaluating No-Reference Video Quality Assessment Algorithms”. In: *IEEE Transactions on Image Processing* 25.7 (2016), pp. 3073–3086.
- [Ord+20] Marta Orduna, César Díaz, Lara Muñoz, Pablo Pérez, Ignacio Benito, and Narciso García. “Video Multimethod Assessment Fusion (VMAF) on 360VR Contents”. In: *IEEE Transactions on Consumer Electronics* 66.1 (2020), pp. 22–31.
- [Osa+09] Osamu, Sei Naito, Shigeyuki Sakazawa, and Atsushi Koike. “Objective perceptual video quality measurement method based on hybrid no reference framework”. In: *2009 16th IEEE International Conference on Image Processing (ICIP)*. 2009, pp. 2237–2240.
- [Pan11] R. Pantos. *HTTP Live Streaming*. 2011. URL: <https://tools.ietf.org/html/draft-pantos-http-live-streaming-13> (visited on 07/07/2017).
- [Pin+11] Maurizio Pintus, Giaime Ginesu, Luigi Atzori, and Daniele D. Giusto. “Objective Evaluation of WebP Image Compression Efficiency”. In: *MobiMedia*. 2011.
- [PSC14] M Pinson, Marc Sullivan, and Andrew Catellier. “A new method for immersive audiovisual subjective testing”. In: *Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*. 2014.
- [PW] Margaret H Pinson and Stephen Wolf. “Comparing subjective video quality testing methodologies”. In: vol. 5150. International Society for Optics and Photonics.
- [R C19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: <https://www.R-project.org/>.
- [Raa+08] A. Raake, M. -. Garcia, S. Moller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann. “T-V-model: Parameter-based prediction of IPTV quality”. In: *2008 IEEE International Conference on Acoustics, Speech and*

Signal Processing. Mar. 2008, pp. 1149–1152. DOI: 10.1109/ICASSP.2008.4517818.

- [Raa+11] Alexander Raake, Jörgen Gustafsson, Savvas Argyropoulos, Marie-Neige Garcia, David Lindgren, Gunnar Heikkilä, Martin Pettersson, Peter List, and Bernhard Feiten. “IP-Based Mobile and Fixed Network Audiovisual Media Services”. In: *IEEE Signal Processing Magazine* 28.6 (2011), pp. 68–79.
- [Raa+17] Alexander Raake, Marie-Neige Garcia, Werner Robitza, Peter List, Steve Göring, and Bernhard Feiten. “A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1”. In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 2017, pp. 1–6.
- [Raa+20a] Alexander Raake, Silvio Borer, Shahid Satti, Jörgen Gustafsson, **Rakesh Rao Ramachandra Rao**, Stefano Medagli, Peter List, Steve Göring, David Lindero, Werner Robitza, Gunnar Heikkilä, Simon Broom, Christian Schmidmer, Bernhard Feiten, Ulf Wüstenhagen, Thomas Wittmann, Matthias Obermann, and Roland Bitto. “Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204”. In: *IEEE Access* 8 (2020).
- [Raa+20b] Alexander Raake, Ashutosh Singla, **Rakesh Rao Ramachandra Rao**, Werner Robitza, and Frank Hofmeyer. “SiSiMo: Towards Simulator Sickness Modeling for 360° Videos Viewed with an HMD”. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 2020.
- [Rao+19a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Werner Robitza, Bernhard Feiten, and Alexander Raake. “AVT-VQDB-UHD-1: A Large Scale Video Quality Database for UHD-1”. In: *21st IEEE International Symposium on Multimedia (IEEE ISM)*. Dec. 2019.
- [Rao+19b] **Rakesh Rao Ramachandra Rao**, Steve Göring, Patrick Vogel, Nicolas Pachatz, Juan Jose Villamar Villarreal, Werner Robitza, Peter List, Bernhard Feiten, and Alexander Raake. “Adaptive video streaming with

Bibliography

- current codecs and formats: Extensions to parametric video quality model ITU-T P.1203". In: *Electronic Imaging* (2019).
- [Rao+20a] **Rakesh Rao Ramachandra Rao**, Steve Göring, Peter List, Werner Robitza, Bernhard Feiten, Ulf Wüstenhagen, and Alexander Raake. "Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analysis and Open Source Implementation". In: *Twelfth IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. Athlone, Ireland, May 2020.
- [Rao+20b] **Rakesh Rao Ramachandra Rao**, Steve Göring, Robert Steger, Saman Zadtootaghaj, Nabajeet Barman, Stephan Fremerey, Sebastian Möller, and Alexander Raake. "A Large-scale Evaluation of the bitstream-based video-quality model ITU-T P.1204.3 on Gaming Content". In: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020.
- [Rao+23] **Rakesh Rao Ramachandra Rao**, Silvio Borer, David Lindero, Steve Göring, and Alexander Raake. "PNATS-UHD-1-Long: An Open Video Quality Dataset for Long Sequences for HTTP-based Adaptive Streaming QoE Assessment". In: *15th International Conference on Quality of Multimedia Experience (QoMEX)*. 2023.
- [Ras+10] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. "Collecting Image Annotations Using Amazon's Mechanical Turk". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Los Angeles: Association for Computational Linguistics, June 2010.
- [Ras17] R. Rassool. "VMAF reproducibility: Validating a perceptual practical video quality metric". In: *IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting*. June 2017, pp. 1–2.
- [RE14] Alexander Raake and Sebastian Egger. "Quality and quality of experience". In: *Quality of experience*. Springer, 2014, pp. 11–33.
- [Rez+20] Yuriy Reznik, Xiangbo Li, Karl Lillevold, Robert Peck, Thom Shutt, and Peter Howard. "Optimizing Mass-Scale Multi-Screen Video Delivery". In: *SMPTE Motion Imaging Journal* 129.3 (2020), pp. 26–38.

- [RGR17] Werner Robitza, Marie-Neige Garcia, and Alexander Raake. "A modular HTTP adaptive streaming QoE model — Candidate for ITU-T P.1203 ("P.NATS")". In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. 2017, pp. 1–6.
- [RGR21a] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. "Enhancement of Pixel-based Video Quality Models using Meta-data". In: *Electronic Imaging, Human Vision Electronic Imaging*. 2021.
- [RGR21b] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. "Towards High Resolution Video Quality Assessment in the Crowd". In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 2021.
- [RGR22] **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. "AVQBits - Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications". In: *IEEE Access* 10 (2022).
- [Rib+11] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer. "CROWDMOS: An approach for crowdsourcing mean opinion score studies". In: *2011 ICASSP*. 2011.
- [RL08] Andreas Rossholm and Benny Lovstroem. "A new low complex reference free video quality predictor". In: *2008 IEEE 10th Workshop on Multimedia Signal Processing*. 2008, pp. 765–768.
- [Rob+18a] Werner Robitza, Steve Göring, Alexander Raake, David Lindegren, Gunnar Heikkilä, Jörgen Gustafsson, Peter List, Bernhard Feiten, Ulf Wüstenhagen, Marie-Neige Garcia, Kazuhisa Yamagishi, and Simon Broom. "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software". In: *9th ACM Multimedia Systems Conference*. Amsterdam, 2018. ISBN: 9781450351928.
- [Rob+18b] Werner Robitza, Dhananjaya G Kittur, Alexander M Dethof, Steve Göring, Bernhard Feiten, and Alexander Raake. "Measuring YouTube QoE with ITU-T P. 1203 under Constrained Bandwidth Conditions". In: *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2018, pp. 1–6.

Bibliography

- [Rob+21] Werner Robitza, **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. “Impact of Spatial and Temporal Information on Video Quality and Compressibility”. In: *13th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. June 2021.
- [Rob+22] Werner Robitza, **Rakesh Rao Ramachandra-Rao**, Steve Göring, Alexander Dethof, and Alexander Raake. “Deploying the ITU-T P.1203 QoE Model in the Wild and Retraining for New Codecs”. In: *Proceedings of the 1st Conference on Mile-High Video*. MHV '22. Denver, Colorado: Association for Computing Machinery, 2022.
- [RT14] Benjamin Rainer and Christian Timmerer. “Quality of Experience of Web-Based Adaptive HTTP Streaming Clients in Real-World Environments Using Crowdsourcing”. In: *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*. VideoNext '14. Sydney, Australia: ACM, 2014. ISBN: 9781450332811.
- [SB06] H. R. Sheikh and A. C. Bovik. “Image information and visual quality”. In: *IEEE Transactions on Image Processing* 15.2 (2006), pp. 430–444.
- [SB13] R. Soundararajan and A. C. Bovik. “Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.4 (2013), pp. 684–694.
- [SB19] Z. Sinno and A. C. Bovik. “Large-Scale Study of Perceptual Video Quality”. In: *IEEE Transactions on Image Processing* (2019).
- [SDF12] Hao Su, J. Deng, and L. Fei-Fei. “Crowdsourcing annotations for visual object detection”. In: (Jan. 2012).
- [Ses+10] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. “Study of Subjective and Objective Quality Assessment of Video”. In: *IEEE Transactions on Image Processing* 19.6 (2010), pp. 1427–1441.
- [Seu+15] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hofffeld, and P. Tran-Gia. “A Survey on Quality of Experience of HTTP Adaptive Streaming”. In: *IEEE Comm. Surveys Tutorials* 17.1 (Mar. 2015), pp. 469–492. ISSN: 1553-877X.

- [SH16] Michael Seufert and Tobias Hossfeld. "One Shot Crowdttesting: Approaching the Extremes of Crowdsourced Subjective Quality Testing". In: Aug. 2016, pp. 122–126.
- [Sha+14a] M. Shahid, J. Sogaard, J. Pokhrel, K. Brunnström, K. Wang, S. Tavakoli, and N. Gracia. "Crowdsourcing based subjective quality assessment of adaptive video streaming". In: *2014 QoMEX*. 2014, pp. 53–54.
- [Sha+14b] Muhammad Shahid, Andreas Rossholm, Benny Lövsström, and Hans-Jürgen Zepernick. "No-reference image and video quality assessment: a classification and review of recent approaches". In: *EURASIP Journal on Image and Video Processing 2014* (2014), pp. 1–32.
- [Sha+22] Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. "Study of the Subjective and Objective Quality of High Motion Live Streaming Videos". In: *IEEE Transactions on Image Processing* 31 (2022), pp. 1027–1041.
- [SHR12] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino. "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC". In: *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. 2012, pp. 127–131.
- [Sid+19] Naty Sidaty, Wassim Hamidouche, Olivier Déforges, Pierrick Philippe, and Jérôme Fournier. "Compression Performance of the Versatile Video Coding: HD and UHD Visual Quality Monitoring". In: *2019 Picture Coding Symposium (PCS)*. 2019, pp. 1–5.
- [Sin+19] Ashutosh Singla, **Rakesh Rao Ramachandra Rao**, Steve Göring, and Alexander Raake. "Assessing Media QoE, Simulator Sickness and Presence for Omnidirectional Videos with Different Test Protocols". In: *26th IEEE Conference on Virtual Reality and 3D User Interfaces*. Osaka, Japan, Mar. 2019.
- [Sin+21] Ashutosh Singla, Steve Göring, Dominik Keller, **Rakesh Rao Ramachandra Rao**, Stephan Fremerey, and Alexander Raake. "Assessment of the Simulator Sickness Questionnaire for Omnidirectional Videos". In: *28th IEEE Conference on Virtual Reality and 3D User Interfaces*. 2021.

Bibliography

- [SLY17] Yule Sun, Ang Lu, and Lu Yu. “Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video”. In: *IEEE Signal Processing Letters* 24.9 (2017), pp. 1408–1412.
- [Sod11] Iraj Sodagar. “The MPEG-DASH Standard for Multimedia Streaming Over the Internet”. In: *IEEE MultiMedia* 18.4 (2011), pp. 62–67.
- [Son+13] Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. “The SJTU 4K video sequence dataset”. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. 2013, pp. 34–35.
- [SRL11] Muhammad Shahid, Andreas Rossholm, and Benny Lövsström. “A reduced complexity no-reference artificial neural network based video quality predictor”. In: *2011 4th International Congress on Image and Signal Processing*. Vol. 1. 2011, pp. 517–521.
- [SRL13] Muhammad Shahid, Andreas Rossholm, and Benny Lövsström. “A no-reference machine learning based video quality predictor”. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. 2013, pp. 176–181.
- [Sta+13] J. Stankowski, T. Grajek, K. Wegner, and M. Domanski. “Video quality in multiple HEVC encoding-decoding cycles”. In: *2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2013, pp. 75–78.
- [Sto22] Julia Stoll. *Number of Netflix paid subscribers worldwide from 1st quarter 2013 to 3rd quarter 2022*. 2022. URL: <https://www.statista.com/statistics/483112/netflix-subscribers/> (visited on 10/19/2022).
- [Sul+12] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. “Overview of the High Efficiency Video Coding (HEVC) Standard”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 22.12 (2012), pp. 1649–1668.
- [TB12] Anthony Tang and Sebastian Boring. “#EpicPlay: Crowd-Sourcing Sports Video Highlights”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. Austin, Texas, USA: Association for Computing Machinery, 2012. ISBN: 9781450310154.

- [Tho+16] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. "YFCC100M". In: *Communications of the ACM* 59.2 (Jan. 2016), pp. 64–73. ISSN: 1557-7317.
- [Tom+10] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi. "Performance comparisons of subjective quality assessment methods for mobile video". In: *2nd Int. Workshop on Quality of Multimedia Experience (QoMEX)*. June 2010, pp. 82–87.
- [Tor+16] Maria Torres Vega, Vittorio Sguazzo, Decebal Constantin Mocanu, and Antonio Liotta. "An experimental survey of no-reference video quality assessment methods". English. In: *International journal of pervasive computing and communications* 12.1 (2016), pp. 66–86. ISSN: 1742-7371.
- [Tra+16a] Huyen T. T. Tran, Nam Pham Ngoc, Anh T. Pham, and Truong Cong Thang. "A Multi-Factor QoE Model for Adaptive Streaming over Mobile Networks". In: *2016 IEEE Globecom Workshops (GC Wkshps)*. 2016, pp. 1–6.
- [Tra+16b] Huyen T. T. Tran, Thang Vu, Nam Pham Ngoc, and Truong Cong Thang. "A novel quality model for HTTP adaptive streaming". In: *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)*. 2016, pp. 423–428.
- [Tra+17] Huyen T. T. Tran, Nam Pham Ngoc, Cuong Manh Bui, Minh Hong Pham, and Truong Cong Thang. "An evaluation of quality metrics for 360 videos". In: *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*. 2017, pp. 7–11.
- [Tu+21] Zhengzhong Tu, Chia-Ju Chen, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. "Efficient User-Generated Video Quality Prediction". In: *2021 Picture Coding Symposium (PCS)*. 2021, pp. 1–5.
- [Utk+20] Markus Utke, Saman Zadtootaghaj, Steven Schmidt, Sebastian Bosse, and Sebastian Möller. "NDNetGaming-development of a no-reference deep CNN for gaming video quality prediction". In: *Multimedia Tools and Applications* (2020), pp. 1–23.

Bibliography

- [Van+16] G. Van Wallendael, P. Coppens, T. Paridaens, N. Van Kets, W. Van den Broeck, and P. Lambert. “Perceptual quality of 4K-resolution video content compared to HD”. In: *8th Int. Conference on Quality of Multimedia Experience (QoMEX)*. June 2016, pp. 1–6.
- [Veg+17] Maria Torres Vega, Decebal Constantin Mocanu, Jeroen Famaey, Stavros Stavrou, and Antonio Liotta. “Deep Learning for Quality Assessment in Live Video Streaming”. In: *IEEE Signal Processing Letters* 24.6 (2017), pp. 736–740.
- [Wan+04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.
- [Wan+21] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. “Rich features for perceptual quality assessment of UGC videos”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 13430–13439.
- [Wei+14] Benjamin Weiss, Dennis Guse, Sebastian Möller, Alexander Raake, Adam Borowiak, and Ulrich Reiter. “Temporal development of quality of experience”. In: *Quality of experience*. Springer, 2014, pp. 133–147.
- [WIA19] Yilin Wang, Sasi Inguva, and Balu Adsumilli. “YouTube UGC Dataset for Video Compression Research”. In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)* (Sept. 2019).
- [Wie+03] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. “Overview of the H.264/AVC video coding standard”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 13.7 (2003), pp. 560–576. DOI: 10.1109/TCSVT.2003.815165.
- [Woj20] Susan Wojcicki. *YouTube at 15: My personal journey and the road ahead*. 2020. URL: <https://blog.youtube/news-and-events/youtube-at-15-my-personal-journey/> (visited on 02/15/2020).

- [Woj21] Susan Wojcicki. *YouTube by the Numbers: Stats, Demographics & Fun Facts*. 2021. URL: <https://www.omnicoreagency.com/youtube-statistics/> (visited on 01/03/2021).
- [WSB03] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. Vol. 2. IEEE. 2003, pp. 1398–1402.
- [WWC19] Sarah Wassermann, Nikolas Wehner, and Pedro Casas. "Machine Learning Models for YouTube QoE and User Engagement Prediction in Smartphones". In: 46.3 (Jan. 2019), pp. 155–158. ISSN: 0163-5999.
- [XJ13] J. Xu and X. Jiang. "Research on Subjective Assessment Method of Ultra High Definition Video Quality". In: *4th World Congress on Software Engineering*. Dec. 2013, pp. 326–330.
- [Yam+21] Kazuhisa Yamagishi, Noritsugu Egi, Noriko Yoshimura, and Pierre Lebreton. "Derivation Procedure of Coefficients of Metadata-Based Model for Adaptive Bitrate Streaming Services". In: *IEICE Transactions on Communications* E104.B.7 (2021), pp. 725–737.
- [YFH19] Shun-Huai Yao, Ching-Ling Fan, and Cheng-Hsin Hsu. "Towards Quality-of-Experience Models for Watching 360° Videos in Head-Mounted Virtual Reality". In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 2019, pp. 1–3.
- [YG13] Kazuhisa Yamagishi and Shan Gao. "Light-weight audiovisual quality assessment of mobile video: ITU-T Rec. P. 1201.1". In: *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSp)*. IEEE. 2013, pp. 464–469.
- [Yim+20] Joong Yim, Yilin Wang, Neil Aylon Charles Birkbeck, and Balu Adsumilli. "Subjective Quality Assessment for YouTube UGC Dataset". In: *2020 IEEE International Conference on Image Processing*. 2020.
- [YKH09] Kazuhisa Yamagishi, Taichi Kawano, and Takanori Hayashi. "Hybrid video quality-estimation model for IPTV services". In: *GLOBECOM 2009-2009 IEEE Global Telecommunications Conference*. IEEE. 2009, pp. 1–5.

Bibliography

- [YLG15] Matt Yu, Haricharan Lakshman, and Bernd Girod. “A Framework to Evaluate Omnidirectional Video Coding Schemes”. In: *2015 IEEE International Symposium on Mixed and Augmented Reality*. 2015, pp. 31–36.
- [Yu+21] Xiangxu Yu, Neil Birkbeck, Yilin Wang, Christos G. Bampis, Balu Adsumilli, and Alan C. Bovik. “Predicting the Quality of Compressed Videos With Pre-Existing Distortions”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 7511–7526.
- [Zad+18] Saman Zadtootaghaj, Nabajeet Barman, Steven Schmidt, Maria G. Martini, and Sebastian Möller. “NR-GVQM: A No Reference Gaming Video Quality Metric”. In: *2018 IEEE International Symposium on Multimedia (ISM)*. 2018, pp. 131–134.
- [Zad+20a] Saman Zadtootaghaj, Nabajeet Barman, **Rakesh Rao Ramachandra Rao**, Steve Göring, Maria G. Martini, Alexander Raake, and Sebastian Möller. “DEMI: Deep Video Quality Estimation Model using Perceptual Video Quality Dimensions”. In: *22nd IEEE International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2020.
- [Zad+20b] Saman Zadtootaghaj, Steven Schmidt, Saeed Shafiee Sabet, Sebastian Moeller, and Carsten Griwodz. “Quality Estimation Models for Gaming Video Streaming Services Using Perceptual Video Quality Dimensions”. In: *Proceedings of the 11th International Conference on Multimedia Systems*. ACM. 2020.
- [Zha+11] L. Zhang, L. Zhang, X. Mou, and D. Zhang. “FSIM: A Feature Similarity Index for Image Quality Assessment”. In: *IEEE Transactions on Image Processing* 20.8 (2011), pp. 2378–2386.
- [Zha+20] Fan Zhang, Angeliki V Katsenou, Mariana Afonso, Goce Dimitrov, and David R Bull. “Comparing VVC, HEVC and AV1 using objective and subjective assessments”. In: *arXiv preprint arXiv:2003.10282* (2020).
- [ZMB17] F. Zhang, A. Mackin, and D. R. Bull. “A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 300–304.

List of Figures

1.1	Factors influencing Quality of Experience.	5
1.2	The MPD hierarchical model. This example shows how the client requests the appropriate representation and plays out the segment (adapted from [Sod11]).	8
1.3	An example HAS session	8
3.1	SI-TI of all the sources used in training and validation in the P.NATS Phase 2 competition.	44
3.2	Bitrate ranges for each encoder–resolution pair used in the P.NATS Phase 2 competition [Raa+20a].	46
3.3	MOS distribution of AVT-PNATS-UHD-1 dataset.	49
3.4	SOS analysis of the AVT-PNATS-UHD-1 dataset.	50
3.5	Thumbnails of source videos in the AVT-VQDB-UHD-1 dataset.	51
3.6	Spatial and temporal complexities SI, TI of all the video contents used in the AVT-VQDB-UHD-1 dataset.	52
3.7	MOS distribution of AVT-VQDB-UHD-1 dataset.	55
3.8	SOS analysis of the AVT-VQDB-UHD-1 dataset.	56
3.9	Inter-test correlation (test_2 and test_3).	57
3.10	Overview of the source videos used in the AV1 dataset.	57
3.11	SI and TI of all the source contents used in the AV1 dataset.	58
3.12	MOS distribution of AV1 dataset.	60
3.13	SOS analysis of AV1 dataset.	60
3.14	MOS comparison between AV1 and H.265.	61
3.15	MOS distribution of PNATS-UHD-1-Long dataset.	63
3.16	Responses to the pre-test questionnaire.	67
3.17	Distribution of browser window height across crowd participants.	68
3.18	Count distribution of how often PVSs were rated	69
3.19	Distribution of MOS in the Lab and Out-of-the-Lab.	70
3.20	SOS analysis of the lab and out-of-the-lab tests.	70
3.21	Scatter plot of the MOS values from lab and out-of-the-lab.	71
3.22	Correlation between lab and out-of-the-lab tests as a function of the number of participants in the out-of-the-lab test.	72

List of Figures

3.23	Overview of the source videos used for the long term audiovisual quality evaluation in the centre-crop crowd test.	73
3.24	SI-TI of all the sources used for the long-term audiovisual quality evaluation in the crowd.	74
3.25	Responses to the pre-test questionnaire.	76
3.26	Distribution of browser window height across crowd participants.	77
3.27	Count distribution of how often PVSs were rated.	77
3.28	Distribution of MOS for the lab and crowd tests.	78
3.29	SOS analysis of the lab and crowd tests.	78
3.30	Scatter plot of the MOS values from lab [Rao+19a] and crowd tests.	79
4.1	General model structure of <i>AVQBits</i> including all four model instances. . .	90
4.2	General model structure of the <i>AVQBits</i> <i>M3</i> / P.1204.3 model.	94
4.3	General model structure of the <i>AVQBits</i> <i>M0</i> model.	98
4.4	General model structure of the <i>AVQBits</i> <i>M1</i> model.	100
4.5	General model structure of the <i>AVQBits</i> <i>H0</i> model.	103
4.6	Scatter plot of <i>AVQBits</i> instances for AVT-VQDB-UHD-1 dataset.	112
4.7	Performance of the correction mapping for all considered video codecs. . . .	116
4.8	General Machine Learning Pipeline.	117
4.9	Frequency of occurrence of features in top 100 performing cases.	121
5.1	Scatter plot of <i>AVQBits</i> instances for PNATS-UHD-1-Long dataset.	129
6.1	MOS distribution of GVS, KUGVD, CGVDS, and Twitch datasets.	137
6.2	Scatter plot of <i>AVQBits</i> instances for the considered gaming datasets.	139
6.3	MOS distribution of 360 Streaming Video Quality Dataset.	146
6.4	Scatter plot of <i>AVQBits</i> instances for 360 Streaming Video Quality Dataset. .	148
6.5	MOS distribution of LIVE-YT-HFR dataset.	151
6.6	Scatter plot of <i>AVQBits</i> instances for LIVE-YT-HFR dataset.	154
6.7	MOS distribution of LIVE-APV dataset.	156
6.8	MOS distribution of <i>LIVE Wild Compressed Video Quality Database</i> dataset. . .	160
6.9	Scatter plot of <i>AVQBits</i> instances for <i>LIVE Wild Compressed Video Quality Database</i>	162
6.10	MOS distribution of the considered dataset.	165
6.11	Scatter plot of <i>AVQBits</i> instances for the considered dataset.	168

List of Tables

3.1	Number of unique footages and SRC files used in the training (TR) and validation (VL) databases in the P.NATS Phase 2 competition [Raa+20a].	43
3.2	Parameter ranges considered in the P.NATS Phase 2 competition [Raa+20a].	45
3.3	Common HRCs used in the P.NATS Phase 2 competition. The video codec is H.264 for all common conditions [Raa+20a].	47
3.4	Training database details [Raa+20a].	47
3.5	Validation database details [Raa+20a].	48
3.6	Source details for the AVT-VQDB-UHD-1 dataset.	52
3.7	Test Design – test_1.	53
3.8	Test Design – test_2 and test_3 (Bit-per-pixel based test).	54
3.9	Test Design – test_4 (Framerate variation test).	54
3.10	Source details for the AV1 dataset.	58
3.11	Test Design - AV1 dataset.	59
3.12	Range of parameters used in the long-duration tests in the P.NATS Phase 2 competition.	61
3.13	Per-source comparison of lab [Rao+19a] and out-of-the-lab test results. . . .	71
4.1	Aggregated features for RF model.	97
4.2	<i>AVQBits</i> / P.1204.3 Quantization-degradation coefficients, PC/TV case. . .	105
4.3	<i>AVQBits</i> / P.1204.3 Quantization-degradation coefficients, MO/TA case. . .	105
4.4	Upscaling- and temporal-degradation coefficients, PC/TV case.	106
4.5	Upscaling- and temporal-degradation coefficients, MO/TA case.	106
4.6	QP-Prediction coefficients for <i>AVQBits</i> M0, PC/TV case.	106
4.7	Quantization-degradation coefficients for <i>AVQBits</i> M0, PC/TV case.	107
4.8	QP-Prediction coefficients for <i>AVQBits</i> M1, PC/TV case.	107
4.9	Quantization-degradation coefficients for <i>AVQBits</i> M1, PC/TV case.	107
4.10	Codec mapping coefficients for <i>AVQBits</i> H0 f, PC/TV case.	108
4.11	Aggregated RMSE on validation and on all databases.	109
4.12	Overall model performance of different models on P.NATS Phase 2 validation databases only.	110

List of Tables

4.13 Performance of the *AVQBits* instances on the AVT-VQDB-UHD-1 dataset. . . 111

4.14 Performance comparison of the *AVQBits* instances with SoA models for tests in the AVT-VQDB-UHD-1 dataset without framerate as dependent variable. 113

4.15 Performance comparison of *AVQBits* instances with SoA models for tests with framerate as independent variable in the AVT-VQDB-UHD-1 dataset. . . 113

4.16 Correction Mapping - Coefficients per codec. 115

4.17 Performance comparison between Hybrid-VMAF and other SoA video quality models. 120

4.18 Performance comparison between Hybrid-VMAF and other SoA video quality models. 121

5.1 Performance of *AVQBits* instances on the PNATS-UHD-1-Long dataset. . . 128

6.1 Overview of the used gaming datasets. 136

6.2 Performance of *AVQBits* instances using the considered gaming datasets. . . 138

6.3 Comparison of performance of *AVQBits* instances with SoA models using the considered gaming datasets. 140

6.4 Performance of FHD-mapped P.1204.3 on the validation datasets. 142

6.5 HRCs for test_1 and test_2. 145

6.6 HRCs for test_3. 145

6.7 Performance of *AVQBits* instances using the 360 Streaming Video Dataset. . . 146

6.8 Comparison of performance of *AVQBits* instances with SoA models using the 360 Video Streaming Quality Dataset. 149

6.9 Comparison of performance of *AVQBits* instances with SoA models using the LIVE-YT-HFR dataset. 153

6.10 Comparison of performance of *AVQBits* instances with SoA models using the LIVE-APV dataset. 157

6.11 Comparison of performance of *AVQBits* instances with SoA models using the LIVE Wild Compressed Video Quality Database. 161

6.12 Comparison of performance of *AVQBits*|M3 and *AVQBits*|M0 with SoA models. 167

B.1 Test Plan for P2STR01. 188

B.2 Test Plan for P2STR02. 189

B.3 Test Plan for P2STR03. 190

B.4 Test Plan for P2STR04. 191

B.5 Test Plan for P2STR05. 192

List of Tables

B.6	Test Plan for P2STR06.	193
B.7	Test Plan for P2STR08.	194
B.8	Test Plan for P2STR09.	195
B.9	Test Plan for P2STR10.	196
B.10	Test Plan for P2STR11.	197
B.11	Test Plan for P2STR12.	198
B.12	Test Plan for P2STR13.	199
B.13	Test Plan for P2STR14.	200
B.14	Test Plan for P2SVL01.	201
B.15	Test Plan for P2SVL02.	202
B.16	Test Plan for P2SVL03.	203
B.17	Test Plan for P2SVL04.	204
B.18	Test Plan for P2SVL05.	205
B.19	Test Plan for P2SVL06.	206
B.20	Test Plan for P2SVL07.	207
B.21	Test Plan for P2SVL08.	208
B.22	Test Plan for P2SVL09.	209
B.23	Test Plan for P2SVL10.	210
B.24	Test Plan for P2SVL11.	211
B.25	Test Plan for P2SVL12.	212
B.26	Test Plan for P2SVL13.	213

List of Acronyms

A list of frequently used acronyms is given in the following table.

ACR	Absolute Category Rating
CBR	Constant Bitrate
CDN	Content Delivery Network
CMAF	Common Media Application Format
CRF	Constant Rate Factor
DASH	Dynamic Adaptive Streaming over HTTP
DNN	Deep Neural Network
FR	Full-Reference
HAS	HTTP-based adaptive streaming
HDR	High Dynamic Range
HFR	High Framerate
HLS	HTTP Live Streaming
HMD	Head-Mounted Display
HRC	Hypothetical Reference Circuit
MOS	Mean Opinion Score
MPD	Media Presentation Description
MPEG	Moving Pictures Expert Group
MSS	Microsoft Smooth Streaming
NR	No-Reference
PCC	Pearson Correlation Coefficient
PVS	Processed Video Sequence
QEB	Quality Equivalent Bitstream
QoE	Quality of Experience
QP	Quantization Parameter
RDO	Rate-Distortion Optimization
RF	Random Forest
RMSE	Root Mean Square Error
RR	Reduced-Reference
SI	Spatial Information
SOS	Standard deviation of Mean Opinion Scores
SRC	Source
SROCC	Spearman Rank Correlation Coefficient
SVR	Support Vector Regression
TI	Temporal Information
UGC	User-Generated Content
V-CNN	Viewport-based Convolutional Neural Networks
VQEG	Video Quality Experts Group
VR	Virtual Reality
