ITU-T Standardized Bitstream-based Video Quality Models

TECHNICAL REPORT

Werner Robitza, Rakesh Rao Ramachandra Rao, Steve Göring,

Alexander Raake

20.08.2020 — v1.0

Audiovisual
Technology
Group

TECHNISCHE UNIVERSITÄT
ILMENAU

Contents

Version History

Version	Date	Editor	Description
v1.0	20.08.2020	TU Ilmenau	First version

1 Highlights

As recently standardized video quality models from ITU-T, the highlights are:

- ITU-T Rec. P.1203 – a set of models to calculate the quality of HTTP Adaptive Streaming sessions
- ITU-T Rec. P.1204 – a set of models to calculate video quality for H.264, H.265 and VP9

Both have been extensively validated in the standardization process, by use of a large number of subjective test databases, and achieve high accuracy results compared to human ratings. Notably, P.1203 is the first standardized model to incorporate effects like stalling in the overall quality prediction, and the models from P.1204, in particular the bitstream-based model P.1203.4, perform better than VMAF.

2 Introduction

2.1 About standardized video quality models from ITU-T

ITU-T's role in video quality model standardization goes back to before 2000. This is when the Video Quality Experts Group (VQEG) conducted its first project in which different video quality models were evaluated against subjective test results. For instance, the FRTV Phase II project was completed in 2003, with four models being recommended to ITU-T for standardization. This resulted in ITU-T Rec. J.144, which includes the Video Quality Metric (VQM) proposed by NTIA – a metric that is still widely used today.

Later VQEG projects led to ITU-T Rec. J.247 and J.341, J.342 and J.343, just to name a few. VQEG's collaboration with ITU has primarily focused on full-reference and reduced-reference models, which require access to the video signals or features, and hybrid models, which need bitstream access. A detailed explanation of these model types is given later.

ITU-T's Study Group 12, Question 14 has been developing standards independently from VQEG, focusing on developing bitstream-based, parametric video quality models that do not require access to any video signals or the reference video, and are therefore very lightweight and usable within a network context, or directly at the client device. Such parametric/bitstream-based models in ITU-T Rec. P.1201 and P.1202 were aimed at IPTV applications with unreliable transmission (i.e., packet loss leading to visual artifacts); the model family ITU-T Rec. P.1203 addressed HTTP Adaptive Streaming. The most recent completed standard is ITU-T Rec. P.1204, a collaborative project between VQEG and ITU-T's Study Group 12, Question 14.

The models described in ITU-T standards have been created by an international consortium of academic and industrial partners. The importance of these standardized models stems from the meticulous attention given to the design of the subjective databases that are the basis for developing and training the models, and the rigorous testing and evaluation of the model performance. The rules under which different model candidates are to be submitted by different institutions, and later validated in terms of performance, are set *before* the databases are available for training, thus creating a level playing field for all parties wanting to participate. In fact, most standards were developed by means of a "competition", in which only the best model candidates would be selected as winners.

The ITU-T-standardized models have been trained and validated extensively. For example, for ITU-T Rec. P.1203, the models were trained on over 1,000 audiovisual sequences that were rated by human viewers, thus over

25,000 individual ratings. For the standard series ITU-T Rec. P.1204 on short-sequence video quality evaluation, almost 5,000 sequences have been rated by around 24 subjects per each of underlying 26 tests, with more than 600 subjects and more than 100,000 individual ratings. The subjective ratings were given in the context of standardized subjective tests conducted in dedicated laboratories, where the environment was set up according to international standards as well (like ITU-T Rec. P.910). Unreliable viewers were rejected, leading to high quality databases that were used for model training.

The standardized video quality models are essential for providing a reliable and valid assessment of video quality, particularly in the context of independent evaluations of network or video streaming providers (e.g., during benchmarking).

2.2 About TU Ilmenau

TU Ilmenau's Audiovisual Technology Group, headed by Prof. Alexander Raake, is actively involved in the standardization of video quality models within ITU-T's Study Group 12, Question 14. Prof. Raake is also co-rapporteur of Question 14. The group and its members, as well as previous associates of Prof. Raake, have been involved in the creation of ITU-T Rec. P.1201, P.1202, P.1203, P.1204 and J.343.1.

An overview of the AVT's group activities in the context of video quality standardization is available on a dedicated website, which features links to model software and tools, open-source video databases, and related articles.

2.3 Different model types

In order to understand the application areas for such models, one has to first consider how different models operate in principle—here shown in an overview in the below figure. Figure 1 shows how an original source video undergoes encoding and packetization and is then sent over a network. The diagram shows the possible information that can be extracted about this video from the perspective of the network or client device.

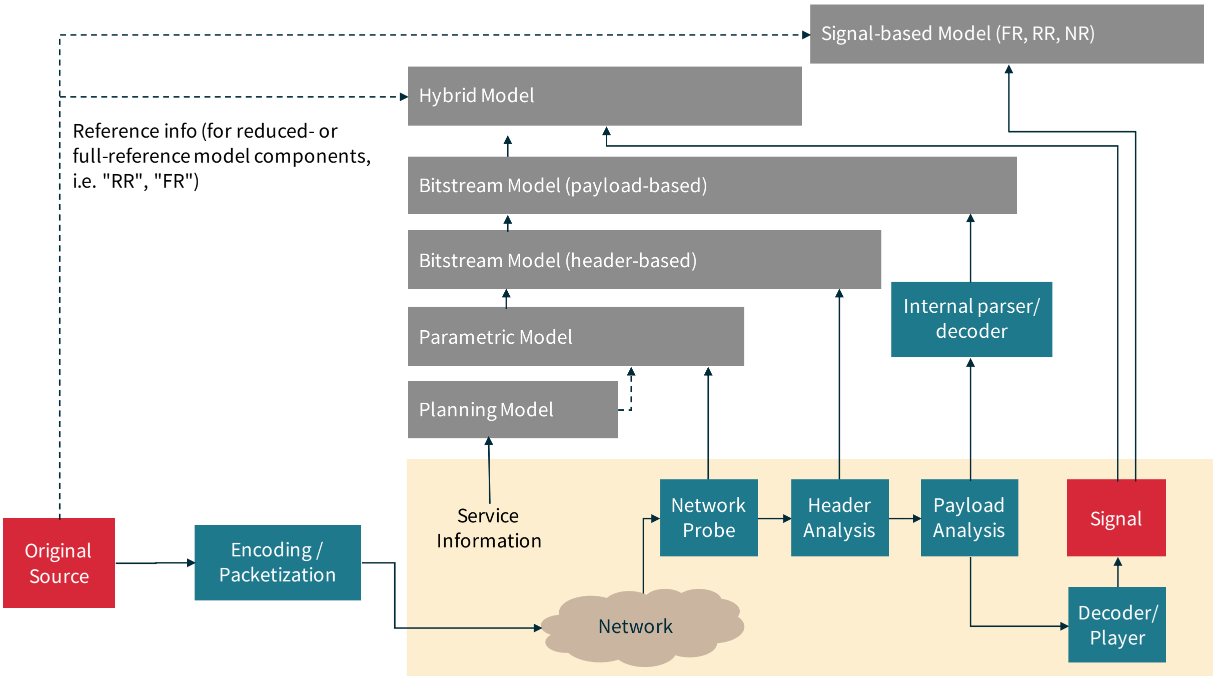

Figure 1: Different model types compared.

From bottom to top (in the grey rectangles), there is more information available for quality prediction. Typically, this allows for more precise quality prediction, at the expense of requiring more information or more computational effort to analyze it.

For example, analyzing pixels may give more accurate quality prediction results than just analyzing frame sizes. However, recent developments show that bitstream-based models can achieve performance comparable to signal-based models for the same application.

The model types shown in the figure are:

- **Planning models:** Operate only on information about the service, that is, hypothetical streams that have not occurred in practice. With planning models, an ISP could for example estimate the required bandwidth for streaming services to deliver acceptable quality.
- **Parametric and bitstream-based models:** Work with parameters describing the transmitted media (e.g. video codec, bitrate), or its actual bitstream—they are the main focus in this document.
- **Hybrid models:** Combine parametric models with signal-based models.
- **Signal-based models:** Require at least the decoded signal from the player (e.g. via screen capture) as pixels. Such a model is then called a no-reference (NR) model. A reduced-reference (RR) model has access to limited features from the original source, and a full-reference (FR) model has full access to the source pixels.

3 ITU-T Rec. P.1203 Description

3.1 Overview

The **ITU-T Recommendation P.1203** is a family of standards that specifies the world's first model to predict the Quality of Experience (QoE) for HTTP Adaptive Streaming (HAS) services. It was released in 2017 and consists of one main and three sub-recommendations:

- ITU-T P.1203 – Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport
- ITU-T P.1203.1: Video quality estimation module (short-term, providing per-one-second output information)
- ITU-T P.1203.2: Audio quality estimation module (short-term, providing per-one-second output information)
- ITU-T P.1203.3: Audiovisual integration and integration of final score, reflecting remembered quality for viewing sessions between 30 s and 5 min duration

Like most other quality models, it outputs quality in terms of Mean Opinion Scores (MOS) on a scale from 1–5, where 1 refers to Bad quality, and 5 to Excellent.

What is special about this model is that it does not just calculate video-only quality (like PSNR, SSIM or VMAF), but that it integrates audio and video, and that it factors in initial loading delay and stalling into the overall quality prediction. Also, it considers the difference between PC/TV display and mobile video viewing, where visual artifacts on mobile screens will be less visible, yielding higher MOS values.

3.2 Module Structure

P.1203 is composed of several modules that each compute different aspects of the overall quality estimation:

- The P.1203.1 and P.1203.2 standards predict video and audio quality in short segments of up to 10 seconds length. These models are metadata-based or bitstream-based. The quality prediction includes effects of degradations that may occur in a video stream caused by lossy compression, temporal or spatial downscaling, i.e. encoding at the server side. The media quality scores are provided to the P.1203.3 integration module on a per-1-second basis.
- The P.1203.3 standard predicts the QoE of an entire video session of up to 5 minutes length. It takes as input the existing short term scores from P.1203.1 and P.1203.2. The final score includes stalling effects due to rebuffering events (including initial loading).

This modular structure can be seen in Figure 2.

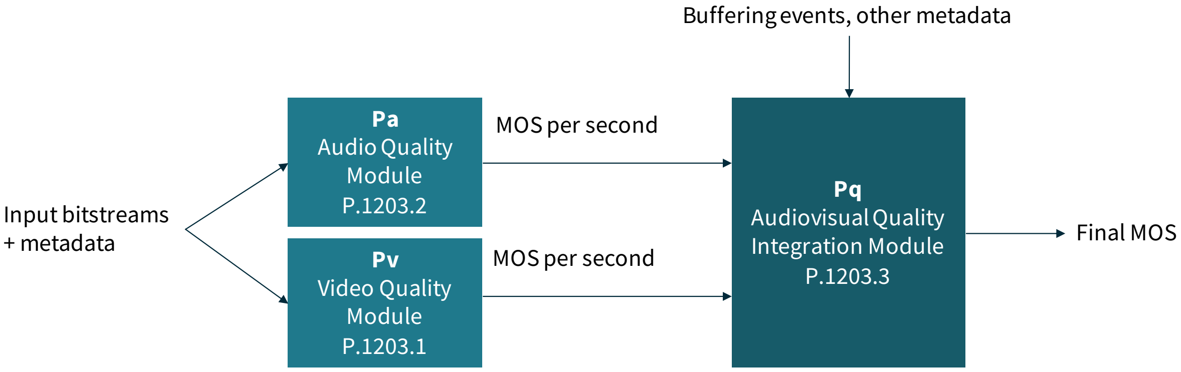

Figure 2: P.1203 model architecture.

In particular, P.1203 is novel because most video quality models do not incorporate fluctuations in quality over time, and cannot handle the impact of initial loading or stalling, although it is crucial for the overall MOS. The P.1203.3 component takes care of those effects.

P.1203 can be used for any type of video streaming on mobile devices or shown on laptops, PCs or TVs, for sequences up to 5 minutes length, with resolutions of up to 1080p HD and frame rates of up to 30 fps. Video must be coded with the H.264 codec; various audio codecs (including AAC) are supported. For other video codecs, an extension developed by TU Ilmenau is available.

3.3 Modes of Operation

P.1203.1, the video quality estimation module, offers four modes of operation, depending on the available information from the audiovisual stream and the required/available computational resources, shown in Figure 3.

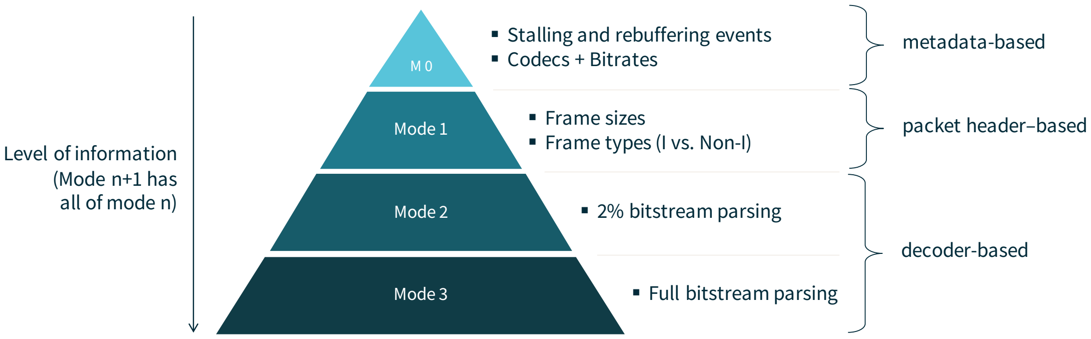

Figure 3: Different modes in P.1203.

P.1203's simplest mode of operation (mode 0) takes as input: audio/video codec, audio/video bitrate, video resolution, and frames per second. Depending on the available data, it offers higher modes of operation that increase prediction accuracy at the expense of being more computationally intensive and requiring input data from more in-depth bitstream inspection.

While Mode 0 has access to basic data, Mode 1 can inspect the packet headers of the transmitted stream to obtain frame sizes and types. Modes 2 and 3 have access to the bitstream itself, where mode 2 only accesses 2% of the stream to reduce computing efforts.

Mode 0 can be used with the information that is available in HAS manifests, as these include data related to video and audio encoding settings, for each respective representation. Mode 3 can be used with the information available in the transmitted bitstream; it can be extracted from the streams before they are transmitted (e.g. on the origin or the CDN), during transmission (e.g. via probes deployed in the network itself), or at the client device itself.

For all P.1203.1 modes, on top of the above information, it is also necessary to know the following:

- Quality switches: which representation the client was playing at any given point in time
- Stalling: whether there was any initial loading or stalling, including the durations of these events

These information are needed to perform the temporal integration of per-second MOS scores for video and audio, in order to estimate the final MOS with the P.1203.3 model.

3.4 Model Performance

Due to its novel architecture, the performance of the model cannot be easily compared to other models like PNSR or VMAF, since those were not developed for sequences longer than 10 s. Also, they do not take into account any stalling/quality fluctuations in their prediction.

According to the official ITU-T document, P.1203 offers a Pearson correlation of 0.81 to 0.89 depending on the mode used. This is the correlation between subjective MOS and the model score. The Root Mean Square Error (RMSE) lies between 0.47 (Mode 0) and 0.33 (Mode 3), where lower numbers are better.

Since the databases on which P.1203 was trained and validated are not publicly available, more detailed performance analyses can only be performed on open-sourced datasets. Figure 4 shows the model performance in terms of subjective MOS versus the model prediction (called O.46) for the four different modes of P.1203.1.

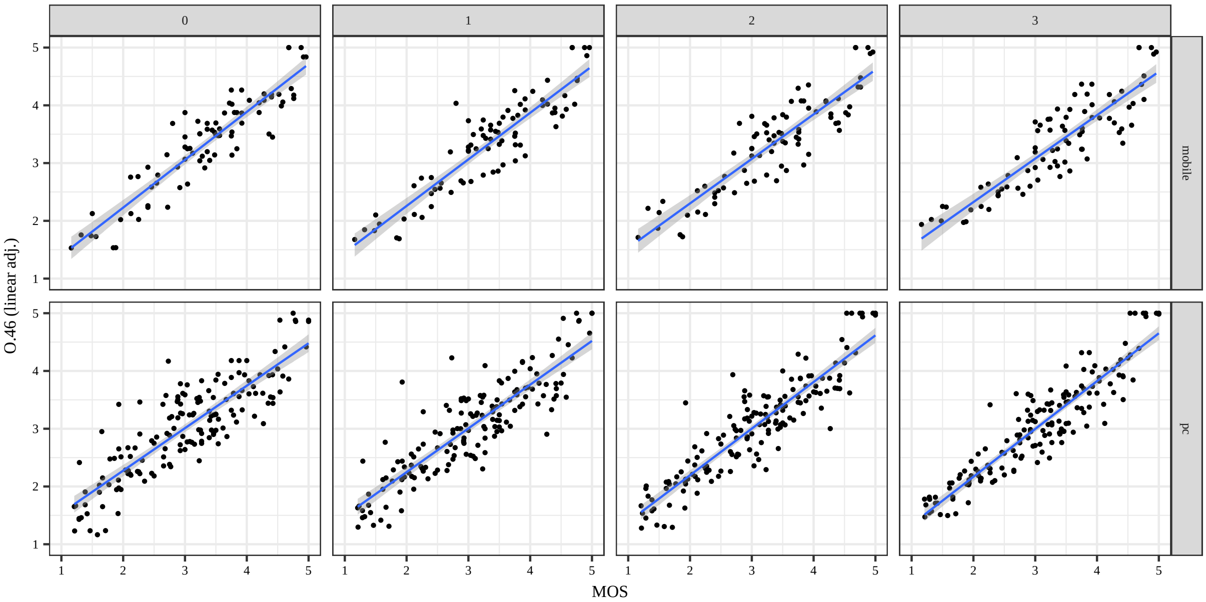

Figure 4: P.1203 model performance across different modes.

The databases used for this comparison are available publicly and have been published in a conference paper. They are part of the official training/validation databases from the P.1203 standardization process.

3.5 Reference Software and Publications

A reference software for ITU-T P.1203 is available at https://github.com/itu-p1203/itu-p1203/ and can be used freely for research purposes. Commercial licensing options are currently under development.

The P.1203 model has been used successfully in various academic publications, including an evaluation conducted by Susanna Sc which won the DASH-IF's Excellence in DASH Award 2020.

4 ITU-T Rec. P.1204.3 Description

The next generation of video quality models were developed as a follow-up to P.1203 in collaboration with the Video Quality Experts Group (VQEG). The standards were published as **ITU-T Rec. P.1204** in 2019:

- ITU-T P.1204.3: Bitstream-based model
- ITU-T P.1204.4: Pixel-based (Full Reference)
- ITU-T P.1204.5: Hybrid model

As can be seen, new model types were considered: instead of only metadata- and bitstream-based models, there are now pixel-based and hybrid models standardized in the same family. Some parts of the standard, such as the metadata-based model, are not available yet and still under development, and are planned to be released in the future.

The newly developed standards enhance the scope of the previously developed models (from P.1203). The new scope includes:

- 4K/UHD video resolution
- new codecs (added support for H.265/HEVC, VP9)
- higher frame rates (up to 60 fps)
- higher bit depths (up to 10 Bit)

In the following, we focus on the P.1204.3 bitstream-based model.

4.1 Bitstream-based Model

P.1203.4 is a bitstream-based model that does not require access to the original source file for calculation. This makes the application much simpler and resource-efficient. Also, decoding of the video is not required.

The model itself operates on features extracted from the video bitstream, and it combines classical approaches for determining video quality with machine learning in order to improve its prediction accuracy. The bitstream is parsed, and features about its quantization parameters, motion vectors, frame sizes etc. are extracted. The P.1204.3 model then integrates those features into a final MOS per video sequence.

4.2 Model Performance

The P.1204 models have been validated in the context of the official ITU-T competition, using multiple subjective databases: 13 training databases were created for developing the models; 13 validation databases were used to validate their performance.

Due to confidentiality agreements in place between the involved parties, the databases or the model performance derived from the official ITU-T databases cannot be shown at this time.

As an alternative, we provide an in-depth analysis of the performance of P.1204.3 compared to popular metrics like PSNR, SSIM, MS-SSIM and VMAF. These comparisons have been performed on the publicly available, dataset AVT-VQDB-UHD1 that was not part of the ITU-T training and validation datasets used during P.1204's development. It consists of four different subjective tests with a total of 756 evaluated sequences, resulting in 19,620 human ratings.

The tests are summarized in the below table:

	Test 1	Test 2	Test 3	Test 4
Sources	6	6	6	8
Codecs	3 (H.264, H.265, VP9)	2 (H.264, H.265)	2 (H.265, VP9)	1 (H.264)
Resolutions	4 (360p, 720p, 1080p, 2160p)	4 (360p, 720p, 1080p, 2160p)	4 (360p, 720p, 1080p, 2160p)	6 (360p, 480p, 720p, 1080p, 1440p, 2160p)
FPS	1 (60fps)	1 (60fps)	1 (60fps)	4 (15, 24, 30, 60fps)

	Test 1	Test 2	Test 3	Test 4
PVSs	180	192	192	192
Participants	29	24	26	25
Display	65” (Panasonic)	55” (LG OLED)	55” (LG OLED)	55” (LG OLED)

The P.1204.3 model was run on all sequences from the four databases. In addition, PSNR, SSIM, MS-SSIM and VMAF were calculated for the same sequences.

Relation between the metrics and subjective MOS

The following plots show the comparison of the predictions from each of these metrics and the actual subjective ratings for all the tested sequences, where each point corresponds to one sequence in the dataset.

As can be seen, P.1204.3 outperforms all other tested metrics, with a very high Pearson correlation (between model score and subjective MOS) of 0.94. From the set of other metrics, only VMAF reaches acceptable performance with 0.87. It is clearly noticeable that PSNR and (MS-)SSIM are not suitable for evaluating the quality of video sequences, given the high nonlinearity of the results.

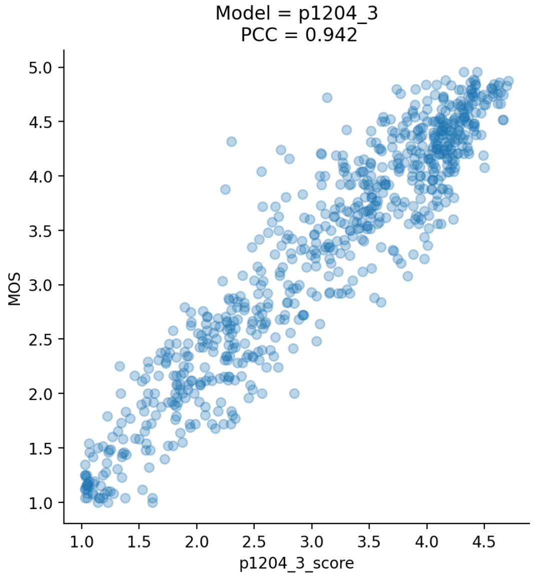

Figure 5: P.1204.3 performance.

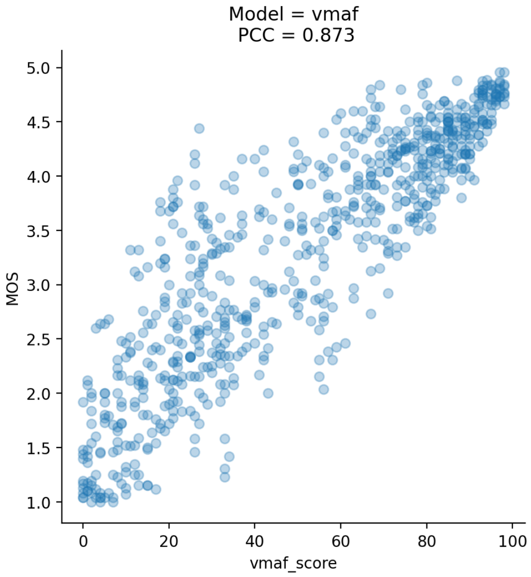

Figure 6: VMAF performance.

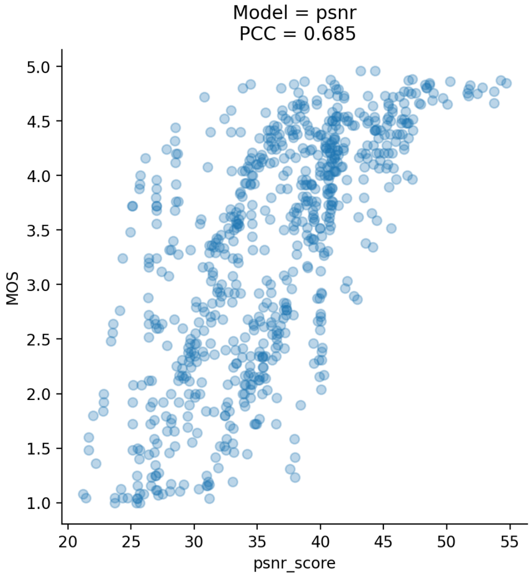

Figure 7: PSNR performance.

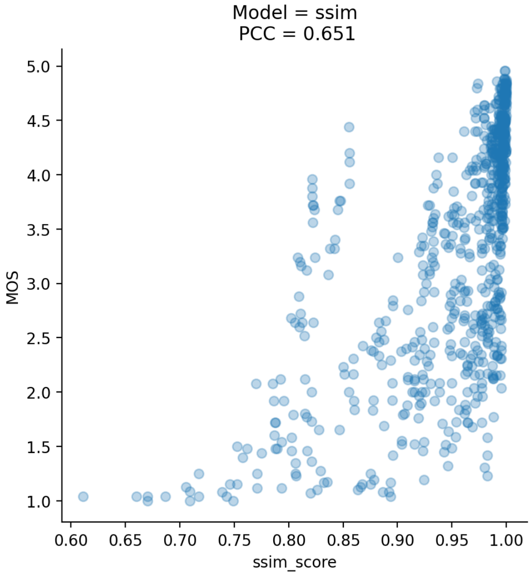

Figure 8: SSIM performance.

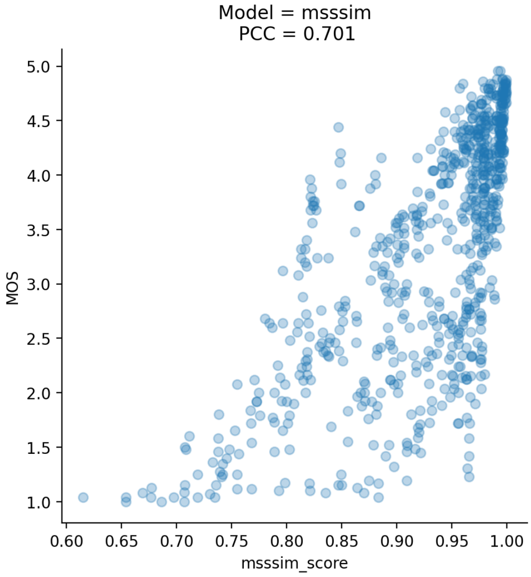

Figure 9: MS-SSIM performance.

Overall performance of P.1204.3 and other metrics

The following table shows the performance of the different metrics on the tested sequences, sorted by Root Mean Square Error (RMSE).

Metric	**RMSE** (lower is better)	**Pearson correlation** (higher is better)
P.1204.3	0.366	0.942
VMAF	0.53	0.873
MS-SSIM	0.774	0.701
PSNR	0.791	0.685
SSIM	0.824	0.651

4.3 Reference Software

A reference software for ITU-T Rec. P.1204.3 was developed by TU Ilmenau and is available online. It can be freely used for non-commercial research purposes.

5 Publications

We have published a number of papers in which P.1203 and P.1204.3 were described, used, or adapted:

- Rakesh Rao Ramachandra Rao, Steve Göring, Werner Robitza, Alexander Raake, Bernhard Feiten, Peter List, and Ulf Wüstenhagen. "Bitstream-based Model Standard for 4K/UHD: ITU-T P.1204.3 – Model Details, Evaluation, Analysis and Open Source Implementation." Twelfth International Conference on Quality of Multimedia Experience (QoMEX). Athlone, Ireland. May 2020.
- Werner Robitza, Alexander M. Dethof, Steve Göring, Alexander Raake, Tim Polzehl, and Andre Beyer. "Are You Still Watching? Streaming Video Quality and Engagement Assessment in the Crowd." Twelfth International Conference on Quality of Multimedia Experience (QoMEX). Athlone, Ireland. May 2020.
- Steve Göring, Christopher Krämmer, and Alexander Raake. "cencro – Speedup of Video Quality Calculation using Center Cropping." 21st IEEE International Symposium on Multimedia (2019 IEEE ISM). Dec 2019.
- 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." 21st IEEE International Symposium on Multimedia (2019 IEEE ISM). Dec 2019.
- 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." Electronic Imaging. 2019.
- 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." 9th ACM Multimedia Systems Conference. Amsterdam. 2018.

- 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." Tenth International Conference on Quality of Multimedia Experience (QoMEX). IEEE. 2018.
- Steve Göring, Alexander Raake and Bernhard Feiten. "A framework for QoE analysis of encrypted video streams." Ninth International Conference on Quality of Multimedia Experience (QoMEX). May 2017.
- 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." Ninth International Conference on Quality of Multimedia Experience (QoMEX). May 2017.
- Werner Robitza, Marie-Neige Garcia, and Alexander Raake. "A modular HTTP adaptive streaming QoE model - Candidate for ITU-T P. 1203 ("P. NATS')." Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017.

Other related publications in which P.1203 was used by other authors include the following:

- Susanna Schwarzmann, Nick Hainke, Thomas Zinner, Christian Sieber, Werner Robitza and Alexander Raake. "Comparing fixed and variable segment durations for adaptive video streaming: a holistic analysis." Proceedings of the 11th ACM Multimedia Systems Conference. 2020.
- Castillo Guzmán, Pau Arce Vila Paola, and Juan Carlos Guerri Cebollada. "Automatic QoE evaluation of DASH streaming using ITU-T Standard P.1203 and Google Puppeteer." Proceedings of the 16th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks. 2019.
- H-F. Bermudez et al. "Live video-streaming evaluation using the ITU-T P.1203 QoE model in LTE networks." Computer Networks 165 (2019): 106967.
- Michael Seufert, Nikolas Wehner, and Pedro Casas. "Studying the impact of HAS QoE factors on the standardized QoE model P.1203." 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2018.
- Satti, Shahid, et al. "P.1203 evaluation of real OTT video services." 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017.

6 Summary

ITU-T has recently published video quality models in the context of HTTP Adaptive Streaming, namely ITU-T Rec. P.1203, which integrates video quality and audio quality scores into a score for an (up to) 5 minute video session, including initial loading and stalling effects, and ITU-T Rec. P.1204, which is a set of high-performance models for UHD/4K 60 fps sequences coded with H.264/HEVC or VP9.

TU Ilmenau and partners have developed several tools in the context of the standardized models, which are ready to use and freely available for research purposes. There is:

- A reference implementation of the ITU-T Rec. P.1203 model
- A reference implementation of the ITU-T Rec. P.1204.3 bitstream model
- A bitstream parser for H.264, H.265 and VP9

All of these tools are available from the overview website, including additional links to databases and software.

Contact

TU Ilmenau

Fachgebiet Audiovisuelle Technik
Helmholtzplatz 2
98693 Ilmenau
Deutschland

Werner Robitza — `werner.robitza@tu-ilmenau.de`
Rakesh Rao Ramachandra Rao — `rakesh-rao.ramachandra-rao@tu-ilmenau.de`
Steve Göring — `steve.goering@tu-ilmenau.de`
Alexander Raake — `alexander.raake@tu-ilmenau.de`

tu-ilmenau.de/mt-avt