Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17741-4

Video expert assessment of high quality video for Video Assistant Referee (VAR): A comparative study

Kjell Brunnström[1,2] · Anders Djupsjöbacka[1] · Johsan Billingham[3] · Katharina Wistel[3] · Börje Andrén[1] · Oskars Ozolins[1] · Nicolas Evans[3]

Received: 19 July 2022 / Revised: 17 October 2023 / Accepted: 23 November 2023

Abstract

The International Football Association Board decided to introduce Video Assistant Referee (VAR) in 2018. This led to the need to develop methods for quality control of the VAR-systems. This article focuses on the important aspect to evaluate the video quality. Video Quality assessment has matured in the sense that there are standardized, commercial products and established open-source solutions to measure it with objective methods. Previous research has primarily focused on the end-user quality assessment. How to assess the video in the contribution phase of the chain is less studied. The novelties of this study are two-fold: 1) The user study is specifically targeting video experts i.e., to assess the perceived quality of video professionals working with video production. 2) Six video quality models have been independently benchmarked against the user data and evaluated to show which of the models could provide the best predictions of perceived quality. The independent evaluation is important to get unbiased results as shown by the Video Quality Experts Group. An experiment was performed involving 25 video experts in which they rated the perceived quality. The video formats tested were High-Definition TV both progressive and interlaced as well as a quarters size format that was scaled down half the size in both width and height. The videos were encoded with both H.264 and Motion JPEG for the full size but only H.264 for the quarter size. Bitrates ranged from 80 Mbit/s down to 10 Mbit/s. We could see that for H.264 that the quality was overall very good but dropped somewhat for 10 Mbit/s. For Motion JPEG the quality dropped over the whole range. For the interlaced format the degradation that was based on a simple deinterlacing method did receive overall low ratings. For the quarter size three different scaling algorithms were evaluated. Lanczos performed the best and Bilinear the worst. The performance of six different video quality models were evaluated for 1080p and 1080i. The Video Quality Metric for Variable Frame Delay had the best performance for both formats, followed by Video Multimethod Assessment Fusion method and the Video Quality Metric General model.

Keywords Contribution · Football · HDTV · PSNR · SSIM · Subjective and objective video quality · Video Assistant Referee (VAR) · Video quality · VIF · VMAF · VQM General · VQM_VFD

Extended author information available on the last page of the article

Published online: 27 December 2023

1 Introduction

With the approval of the International Football Association Board to allow trials for the use Video Assistant Refereeing (VAR) technology in the game of football the Fédération Internationale de Football Association (FIFA) expressed the need to develop technical guidelines (minimum requirements) for VAR systems to approve them for the use in the game. The key questions raised by FIFA in this relation were focused on the video quality i.e., the processing of images and challenges linked to coding and decoding, the synchronisation, and the re-formatting of broadcast feeds. RISE Research Institutes of Sweden in collaboration with FIFA, therefore, developed objective test methods that can be used to measure the latency, synchronicity, and video quality of VAR-systems at three different measurement points [1], see Fig. 1. The latency and synchronisation are not covered by this article, for further information see Brunnström et al. [1].

For measurement of video quality, the following Measurement Points (MP) were identified as being important:

- MP1 is where the camera signals enter the Video Operation Room and exit the VAR-system. MP2 is where the video is sent back to the Outside Broadcast van or to the broadcast provider.
- Encoding, conversion of different video formats and possible integration of different image sources into single video feeds i.e., the video quality of the resulting output of the VAR system at MP1.
- Measuring the output video quality from a VAR system back to the broadcaster for transmission on-air at MP2.

Video Quality assessment has matured in the sense that there are standardized, commercial products and established open-source solutions to measure video quality in an objective way [2–5]. Furthermore, the methods to experimentally test and evaluate the Quality of Experience (QoE) [6, 7] of a video are also widely accepted in the research community and in the broadcasting industry. These are in most cases based on standardized procedures [8–17].

Previous research and development have primarily focused on the end-user quality assessment. It is even specified in a Recommendation from the International Telecommunication Union (ITU) that a test persons should be naïve see e.g. [9]. This is not so surprising

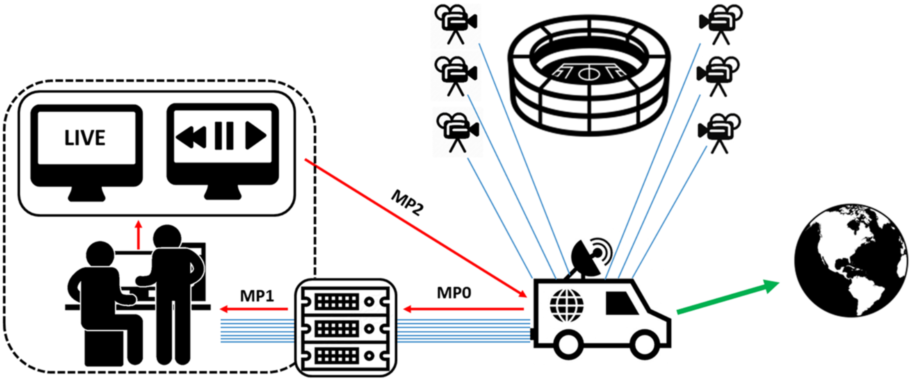

Fig. 1 Schematic overview of the VAR setup. Three Measurement Points (MP) are indicated MP0, MP1, and MP2, for the evaluation of latency, synchronization, and video quality

since it is where the video content is usually consumed and where most of the quality-related compromises must be done. However, the assessment and measurement of the video quality in the contribution phase is less studied. The European Broadcasting Union defines ‘High Quality’ as “a quality that is transparent to the intention of the programme maker and without visible artefacts, insofar as it is reasonably economic and technically practical to achieve.” [18].

The research questions that this article is addressing are:

1) What levels of video quality are considered good or better for expert video professionals especially targeting sports content?
2) Which objective quality models that have been developed by using video quality data from naïve test persons and most likely a wider and lower quality range, can well predict the quality in a high-quality range as defined by expert professionals from the broadcast industry?

The novelty and the aim of this research and paper is twofold:

- Firstly, a user study was conducted specifically targeting video experts. That is using professionals whose main occupation is video production. The study measured how they perceived the quality of the shown videos. This gives valuable insight into the video quality needed in the contribution phase.
- Secondly, the objective quality methods performance in predicting video experts’ opinions have not been performed previously. In this article, we are not presenting our own method in comparison to others. This is to be able to be unbiased and present an impartial evaluation. There is a tendency to publish the creators’ solution as the best performing method, with usually inflated prediction performances and it is unclear whether all methods that have been compared to had received the same chance. In independent evaluations, the performances usually drop drastically [19], compared to the performances in the articles where the methods were originally presented. This has also been shown repeatedly by the work of the Video Quality Experts Groups (VQEG) [20–27].

Video quality models are objective, mathematical models that approximate results from user quality assessments, in which human observers are asked to rate the quality of a video. In the literature various models and approaches are suggested [28, 29]. They are very often subdivided into groups depending on the amount of reference information that the models are using.

- Full Reference (FR) – access to high quality or non-distorted versions of the video to compare with the distorted video.
- Reduced Reference (RR) – key quality parameters are computed from the reference and the distorted video, which are compared [30].
- No Reference (NR) – no access to the reference [30].

Many models have been proposed in the literature, e.g., [31–34]. The Video Quality Experts Group (VQEG) has been involved in several efforts to evaluate the performance of models proposed [20, 21, 23–25, 29], which later has been standardized by the ITU [29, 31, 33]. In this research, we primarily focused on the FR models as they are more accurate than the NR models. Additionally, FR models are more likely to cover the high-quality range that is required for the use in VAR systems [35, 36]. A complication is that it requires a reference signal i.e., a video sequence taken at MP 0, that can be compared to the same video sequence at MP 1 and MP 2. For a working solution see [1]. In this article, we have chosen to present the performance of open source and published models. The performance of some commercial products were also investigated. As they did not perform better than the best-performing models in our study, we have made the decision not to disclose their names and performances.

The evaluation of the objective models in this work is based on data from a user study performed using standardized procedure as recommended by the ITU [8–10]. The performance comparison between the objective quality models was also following recommendations by the ITU [37].

The rest of the article is organized starting with on overview of related work in Section 2. The method used is then described in Section 3, followed by the Results in Section 4. The article is concluded with a Discussions and Conclusions in Sections 5 and 6 respectively. The abbreviations used are summarised in Table 1 and the symbols in Table 2.

2 Related work

Fully independent evaluations of objective video quality methods are rare. Many of the independent evaluations that have been conducted were done by VQEG and recently also by the ITU-T Study Group 12, either independent or in collaboration with VQEG. For the work of VQEG we will cover only those evaluations that most resembles the current work. Please, see www.vqeg.org for a full list of evaluations performed.

In VQEG phase I [25] objective video quality methods were evaluated for standard definition TV. It was a large test where about 26000 user opinions were collected, that was divided into four different parts 60 Hz high and low-quality and 50 Hz high and low quality. The high-quality range was with bitrates between 3 Mb/s and 50 Mb/s, whereas the low-quality range had bitrates between 768 kb/s and 4.5 Mb/s. There were nine methods evaluated. The effort was at the time seen as a failure as no method proved to be statistically better than Peak Signal to Noise Ratio (PSNR) [38]. The high-quality range of this work was similar to the current work, although with another video format Standard Definition TV (SDTV) as compared to High Definition TV (HDTV) in this study. The experimental method was also different as in VQEG phase I the Double Stimulus Continuous Quality Scale (DSCQS) [8] was used as compared to Absolute Category Rating (ACR) with Hidden Reference (ACR-HR) [9] removal that was used in this work. Furthermore, naïve observers were used in VQEG phase I and not video experts as in this article. The work was still very important as it showed very clearly for the first time the drop in performance of the video quality methods when they were subjected to a fully independent evaluation.

VQEG phase II [24] was not as large as phase I but was designed to be more challenging for the metrics. It did not contain the high-quality range of the videos. The experimental method was DSCQS and the observers were naïve.

VQEG HDTV [20] is in a sense the experiments most similar to current investigation. The video formats were 1080p at 25 and 29.97 frames-per-second and 1080i at 50 and 59.94 fields-per second. The encoding considered were MPEG 2 and H.264 with both encoding distortions and transmission errors. The bitrates range were between $1-30$ Mbit/s. The experimental method for getting the user rating was the same is in the current study i.e., ACR-HR. The observers were naïve as a contrast to the current study.

Wulf and Zölzer [39] evaluated three different objective quality metrics on three open video quality databases but did not perform their own user study. The majority of

Table 1 Abbreviations

ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
ANOVA	Analysis of Variance
AVC	Advanced Video Coding
BGR	Blue, Green and Red
DSCQS	Double Stimulus Continuous Quality Scale
FIFA	Fédération Internationale de Football Association
FR	Full Reference
H.264	Name of the codec specified in ITU-T Rec. H.264
HDTV	High-Definition TV
HRC	Hypothetical Reference Circuit
HSD	Honestly Significant Difference
ITU	International Telecommunication Union
JPEG	Joint Photographic Experts Group
MJPEG	Motion JPEG
MOS	Mean Opinion Scores
MP	Measurement Point
NR	No Reference
PCC	Person Correlation Coefficient
PSNR	Peak Signal to Noise Ratio
PSNR	Peak Signal to Noise Ratio
PVS	Processed Video Sequence
QoE	Quality of Experience
RMSE	Root Mean Square Error
RR	Reduced Reference
SCC	Spearman Correlation Coefficient
SDTV	Standard Definition TV
SOS	Standard deviation of Opinion Scores
SRC	Source video
SSIM	Structural Similarity Index
VAR	Video Assistant Refereeing
VIF	Visual Information Fidelity
VMAF	Video Multimethod Assessment Fusion
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
VQM_VFD	VQM Variable Frame Delay
yadif	Yet Another DeInterlacing Filter

Table 2 Symbols

a	A parameter that characterizes the relationship between SOS^2 and the MOS
p	Probability of significance
q	The ratio between the maximum RMSE squared and the minimum RMSE squared

the data comes from VQEG HDTV and the rest from Ecole Polytechnique Fédérale de Lausanne and Politecnico di Milano [40, 41] as well as Laboratory for Image and Video Engineering [42].

Sedano et al. [43] used the same databases as Wulf and Zölzer and evaluated six different full-reference methods for the purpose of developing a light-weight parametric no-reference model. The paper presented a method of using full-reference methods that was tested for a specific scope for the development of no-reference modules for that scope.

Raake et al. [29] did a comprehensive evaluation of Bitstream-, Pixel-Based and Hybrid Video Quality methods for the development of three standards for video quality assessment for video formats up to 4 K resolution involving video codecs such as Advanced Video Coding (AVC) (H.264) [44], H.265/HEVC [45] and VP9. The user studies were using ACR as in the current study, but the ratings were based on naïve observers.

Ling et al. [46] performed a study for high quality video involving both HD and UHD videos. The user studies used the Degradation Category Rating (DCR) method instead of ACR. Four different objective quality methods were independently benchmarked in the study. The viewers were non-experts.

Pinson [19] did a comprehensive independent evaluation of no-reference models that had been published in the literature. The general conclusion was that the performance that had been published by the authors of the models did not come near the performance when they were independently tested. The main differences here are the focus on no-reference models and video quality databases based on naïve observers.

In summary (see Table 3) previous studies have been based on using non-expert or naïve observers in the investigations. This is not surprising since this is a common practice and recommended by the ITU. In our case, the videos used were close in quality to what could be expected in the contribution phase in broadcasting, the appropriate observer should be an expert in judging video quality.

3 Method

3.1 Video quality user study

3.1.1 Procedure

For high video quality where the differences are small, there are arguments for a comparative methods like the Double Stimulus Impairment Scale (DSIS) or Double-stimulus continuous quality-scale (DSCQS) [8]. However, these will reduce the number of stimuli or videos that can be scored during an experiment, since each trial takes either almost double the time for DSIS or up to four times as long when using DSCQS [8], as compared to a single stimulus method. It is also important to get a wide variety of video rated to have a rich data to evaluate objective methods against. We, therefore, decided to use a single stimulus method: ACR-HR method [8–10]. It was shown by Lee et al. that the results correlate well with DSCQS [11]. This approach has been successfully employed previously e.g., VQEG HDTV [20].

This method uses single stimulus procedure. One video is presented at a time to the user and they are asked to provide their rating on a five graded category scale for each video after the video stops, see Fig. 2 The ratings were provided in this study via a voting interface on the screen, see Fig. 3, asking the user to “judge the video quality of the video?” (In

Table 3 Summary of main pros and cons of previous studies. The pros and cons are here meaning suitability for our purpose and cons are not pointing out flaws in the investigation

Study	Year	Pro	Cons
VQEG Phase I [25]	2000	Large investigation also involving bitrates comparable to current study. Data published	Naïve observers. STDV. Different experimental methods
VQEG Phase II [24]	2003	More focused investigation than VQEG Phase I	Naïve observers. STDV. Different experimental methods. Videos were not published
VQEG HDTV [20]	2010	HDTV video format both progressive and interlaced. Same experimental method. Data partly made public	Naïve observers. Lower bitrates
Wulf and Zölzer [39]	2012	Partly HDTV	Did not perform their own study and based on naïve observers
Sedano et al. [43]	2014	Partly HDTV	Did not perform their own study and based on naïve observers
Raake et al. [29]	2020	Both HDTV and UHD. Comprehensive study	Naïve observers. Lower bitrates. Data not published
Ling et al. [46]	2020	Both HDTV and UHD. Target high quality assessment	Naïve observers. Different experimental methods
Pinson [19]	2022	Comprehensive study. Show significant performance reduction for independently evaluated methods	Targets no-reference methods. Did not perform their own study

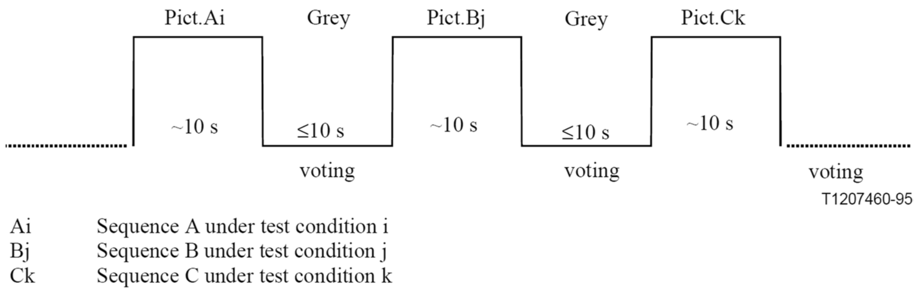
Pict.Ai
Grey
Pict.Bj
Grey
Pict.Ck
~10 s
≤10 s
~10 s
≤10 s
~10 s
voting
voting
voting
T1207460-95
Ai Sequence A under test condition i
Bj Sequence B under test condition j
Ck Sequence C under test condition k

Fig. 2 Stimulus presentation in the Absolute Category Rating (ACR) method (Figure borrowed from ITU-T P.910 [8])

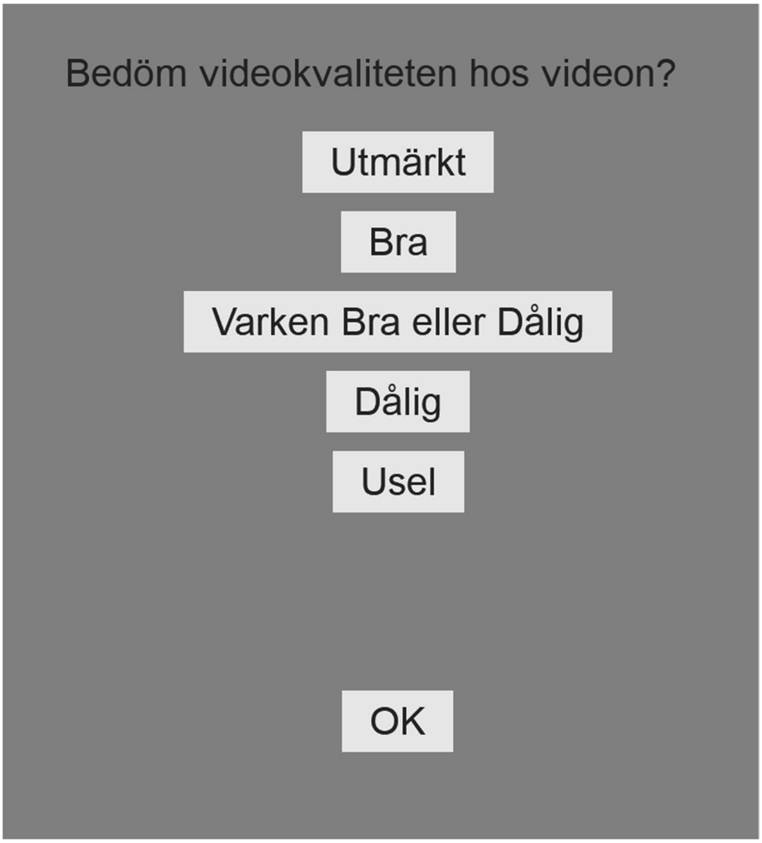
Bedöm videokvaliteten hos videon?

Fig. 3 The voting interface used in the experiment

Swedish: “Bedöm videokvaliteten hos videon?”). The rating scale used was the five graded ACR quality scale [8, 9] (Swedish language translation in parentheses):

5 Excellent (Utmärkt)
4 Good (Bra)
3 Fair (Varken Bra eller Dålig)
2 Poor (Dålig)
1 Bad (Usel)

High-quality or pristine reference videos were mixed in among all the other videos and rated as the others. Viewers saw each presentation of a video once and did not have the option of re-playing a presentation. However, they could change their voting as many times as they wanted before submitting their final answer, as described in ITU-T P.913 [10].

To evaluate objective video quality models for different formats that are used in the TV production of football games the users were asked to provide their rating on three different video formats:

1) Full size 1920×1080 video based on progressive source (1080p).
2) Full size 1920×1080 video based on interlaced source (1080i).
3) Quarter size 960×540 video based on interlaced source (540i).

The order in which the different video formats were played to the users, as well as the order of the single video sequences within each video format were randomized for each participant. For the video playback and randomisation, the VQEGPlayer [32] was used. The time required by the users to watch the videos and provide the rating for all three sessions was in total approximately 45 min, with short breaks between each session. The total time required for each user, including instructions, visual testing, training, pre-, and post-questionnaire, was about 1.5 h.

There were 60 Processed Video Sequences (PVS) to be evaluated per session. These consisted of 6 different SouRCe sequences (SRC), i.e., different content that each of them was processed with 10 different error conditions, so called Hypothetical Reference Circuits (HRC), see also below in Sessions 3.1.4 and 3.1.50. Each video was 10 s and with an average estimated voting time of 5 s, a trial was about 15 s. This gave a total time for a session of approximately 15 min.

Instructions were written out for the participants to read, to ensure that the instructions given were as similar as possible. Some explanations and backgrounds were given verbally, especially in response to any questions and uncertainty of the task to perform. Before the actual test a training session was carried out to familiarize the test persons with the procedure and to show the range of qualities involved. A very good, an intermediate and a bad quality version of a video that was not used in the actual test were shown to the test persons. Test persons were allowed to ask questions after the training session if something was still unclear.

3.1.2 Set-up of test room

To create a controlled and uniform environment for the participants the test room was set-up to comply with the requirements of the ITU-R Rec. BT.500-15 [8]. A high-end consumer-grade 65″ 4 K TV (Ultra HD, LG OLED65E7V) was used for the experiments, having a resolution of 3840×2160 pixels. As the videos used in the experiment had a lower

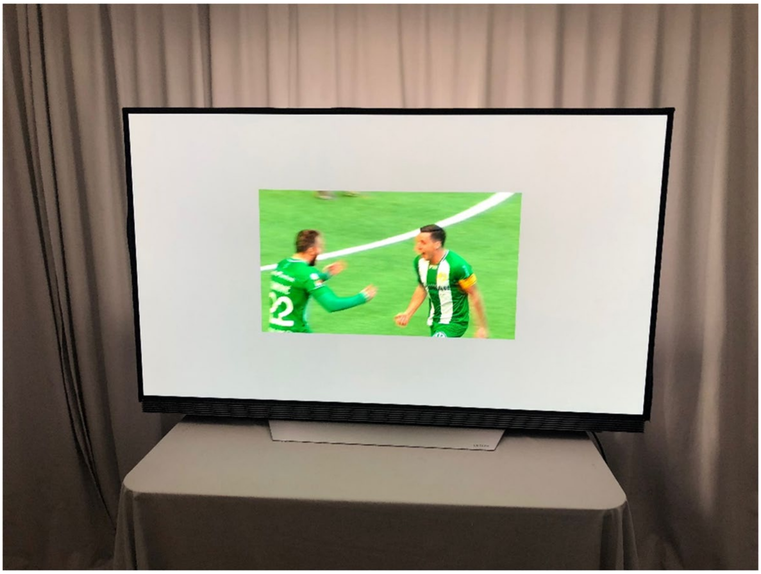

Fig. 4 The video with resolution 1920×1080 pixels was presented in the centre of the screen with a grey surround

resolution (1920×1080 and 960×540) than the screen, the video was displayed pixel matched in the centre of the screen with a grey surround, see Fig. 4. The interlaced 1080i video was deinterlaced in software (see Section 3.1.5) and the deinterlacing of the TV was not used. The TV was characterized e.g., luminance, contrast, colours, and colour temp, and verified to conform to ITU-R BT.709 [47].

Different video formats and display resolution have been optimized to be viewed from a specific viewing distance, roughly corresponding to an angular pixel pitch of 1 min of arc. For HD (1920×1080) this is 3H, where H=Picture Height (height of the video window, not the physical display), corresponding to 120 cm in our case. The users were requested to keep this viewing distance while evaluating the videos. This distance was marked on the floor and the chair, which was an office chair on wheels that had been equipped with a wheel locking mechanism. The position of each participant was also adjusted, so that their eyes were vertically and horizontally centred in relation to the TV screen centre when looking straight ahead.

3.1.3 Test persons

In the experiment, 25 Swedish-speaking video experts participated, see further Section 4.1.1.

The term ‘expert’ has been used in the sense that the viewers’ work involves video picture quality, picture production in broadcasting, creation of film or video, post-production, etc. They were recruited by an external company Feebackfrog AB (www.feedbackfrog.com), which has been used previously by RISE for recruitment of participants in Usability testing. Feedbackfrog took all the contacts with the participants, scheduled the time to experiment, and, finally, also expedited the compensation after the completion of the experiment. The anonymity of the data collected was, therefore, ensured as no contact information of the participants were digitally stored at RISE. The participants were compensated with gift cards to a value of 1500 SEK (about 150 EUR) each.

All viewers were tested prior for the following:

- Visual acuity with or without corrective glasses (Snellen test).
- Colour vision (Ishihara test).

3.1.4 Source video sequences

To rate the video quality a set of six different source video sequences was shown to the expert panel. The video formats selected for the SRC were:

- 1920 × 1080 progressive 50 frames-per-seconds (1080p)
- 1920 × 1080 interlaced 50 fields-per-seconds (1080i)

All SRC (Tables 4 and 5) were obtained as uncompressed videos during live football broadcast productions, as well as from Swedish Television's, "Sveriges Television (SVT)", production Fairytale that was produced for research and standardization purposes [48]. From all collected videos, video clips of the length of 14 s were extracted.

The collection of the football related videos took place in Madrid, at the Estádio Municipal de Butarque on 30 April 2019, during a FIFA test event for Virtual Offside Line Technology where both video formats, 1080p and 1080i, were recorded. The second football video was collected during a match of the Swedish national football League at the Tele2 Arena in Stockholm on 14th of May 2019. Due to the broadcasting of the football match that was conducted in 1080i format only this video format was collected. In both cases the equipment to collect the videos was a Blackmagic Design UltraStudio HD Mini grabbing box connected to a Lenovo Carbon X1 6th gen computer (CPU: Intel Core i7-8550U, RAM: 16 GB) and an external 500 GB SSD hard drive (Samsung X5) with Thunderbolt connection. The software used was Blackmagic Design Media Express Version 3.5.7 and the format was uncompressed AVI 16 bit 4.2.2 uyvy.

The SRC produced from the SVT production are clips from a 6.5-min-long video Fairytale [48], that was professionally filmed and produced on 65 mm analogue film in 50 fps (slow motion up to 100 fps) and scanned frame by frame while colour correcting and applying film grain noise reduction, to produce the 4 K (3840 × 2160 progressive, 16 bit per colour) Master. The 1080p version was produced by downsampling the Master using a sinc filter.

The interlacing was also carefully done, by starting with a 2164-line (3840 × 2164p/50) raster. Every second frame was shifted two lines downwards and then they were cropped to 2160 lines. The shifted half frames were filtered down to 540 each, by Shake's box filter vertically and Shake's sinc filter horizontally to reduce the resolution to 1920 [49, 50]. For more details on the production see [48].

When selecting the SRC the focus was primarily on football content. To ensure diversity in the video sequences the content was mixed with the Fairytale content. To be representative for the future use of the test method, in the analysis of the video quality of VAR

Table 4 Description of the source video sequence of the 1080p format

Name	*Description*	*Origin*	*Image*
Butarque_follow_ball_2 (SRC1)	The camera follows the football in an almost empty stadium. Very few people. Motion mostly camera pan.	Grabbed at Estadio Municipal de Butarque.	
Butarque_follow_player_2 (SRC2)	The camera follows one football player in an almost empty stadium. Very few people. Motion mostly camera pan.	Grabbed at Estadio Municipal de Butarque.	
Butarque_passing_1 (SRC3)	Three players are passing the football. Some cameras pan and zoom at the same time as players are moving around.	Grabbed at Estadio Municipal de Butarque.	
SVT_CrowdRun (SRC4)	Many people running at the same time. A lot of spatial details and motion. Almost a static camera. This sequence is defined in [48] and classified as “difficult” on coding difficulty.	SVT Fairytale [48].	
SVT_RunRun (SRC5)	Many people running at the same time. Slow-motion, scene change. This sequence is not defined in [48].	SVT Fairytale [48].	
SVT_ParkJoy (SRC6)	A few people running at a distance, the camera is moving with them, and trees are passing in the foreground. High contrasts. This sequence is defined in [48] and classified as “difficult” on the coding difficulty.	SVT Fairytale [48].	
SVT_Searching (used only in the training session)	Some people move back and forth. A lot of leaves and flowers. Colourful. This sequence is not defined in [48].	SVT Fairytale [48].	

Technology, a special requirement for the videos were that they had to contain overall a lot of motion to be representative for football.

Final SRC were cut to 10 s, but the source scenes used for PVS creation were 14 s to accommodate for any transient behaviour of the encoder.

Table 5 Description of the source video sequence of the 1080i format

Name	*Description*	*Origin*	*Image*
Football_celebrating_1 (SRC1)	The camera follows a player after he has scored a goal. One person up to a group of people. Close-ups. Motion mostly camera pan.	Grabbed at Tele2 Arena.	
Football_goal_1 (SRC2)	Overview over football goal area, with several players there. Many people at a distance. Motion from camera pan and player's motion.	Grabbed at Tele2 Arena.	
Football_greetings (SRC3)	Players are greeting the referees; children are greeting each other, and some players are moving in the background. Fixed camera, close-ups, and motion primarily from people motion.	Grabbed at Tele2 Arena.	
Football_play_1 (SRC4)	Overview over football goal area, with several players there. Many people at a distance. Motion from camera pan and player's motion.	Grabbed at Tele2 Arena.	
SVT_CrowdRun (SRC5)	Many people running at the same time. A lot of spatial details and motion. Almost a static camera. This sequence is defined in [48] and classified as "difficult" on coding difficulty.	SVT Fairytale [48].	
SVT_ParkJoy (SRC5)	A few people running at a distance, the camera is moving with them, and trees are passing in the foreground. High contrasts. This sequence is defined in [48] and classified as "difficult" on the coding difficulty.	SVT Fairytale [48].	
Football_goal_2(used only in the training session)	Overview over football goal area, involving several players there. Many people at a distance. Motion from camera pan and player's motion.	Grabbed at Tele2 Arena.	

3.1.5 Processed video sequence generation

A Hypothetical Reference Circuit (HRC) is a processing performed by a system or software that can introduce degradations to the video e.g., video encoding, scaling, transmission, etc. [10]. The processing performed in this experiment has been similar for the different formats that were used in the experiment, but with differences adjusted to the format as well as targeting specific aspects of the format. There were 10 different HRCs per video format (including the reference) and each HRC was applied to each SRC for each of the formats, making 60 PVSs per format [10]. All PVSs was 10 s long.

A summary of the HRCs is the following:

- 1080p: H.264 (80 Mbit/s – 10 Mbit/s) and Motion JPEG (80 Mbit/s – 20 Mbit/s), see also Table 6.
- 1080i: H.264 (50 Mbit/s – 10 Mbit/s), Motion JPEG (80 Mbit/s – 20 Mbit/s) and bad deinterlacing, see also Table 7.
- 540i: H.264 (50 Mbit/s – 10 Mbit/s) and different scaling algorithms see also Table 8.

Table 6 The HRCs or processing applied to the 1080p format

HRC	Processing	Bitrates	Actual processing
1	No processing	Uncompressed	
2	Motion JPEG	80 Mbit/s	FFMPEG -c:v mjpeg -b:v 80 M
3	Motion JPEG	60 Mbit/s	FFMPEG -c:v mjpeg -b:v 60 M
4	H.264	80 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 80 M
5	H.264	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50 M
6	H.264	30 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 30 M
7	H.264	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20 M
8	Motion JPEG	40 Mbit/s	FFMPEG -c:v mjpeg -b:v 40 M
9	Motion JPEG	20 Mbit/s	FFMPEG -c:v mjpeg -b:v 20 M
10	H.264	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10 M

Table 7 The HRCs or processing applied to the 1080i format

HRC	Processing	Bitrates	Actual processing
1	No processing	Uncompressed	
2	Motion JPEG	80 Mbit/s	FFMPEG -c:v mjpeg -b:v 80 M
3	H.264	80 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 80 M
4	H.264	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50 M
5	H.264	30 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 30 M
6	H.264	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20 M
7	Motion JPEG	40 Mbit/s	FFMPEG -c:v mjpeg -b:v 40 M
8	Motion JPEG	20 Mbit/s	FFMPEG -c:v mjpeg -b:v 20 M
9	H.264	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10 M
10	Deinterlacing	Uncompressed	Blend fields, double the frame rate, top-field-first, VirtualDub (version 1.10.4)

Table 8 The HRCs or processing applied to the 540i format

HRC	*Processing*	*Bitrates*	*Actual processing*
1	Scaling (lanczos)	Uncompressed	FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags lanczos
2	H.264 + Scaling (lanczos)	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags lanczos
3	H.264 + Scaling (bilinear)	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags bilinear
4	H.264 + Scaling (neighbor)	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags neighbor
5	H.264 + Scaling (lanczos)	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags lanczos
6	H.264 + Scaling (bilinear)	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags bilinear
7	H.264 + Scaling (neighbor)	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags neighbor
8	H.264 + Scaling (lanczos)	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags lanczos
9	H.264 + Scaling (bilinear)	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags bilinear
10	H.264 + Scaling (neighbor)	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10 M FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vf scale = 960:-1 -sws_flags neighbor

The experimental program VQEGPlayer [51], takes the raw format Blue, Green and Red (BGR) and not AVI as a file format, and a colour space conversion was, therefore, applied to PVS:s as a the last step. In addition, the interlaced video clips (1080i and 540i) were deinterlaced in software.

Deinterlacing was performed using FFMPEG's [52] Yet Another Deinterlacing Filter (yadif) mcdeint (motion compensating deinterlacing) "slow processing option", with the command below.

- FFMPEG -vcodecrawvideo -pix_fmt uyvy422 -vfyadif=1:0:0,mcdeint=3:0:1

The last step in the preparation of PVS was performed by the program iconvert, written by Prof. Marcus Barkowsky, Deggendorf Institute of Technology (DIT), Germany. The program performed format and colour space conversion, as well as cutting off the 2 first and 2 last seconds of the PVS down to 10 s.

For 1080p and 1080i the following command was used

- iconvert.exe uyvy422:$outfile1@1920×1080,frames=100-600 -fyuv444_to_rgb888@mode=709

For 540i the following command was used

- iconvert.exe uyvy422:$outfile1@960×540,frames=100-600 -fyuv444_to_rgb888@mode=709

A summary of hardware and software used can be seen in Table 9.

3.2 Objective video quality assessment methods

In the evaluation we were interested to study how well the models could predict the data obtained from the users. We have followed standardised procedures from the ITU in this regard [37]. We studied the overall performance by calculating the Pearson Correlation Coefficient (PCC) [37] and the Root Mean Square Error (RMSE) [37], between the user data Difference Mean Opinion Scores (DMOS) and the scores of the objective models i.e. the Predicted DMOS (PDMOS). The PCC measures the linear relationships between the DMOS and the PDMOS. The RMSE gives the accuracy.

The DMOS was calculated by subtracting for each user its rating of the reference from the rating of the distorted video. To get the values on the same scale as the Mean Opinion Scores (MOS) i.e., 1–5, the following formula was used:

$$Dscore_{kij} = 5 - \left(SRC_k - PVS_{kij}\right) \vee k = 1 \dots P, i = 2 \dots Q, j = 1 \dots N \tag{1}$$

Here the index k was running over all SRCs $(P=6)$ and i over all HRCs $(Q=10)$ and j over all participants $(N=25)$. The reference for source k was denoted $SRC_k=PVS_{k1j,}$ which was then the PVS for the HRC with index 1.

$$DMOS_{ki} = \frac{1}{N} \sum\nolimits_{j=1}^{N} Dscore_{kij} \tag{2}$$

Table 9 Hardware and software used

Hardware	Type	Software	Settings
High-end consumer-grade 65″ 4 K TV	LG OLED65E7V		Resolution: 3840×2160; Colour temp: 6500 K; Gamma: 2.2; Peak white full screen: 300 cd/m^2; Frame rate. 50 Hz
User test computer	CPU Intel Core i7-6850 K, RAM 24 GB	Windows 10. VQEGPlayer	Playout resolution: 1920×1080 or 960×540 centred with grey surround; Frame rate: 50 Hz
Video grabbing hardware	Blackmagic Design UltraStudio HD Mini	Blackmagic Design Media Express, Version 3.5.7	Avi format uncompressed 8 bit 4:2:2
Grabbing computer	Lenovo Carbon X1 6th gen computer (CPU: Intel Core i7-8550U, RAM: 16 GB)	Windows 10	

The PCC measures the linear relationship between the model scores and the DMOS. For simplicity we drop the index to keep track of the different sources and let the index i be taken over all PVSs i.e., $i = 1 \ldots M$. where $M = 6 \cdot 9 = 54$.

$$PCC = \frac{\sum_{i=1}^{M}(DMOS_i - \overline{DMOS}) \cdot (PDMOS_i - \overline{PDMOS})}{\sqrt{\sum_{i=1}^{M}(DMOS_i - \overline{DMOS})^2} \cdot \sqrt{\sum_{i=1}^{M}(PDMOS_i - \overline{PDMOS})^2}} \tag{3}$$

where $\overline{DMOS}$ and $\overline{PDMOS}$ were the mean values of respective entity. The calculation was done using the Matlab function corr [53]. The RMSE was calculated using:

$$RMSE = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(DMOS_i - PDMOS_i\right)^2} \tag{4}$$

where M is the number of points that are compared and minus 1 is to get an unbiased estimate of the RMSE according to ITU-T Rec. P. 1401 [37].

As the relationships very often are not linear it is recommended to linearize the dependency by fitting a 3rd order monotonic polynomial to the data [37]. We also use this to map the value range of the PDMOS into the range of the DMOS i.e., 1 to 5 so that RMSE can be calculated. According to [37] we have used to following mapping:

$$PDMOS = a + b \cdot PDMOS_{raw} + c \cdot PDMOS_{raw}^2 + d \cdot PDMOS_{raw}^3 \tag{5}$$

where

$$RMSE(DMOS, PDMOS) \rightarrow min\ and\ f\left(PDMOS_{raw}\right) = monotonous\ between\ PDMOS_{min} and\ PDMOS_{max}.$$

Finally, a statistical hypothesis test was applied to the RMSE values. The null hypothesis, H_0, was that there was no statistical difference between two RMSE values, and the alternative hypothesis, H_1, was that there was a statistical difference. The test was based on forming an F ratio between the larger RMSE value squared divided with the smaller RMSE value squared. The degrees of freedom was the number of points in the RMSE calculation, minus 4 due to the 3rd order monotonic polynomial fit i.e. 54 – 4 = 50 [37].

$$q = \frac{RMSE_{max}^2}{RMSE_{min}^2} \tag{6}$$

The q-value was then compared to $F(0.05, 50, 50)$ for 95% confidence and if q > F the difference was considered significant. We used Holm [54] to handle multiple comparison. The Spearman Correlation Coefficient (SCC) was also calculated [55], using Matlab function corr [53].

Another criterion for a useful method is whether it can handle registering the distorted video with the reference. That is if some offset has occurred either spatially or temporally by the distortion introduced. This can have a severe impact on some computations. For instance, the Peak Signal to Noise Ratio (PSNR) computes the difference pixel by pixel and a small shift can, therefore, have a big impact on the score, which is not reflected by a quality difference. The Video Quality Metric (VQM) [31] software also contains a rather sophisticated registration algorithm that could be run separately from the core algorithm and could, therefore, be used for pre-processing of videos to evaluate the performance of other algorithms. This registration method has been standardized by the ITU (ITU-T Rec J.144) [5].

3.2.1 Video Multimethod Assessment Fusion

Video Multimethod Assessment Fusion (VMAF) [3] was developed by Netflix and the University of Southern California. It is not standardised but has shown good performance in recent evaluations. The method is available as open-source software (https://github.com/Netflix/vmaf/releases). VMAF fuses the results from more basic image quality-based metrics, but also from motion estimation calculations. The fusing was done using a Support Vector Machine (SVM) regressor [56]. In [3] presents three metric used:

- Visual Information Fidelity (VIF) – see below. In VMAF the metrics output from different scales are kept separate and not combined over four scales as in the original formulation.
- Detail Loss Metric (DLM) [57] – this is an algorithm that measures loss of useful information and added impairments that will distract the user. In the original algorithms a combined score is produced, but in VMAF the elementary metrics are used.
- Motion is temporal estimation based on the average absolute pixel difference of the luminance component between consecutive frames.

The version used was 1.3.15 with the following command: vmafossexec.exe yuv422p 1920 1080 reference.yuv distorted.yuv.\vmaf-1.3.15\model\vmaf_v0.6.1.pkl –log results.csv –log-fmt csv.

3.2.2 Video Quality Metric

The Video Quality Metric (VQM) was developed by the Institute for Telecommunication Sciences (ITS), the research laboratory of the National Telecommunications and Information Administration (NTIA) (USA). It was standardized by the ITU for SDTV (ITU-T Rec. J.144) [5], and has shown very good performance, especially the updated version VQM for Variable Frame Delay (VQM_VFD) [2]. It is available as open-source software.

The VQM General model was designed to be a reduced reference model but has in most cases been compared to FR-models and the performance has been as good or even better. It was designed for a wide variety of video systems and bitrates. VQM was also tested and evaluated extensively during its development phase [31]. The model was developed and optimized for the codecs and video system of its time i.e., before 2004. After performing a rigorous calibration that will contain spatial and temporal alignments, calculations of valid region as well as gain and level offset estimation, the video quality is estimated as a linear function of some quality features.

$$\begin{aligned} VQM\_General = & -0.2097 \cdot si\_loss + 0.5969 \cdot hv\_loss + 0.2483 \cdot hv\_gain + 0.0192 \cdot chroma\_spread \\ & + 2.3416 \cdot si\_gain + 0.0431 \cdot ct\_ati\_gain + 0.0076 \cdot chroma\_extreme \end{aligned} \tag{7}$$

where *si_loss* is the loss of spatial information such as blur; *hv_loss* is the degree of that the horizontal and vertical edges have shifted towards a more diagonal orientation; *hv_gain* detects the opposite shift of edges compared to *hv_loss*; *chroma_spread* computes the changes in the spread of 2D colour samples; *si_gain* measures quality improvements based on e.g. edge sharpening and *chroma-extreme* finds colour impairments due to transmission errors. The command used for VQM General (version 30) was: matlab.exe -r cvqm (‘reference.avi’, ‘distorted.avi’, ’progressive’, ’frcal’, ’general’).

VQM_VFD was designed to cope with videos that can have dropped frames or in other ways have variations in their frame rate. The key difference with this updated model is the neural networks module weighting together the extracted perceptual features of the model [2], as shown in Equation above.

The command used for VQM_VFD (version 12) was: matlab.exe -r vfd (‘distorted.avi’, ‘reference.avi’, ’progressive’, ’results.txt’).

3.2.3 Peak Signal to Noise Ratio

Peak Signal to Noise Ratio (PSNR) is the most used video quality assessment method. It is standardized by ITU in ITU-T Rec J.340 [38]. It has been disproven many times for not providing a good estimate of perceived visual quality. It has merits in the high-quality range, but its output values cannot be compared between different source content, which makes it hard to attach requirement levels to that the system should fulfil.

Open source implementations are provided by Ecole Polytechnique Fédérale de Lausanne with VQMT: Video Quality Measurement Tool (VQMT) [58] and within the VQM software [59].

This implementation calculates PSNR frame by frame by first forming the Mean Square Error of the difference between the reference and the distorted frame

$$MSE = \frac{1}{m \cdot n} \sum\nolimits_{i=0}^{m-1} \sum\nolimits_{j=0}^{n-1} \left[reference(i,j) - distorted(i,j)\right]^2 \tag{8}$$

$$PSNR = 10 \cdot log_{10}\left(\frac{255^2}{MSE}\right)dB \tag{9}$$

The calculations were performed in the colour space YUV 4:2:2. The different channels carried equal weights in the calculations. The command used for calculating PSNR was: vqmt.exe reference_video.yuv distorted_video.yuv 1080 1920 500 yuv422 results.txt PSNR.

The overall result was obtained by taking the average over all frames.

3.2.4 Structural Similarity Index

The Structural Similarity Index (SSIM) is an image quality method that has been popular since it was first published [32]. It was developed to improve compared to PSNR but is still very simple. The idea is to compare the structure between two images. It is defined for images and does not specify how it should be used for video. The paper [32] is one of the most cited papers in the image processing community. Here we have used the implementation in VQMT [58]. A usual form to calculate the SSIM is

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{10}$$

where x and y represent the different images that are compared, μ is the mean and σ standard deviation. C_1 and C_2 are constants to avoid instability in the calculations. The command used was the same as above for PSNR and calculated at the same time and the colour space was treated the same. The overall result was obtained in the same way by taking the average over all frames.

3.2.5 Visual Information Fidelity

Visual Information Fidelity model has an information-theoretic approach to the image and video quality assessment problem. The model combines an image fidelity criterion of shared image information between the reference and the degraded image, with a measure of the image information that is present in the reference and calculates how much of it that can be extracted from the degraded image [34]. Here we have used the implementation in VQMT [58], which implements the pixel domain version.

In the VIF method the image is divided into subbands for instance with the discrete wavelet transform [60, 61] or using the method of steerable pyramid [62]. A VIF metric is calculated for each subband using

$$VIF_i = \frac{\left(2 \cdot \mu_{ref} \cdot \mu_{dist} + C1\right) \cdot \left(2 \cdot \sigma_{ref,dist} + C2\right)}{\left(\mu_{ref}^2 + \mu_{dist}^2 + C1\right) \cdot \left(\sigma_{ref}^2 + \sigma_{dist}^2 + C2\right)} \tag{11}$$

where μ_{ref} for the subband the average values of the luminance of the reference and μ_{dist} the same for the distorted image. $\sigma_{ref,dist}$ is the covariance between the reference and the distorted image. σ_{ref} and σ_{dist} are the standard deviations for the respective images for the subbands. The total VIF is obtained by multiplying all the subband VIFs together. For N subbands

$$VIF = \prod\nolimits_{i=1}^{N} VIF_i \tag{12}$$

The command used was the same as above for PSNR and calculated at the same time and the colour space was treated the same. The overall result was obtained in the same way by taking the average over all frames.

4 Results

4.1 Video quality user study

4.1.1 Characterisation of expert panel

In total 25 video experts participated: 23 males and 2 females. The number of females was low, but this was expected considering the area of investigation. The average age of the test users was 37.8 years, with a standard deviation of 10 years. The oldest was 65 years of age and the youngest was 24 years. All participants were working in or studying in the TV area as producers, technicians, and photographers, or similar, thus considered as experts in evaluating video quality. This was shown in their answer to the question of their experience of video evaluation where 1 meant “no experience” and 5 meant “expert”. This question had an average of 4.5 and a standard deviation of 0.58 and only two users scored below 4 one put 3.5 and the other 3.

Very few had the experience of being scientific video test viewers. The average was as low as 1.48 (standard deviation 0.75) where 1 meant no experience and 5 a lot. Only one subject put a 4 and one put a 3. About 1/3 of the subjects put 1.

All test persons had a good visual acuity as expected for such professionals, average 1.09/1.06 (right/left eye), standard deviation 0.18/0.20, max 1.4, and min 0.6 on one eye. About half of them wore glasses or lenses. All had an accurate colour vision.

The test persons were asked to rate how typical they experienced the distortions in the videos from 1 to 5, where 1 was not at all and 5 was very typical. The mean was 4 with a standard deviation of 1, showing that the test persons were very familiar with the distortions.

The test persons were also asked if the range of resolutions was typical on the same scale as above and mean was 3.4 with standard deviation 1.3. This indicates that the range was not completely typical, but more typical than not.

4.1.2 Statistical analysis of users' rating data

The scale responses were given numerical values when analysed using the following: Bad = 1, Poor = 2, Fair = 3, Good = 4 and Excellent = 5.

Characterization of the quality of the video clips is the Mean Opinion Scores (MOS) which is the mean over the ratings given by the users

$$MOS_{pvs} = \frac{1}{N}\sum\nolimits_{i=1}^{N} X_i \tag{13}$$

where μ_{ij} is the score of user i for PVS j. N is the number of users and M is the number of PVSs.

The statistical analysis that was performed was by first applying a repeated measures Analysis of Variance (ANOVA) and then performing a post-hoc analysis based on Tukey Honestly Significant Difference (HSD) [63, 64]. The repeated measures ANOVA will determine whether there are any significant differences between the MOS, and the post-hoc test will identify the specific pairs for which these significant differences exist.

Another characterisation of the data is the Standard deviation of Opinion Scores (SOS) as suggested by Hossfeld, et al. (2011) [65], which is now widely used. They demonstrated that there is a squared relationship between the SOS and MOS, dependent on only a single parameter a, which can be used to characterize an experimental study. For instance, they showed that typical image and video quality experiments has an a between 0.15, whereas a web surfing experiment would be about 0.25 and gaming above 0.3. The objective of this method is to determine the a value that provides the best fit to the data.

$$SOS^2(x) = a \cdot \left(-x^2 + 6x - 5\right) \tag{14}$$

4.1.3 1080p

The mean quality of the source video clips (SRCs) for 1080p taken over all degradations (HRCs) and users is shown in Fig. 5 left graph (see also Table 4). The MOS values are close to 3 and slightly above indicating that they have about the same number of high-quality degradations as low degradation with slightly more high quality than low quality. SRC2 (Butarque_follow_player_2) had the overall highest quality, indicating that this video was the least challenging to encode. SRC4 (SVT_CrowdRun) had the overall lowest quality, indicating that this video was the most challenging to encode, followed by SRC6

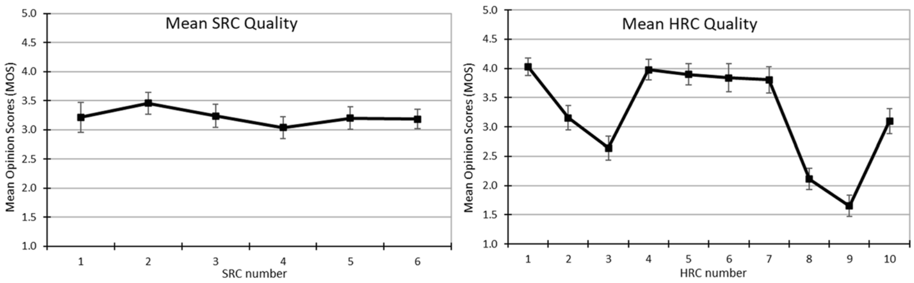

Fig. 5 (left) Mean quality of the source video for 1080p (y-axis) clips (SRCs, x-axis) for 1080p taken over all degradations (HRCs) and users, see also Table 4. (right) Mean quality (y-axis) of the degradations (HRCs, x-axis) taken over all source video clips (SRCs) and subjects, see also Table 6. The error bars represent 95% confidence intervals

(SVT_ParkRun). These two had also statistically significantly lower quality than SRC2, having p-values lower than 0.05 ($p = 0.00025$ and $p = 0.038$ respectively).

In Fig. 5 right graph (see also Table 6) the mean quality of the degradations (HRCs) taken over all source video clips (SRCs) and users are shown. It can be noted that there are some HRCs (4–7) that have almost the same quality (MOS) as the reference (HRC1). These are not statistically significantly different from the reference. The other HRCs were statistically significantly different from the HRC1 with $p = 0.00001 < 0.05$.

A breakdown of the HRCs into the different processing schemes and bitrates that were applied to the SRCs is shown in Fig. 6. The encoding performed by Motion JPEG is shown in solid black and the H.264 in dashed black curve. The MOS of the reference is marked as a red line without tying it to the bitrate to not make the x-axis too long. The quality drops fast with lower bitrates for MJPEG, whereas the quality for H.264 is indistinguishable from the reference down to about 20 Mbit/s.

In Fig. 7 the relationship between SOS^2 and MOS along with the fitted curve that estimates the parameter a for 1080p is shown. The parameter was estimated to be 0.19. Here

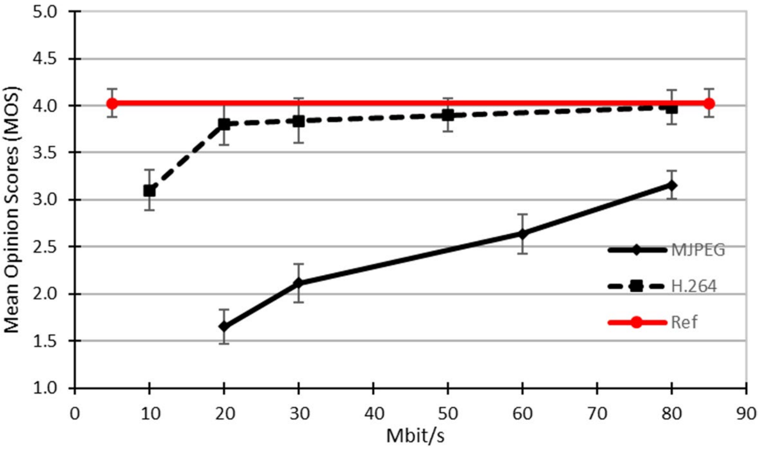

Fig. 6 The mean quality for 1080p (y-axis) of the degradations (HRCs) taken over all source video clips (SRCs) and users, divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as a red line without tying it to the bitrate to not make the x-axis too long. The error bars represent 95% confidence intervals

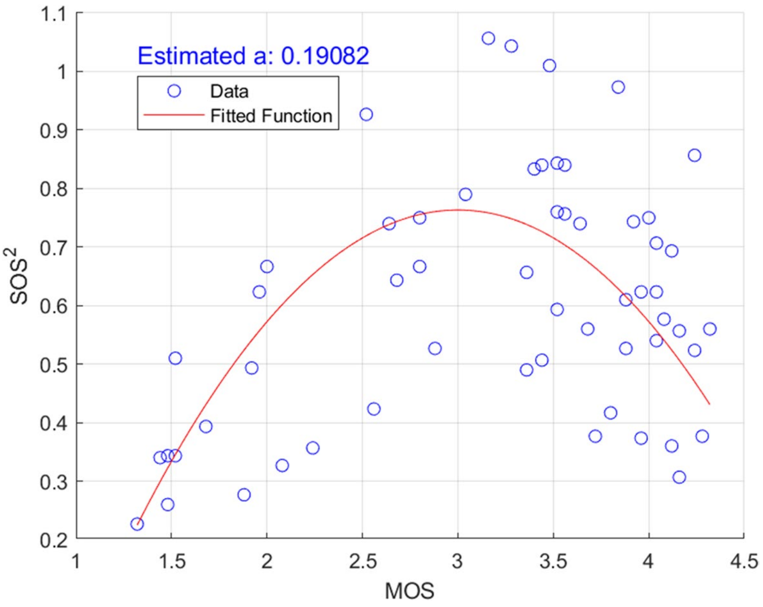

Fig. 7 The relationship between SOS^2 and MOS along with the fitted curve that estimates the parameter *a* for 1080p

we can also see that there more videos rated with higher quality than with lower quality. The variance is also highest about 3 to 3.5 showing that in this range the test persons were most uncertain about the quality.

4.1.4 1080i

In the same way as for 1080p above, the mean quality of the source video clips (SRCs) 1080i taken over all degradations (HRCs) and users is shown in Fig. 8 left graph (see also Table 5).

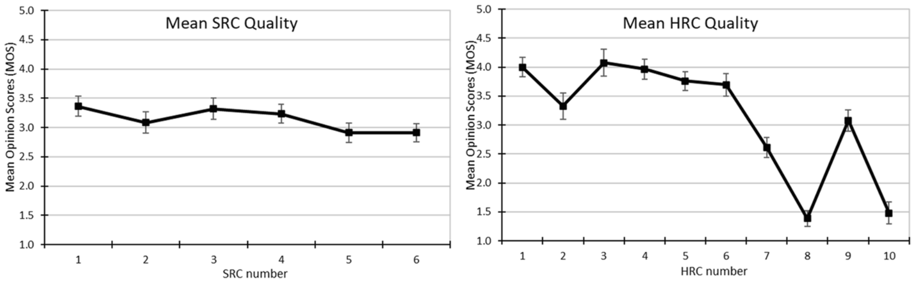

Fig. 8 (left) Mean quality for 1080i (y-axis) of the source video clips (SRCs, x-axis) for 1080i taken over all degradations (HRCs) and users, see also Table 5. (right) Mean quality (y-axis) of the degradations (HRCs, x-axis) taken over all source video clips (SRCs) and users, see also Table 7. The error bars represent 95% confidence intervals

The MOS values are also in this case slightly above 3. SRC5 (SVT_CrowdRun) and 6 (SVT_ParkJoy) have statistically significantly lower ($p<0.05$) MOS than SRC1 (Football_celebrating_1, $p=0.0001$). SRC3 (Football_greetings_1, $p=0.0001$) and SRC4 (Football_greetings_1, $p=0.0007$) and SRC2 (Football_goal_1) have lower MOS than SRC1 ($p=0.005$) and SRC3 ($p=0.02$).

In Fig. 8 right graph (see also Table 7) the mean quality of the degradations (HRCs) taken over all source video clips (SRCs) and users are shown. Also, for 1080i there are some HRCs (3–5) that have almost the same quality (MOS) as the reference (HRC1). These are not statistically significantly different from the reference. However, HRC6 is statistically significantly lower, even if this can be hard to see from the graph ($p=0.03$). The other HRCs, were also statistically significantly different from the HRC1 with $p=0.00001<0.05$.

The MOS plotted against the bitrates for the different HRCs are shown in Fig. 9. The encoding performed by MJPEG is shown in solid black and the H.264 in dashed black curve. The MOS of the reference is marked as a red line without tying it to the bitrate to not make the x-axis too long. The HRC10 was not an encoding error condition and was just a very simple deinterlacing applied directly to the uncompressed video and its MOS has been drawn in a similar way as the reference HRC1, as a yellow line across the graph. This error condition was not liked very much by the users and thus received very low ratings. The quality drops fast with lower bitrates for MJPEG, whereas the quality for H.264 is indistinguishable from the reference down to about 30 Mbit/s, but in contrast to 1080p 20 Mbit/s is statistically significantly lower for 1080i ($p=0.03<0.05$).

In Fig. 10 the relationship between SOS^2 and MOS along with the fitted curve that estimates the parameter a for 1080i is shown. The parameter was estimated to 0.18, almost the same as for 1080p. Here we can also see that there are more videos rated with higher quality than with lower quality and even more compared to 1080p. The variance was in this case highest in the range 3.5 to 4 further indicating a more skewed distribution than 1080p.

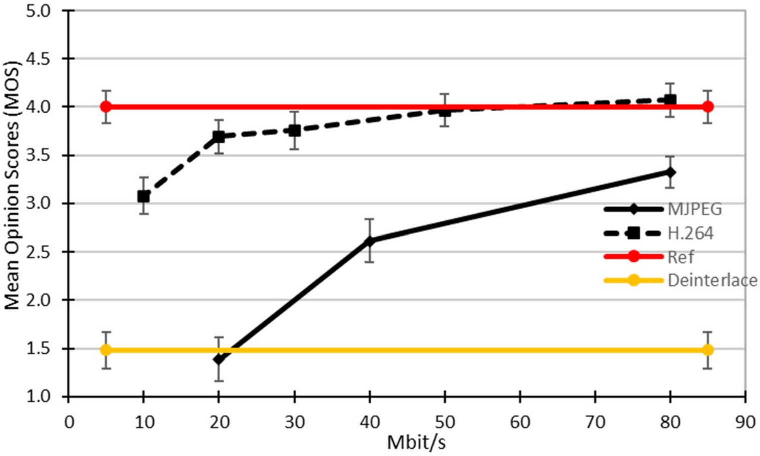

Fig. 9 The mean quality for 1080i (y-axis) of the degradations (HRCs) taken over all source video clips (SRCs) and users, divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as a red line without tying it to the bitrate to not make the x-axis too long. Similarly, the error conditions based on simple deinterlacing on an otherwise uncompressed video are shown as a yellow line. The error bars represent 95% confidence intervals

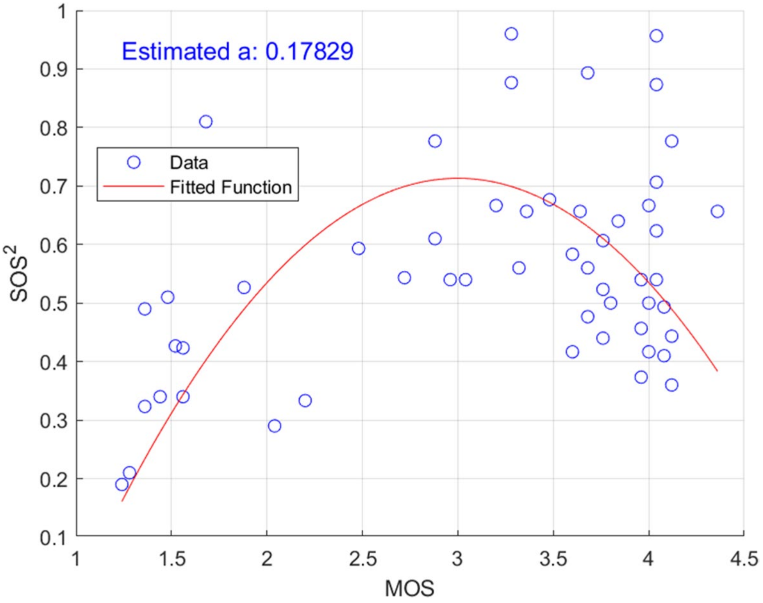

Fig. 10 The relationship between SOS^2 and MOS along with the fitted curve that estimates the parameter a for 1080i

4.1.5 540i

The mean quality of the source video clips (SRCs) for 540i taken over all degradations (HRCs) and users is shown in Fig. 11 left graph (see also Table 5). The average quality i.e., the MOS values are here more above 3 than what could be observed for 1080p (Fig. 5) and 1080i (Fig. 8). This is most likely due to the smaller format, which makes it harder to see the degradations. Overall highest quality had SRC3 (Football_greetings_1), which had a statistically significantly higher MOS than all the other SRCs ($p < 0.05$). The MOS of SRC6 (SVT_ParkJoy) was statistically significantly lower ($p < 0.05$) than SRC5 (SVT_CrowdRun, $p = 0.0003$) and SRC4 (Football_play_1, $p = 0.0002$) as well.

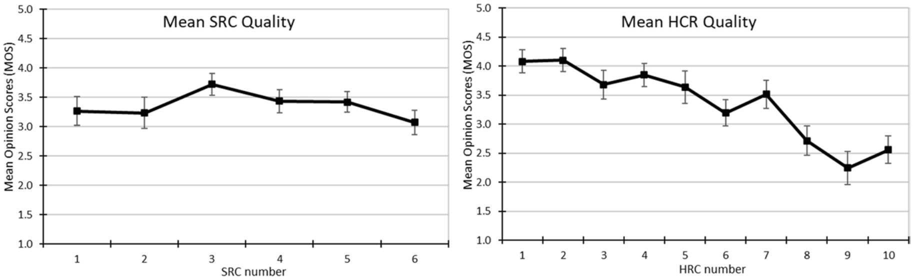

Fig. 11 (left) Mean quality (y-axis) of the source video clips (SRCs, x-axis) for 540i taken over all degradations (HRCs) and users, see also Table 5. (right) Mean quality (y-axis) of the degradations (HRCs, x-axis) taken over all source video clips (SRCs) and users, see also Table 8. The error bars represent 95% confidence intervals

In Fig. 11 right graph (see also Table 8) the mean quality of the degradations (HRCs) taken over all source video clips (SRCs) and users are shown. The different bitrates applied with H.264 come in groups of three, so HRC2-4 is 50 Mbit/s, HRC5-7 is 20 Mbit/s and HRC8-10 is 10 Mbit/s. The latter group was statistically significantly lower quality than the other, including the reference HRC1 ($p < 0.05$). For the other groups, it depends on which scaling method was used whether there was a difference or not. HRC6 bilinear 20 Mbit/s was statistically significantly worse than 50 Mbit/s ($p < 0.05$), but not Lanczos HRC5 and nearest neighbur HRC7. HRC3 bilinear 50 Mbit/s was statistically significantly worse than the reference HRC1 and Lanczos 50 Mbit/s HRC2. Lanczos were statistically significantly better than bilinear for all the bitrates.

A breakdown of the HRCs into the different processing schemes and bitrates applied to the SRCs is shown in Fig. 12. The different scaling methods are drawn as a separate curve, where Lanczos is drawn in solid black, bilinear in dashed black, and nearest neighbour in yellow. The MOS of the reference is marked as a red line without tying it to the bitrate to not make the x-axis too long. The graph shows a clear drop in quality for 10 Mbit/s, but not so severe for 20 Mbit/s (both levels were significant with $p < 0.001$), and hardly any quality decrease for 50 Mbit/s. The difference in quality between Lanczos and bilinear were statistically significant for 50 Mbit/s ($p = 0.003$), 20 Mbit/s ($p = 0.0002$) and 10 Mbit/s ($p = 0.0005$).

In Fig. 13 the relationship between SOS^2 and MOS along with the fitted curve that estimates the parameter a for 540i is shown. The parameter was estimated to 0.22, which is higher than both 1080p and 1080i. The distribution of qualities was more evenly spread than for the other formats. The variance is also highest approximately 3 showing that in this range the test persons were mostly uncertain about the quality.

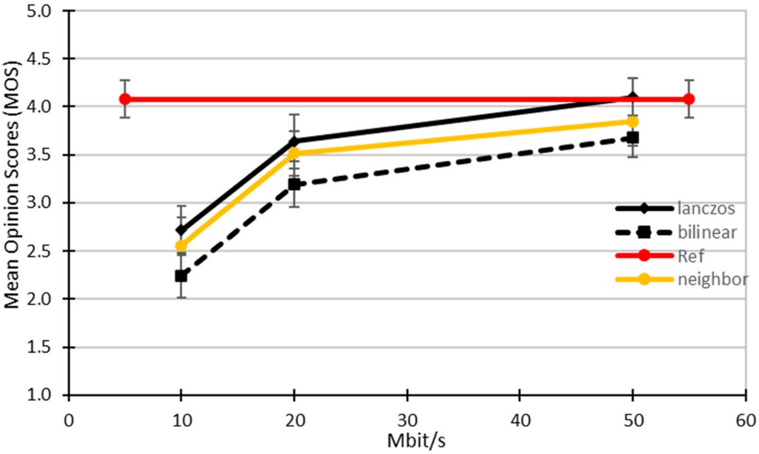

Fig. 12 The mean quality (y-axis) of the degradations (HRCs) taken over all source video clips (SRCs) and users, divided into the different scaling methods (Lanczos in a solid black curve, bilinear in dashed black curve, and nearest neighbour in yellow curve). The scaling was combined with H.264 encoding. The MOS of the reference is marked as a red line without tying it to the bitrate to not make the x-axis too long. The error bars represent 95% confidence intervals

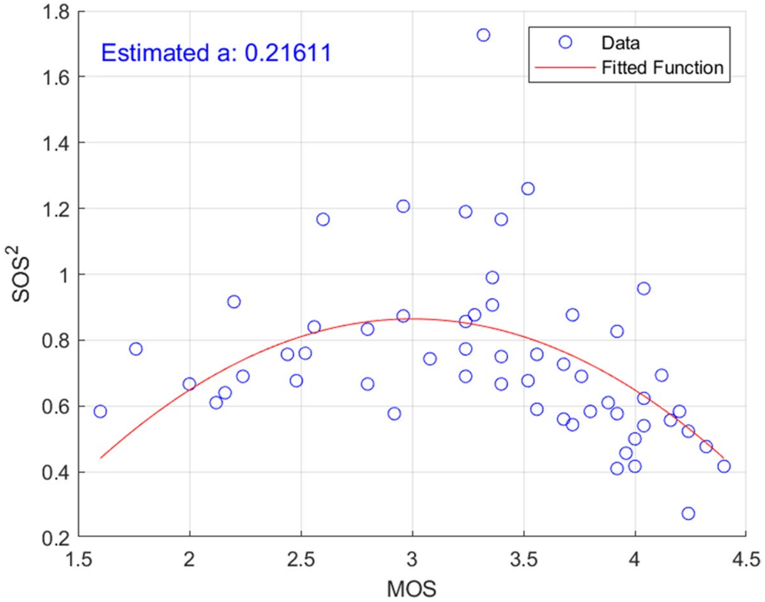

Fig. 13 The relationship between SOS^2 and MOS along with the fitted curve that estimates the parameter *a* for 540i

4.2 Objective video quality models evaluation

The objective models presented in Section 3.2, were evaluated for their performance on the video format 1080p and 1080i, using the method described there.

4.2.1 Video Multimethod Assessment Fusion

The performance of VMAF was very good using the default training. The results was a PCC of 0.89 and RMSE of 0.44 for 1080p (see Fig. 14 and Table 10) and a PCC of 0.89

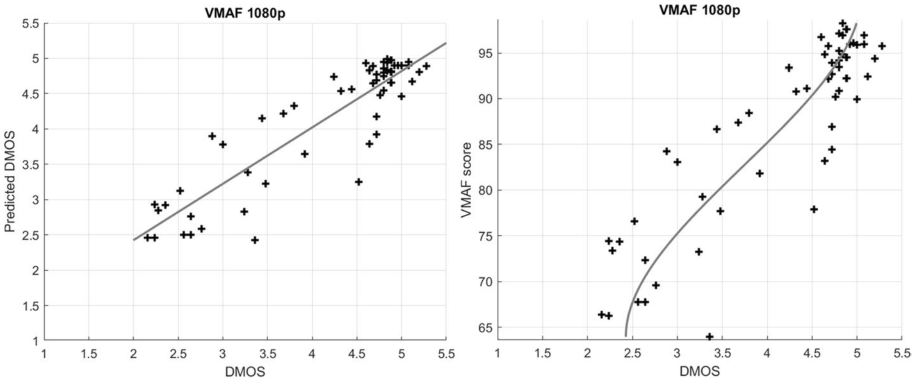

Fig. 14 Scatterplots of VMAF performance for 1080p. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

Table 10 Performances of the different models for 1080p. PCC (raw) is the Pearson Correlation Coefficient calculated on the raw scores. PCC (fit) is the linear correlation based on the 3rd order monotonic polynomial fit. SCC is the Spearman rank-order correlation or Spearman's rho. RMSE is the Root Mean Square Error

Model	PCC (raw)	PCC (fit)	SCC	RMSE
VMAF	0.882	0.893	0.828	0.441
VQM_VFD	-0.956	0.964	0.886	0.260
VQM_General	-0.806	0.866	0.806	0.492
SSIM	0.649	0.768	0.743	0.629
PSNR	0.727	0.748	0.712	0.652
VIF	0.735	0.753	0.719	0.646

and RMSE of 0.47 for 1080i (see Fig. 15 and Table 11), based on the 3rd order monotonic polynomial fit which (see Tables 14 and 15). Scatterplots are shown in Fig. 15. This implementation of VMAF needs external registration software, i.e., the encoding may shift the reference video and the degraded video a few frames which needs to be adjusted before running the model otherwise very low scores are obtained.

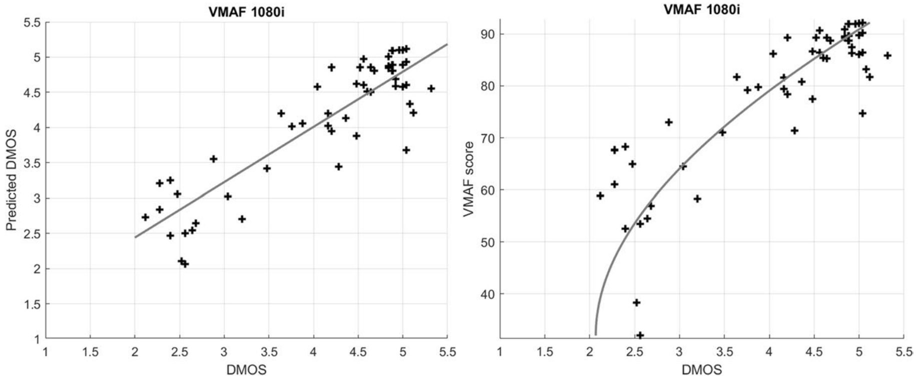

Fig. 15 Scatterplots of VMAF performance for 1080i. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

Table 11 Performances of the different models for 1080i. PCC (raw) is the Pearson Correlation Coefficient calculated on the raw scores. PCC (fit) is the linear correlation based on the 3rd order monotonic polynomial fit. SCC is the Spearman rank-order correlation or Spearman's rho. RMSE is the Root Mean Square Error

Model	PCC (raw)	PCC (fit)	SCC	RMSE
VMAF	0.852	0.886	0.763	0.468
VQM_VFD	-0.875	0.914	0.773	0.409
VQM_General	-0.848	0.865	0.700	0.506
SSIM	0.568	0.662	0.563	0.756
PSNR	0.743	0.806	0.672	0.597
VIF	0.746	0.798	0.649	0.607

4.2.2 Video Quality Metric

VQM General model (ITU-T Rec. J.144) was evaluated at RISE and the performance was good, with a PCC of -0.87 and RMSE of 0.49 for 1080p (see Fig. 16 and Table 10) and a PCC of -0.86 for 1080i and RMSE of 0.51 (see Fig. 17 and Table 11), based on the 3rd order monotonic polynomial fit (see Tables 14 and 15). Negative values indicate that the slope of the line is negative, meaning in this case if the quality is low the model gives higher values than if the quality is high.

VQM has also an improved model, called VQM_VFD, where VFD stands for Variable Frame Delay. RISE evaluated this version too, which had a really good performance and was the top-performing model in this evaluation, with a PCC of -0.96 and RMSE of 0.26 for 1080p (see Fig. 18 and Table 10) and a PCC of -0.91and RMSE of 0.41for 1080i, (see Fig. 19 and Table 11), based on the 3rd order monotonic polynomial fit (see Tables 14 and 15).

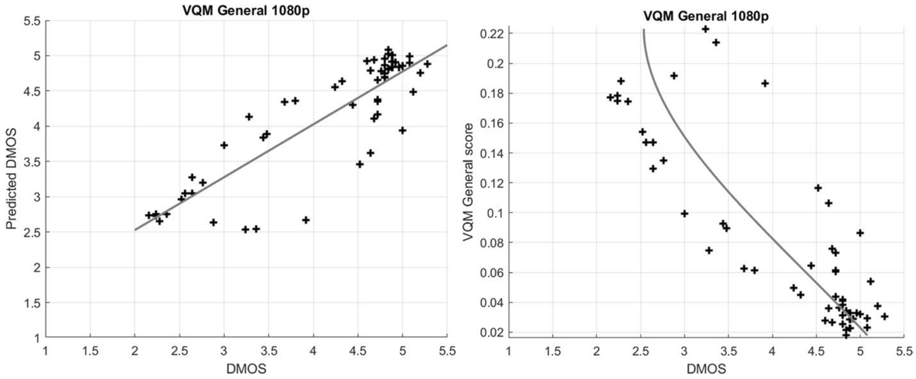

Fig. 16 Scatterplots of VQM General performance for 1080p. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

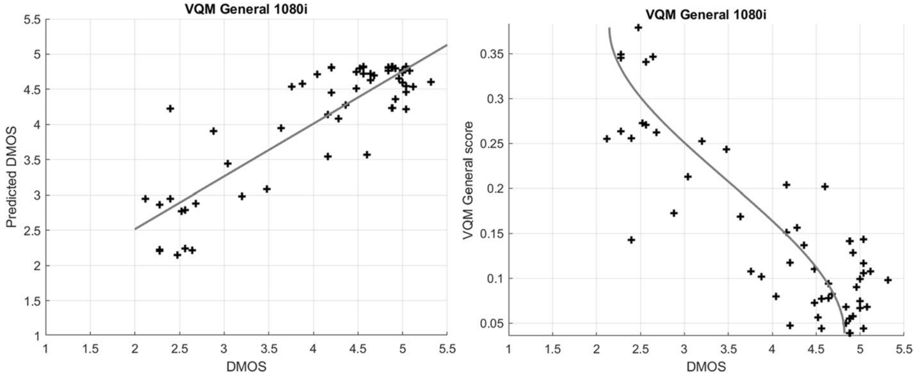

Fig. 17 Scatterplots of VQM General performance for 1080i. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

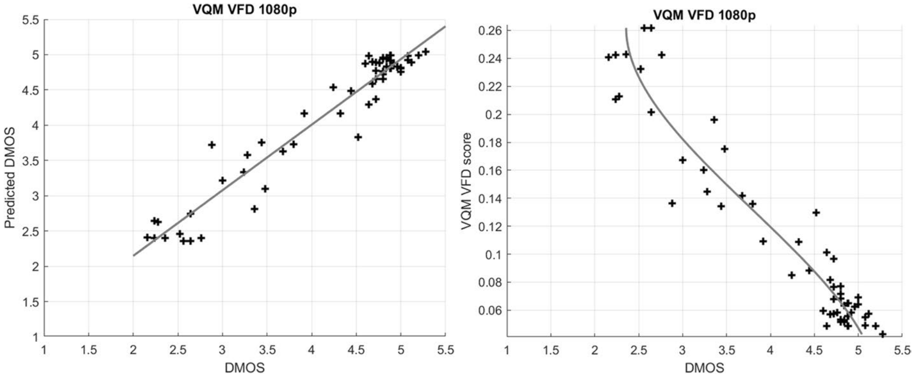

Fig. 18 Scatterplots of VQM_VFD performance for 1080p. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

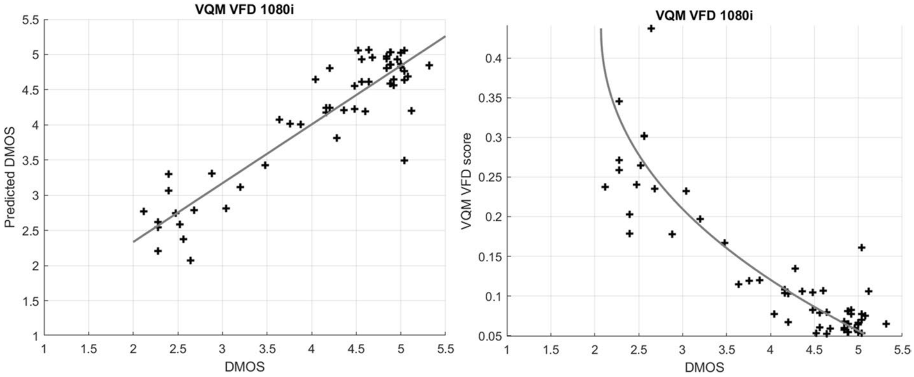

Fig. 19 Scatterplots of VQM_VFD performance for 1080i. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

VQM quality metrics (both the General model and VFD) do not need external registration software and the registration algorithm can be used on its own.

4.2.3 Peak Signal to Noise Ratio

PSNR was evaluated with a PCC of -0.75 and RMSE of 0.65 for 1080p (see Fig. 20 and Table 10) and a PCC of -0.81 and RMSE of 0.60 for 1080i, (see Fig. 21 and Table 11), based on the 3rd order monotonic polynomial fit (see Tables 14 and 15). In the raw format (Fig. 21 right graph), there is a group of outliers belonging to HRC10 (simple deinterlacing), that decrease the performance. The 3rd-order monotonic polynomial fit partly compensated for it and included them into the main group.

The implementation of PSNR needs external registration software but could also be calculated with the VQM software and then the registration could be done using that [38].

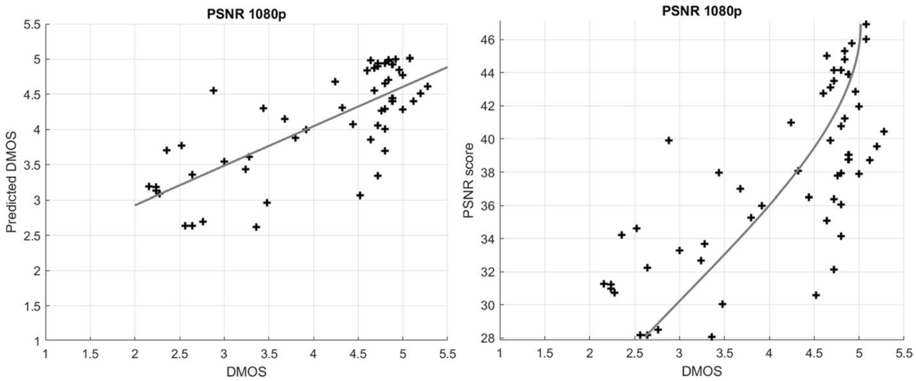

Fig. 20 Scatterplots of PSNR performance for 1080p. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

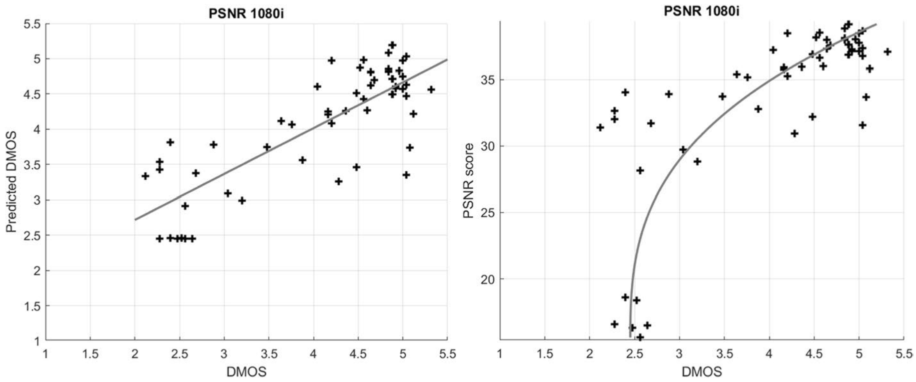

Fig. 21 Scatterplots of PSNR performance for 1080i. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

4.2.4 Structural Similarity Index

SSIM performance was evaluated to a PCC of -0.77 and RMSE of 0.63 for 1080p (see Fig. 22 and Table 10) and a PCC of -0.66 and RMSE of 0.76 for 1080i, (see Fig. 23 and Table 11), based on the 3rd order monotonic polynomial fit (see Tables 14 and 15). SSIM has an especially low performance for 1080i. This is due to the HRC10 which SSIM had large problems with, and the 3rd order monotonic polynomial fit was not able to adequately compensate for it.

This implementation of SSIM needs external registration software.

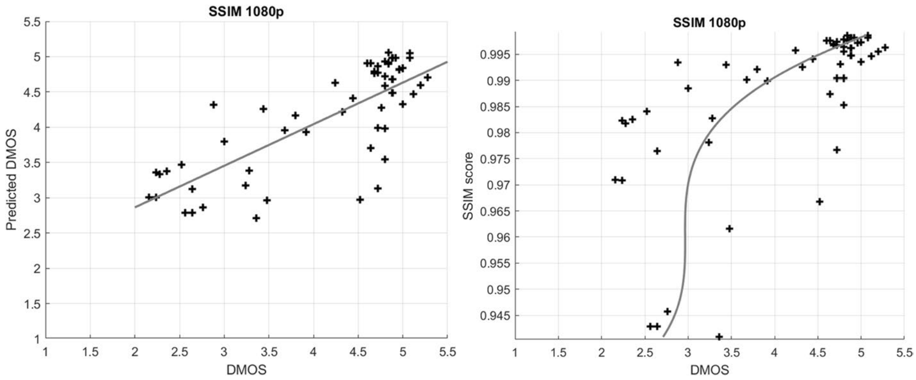

Fig. 22 Scatterplots of SSIM performance for 1080p. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

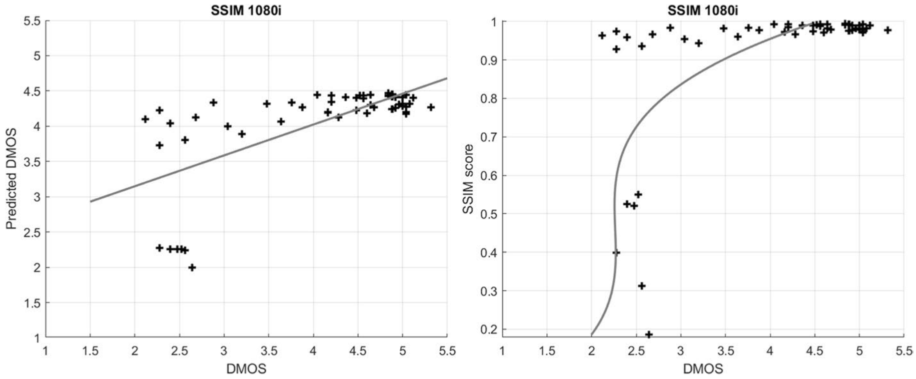

Fig. 23 Scatterplots of SSIM performance for 1080i. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

4.2.5 Visual Information Fidelity

VIF performance was evaluated to a PCC of -0.75 and RMSE of 0.64 for 1080p (see Fig. 24 and Table 10) and a PCC of -0.80 and RMSE of 0.61 for 1080i, (see Fig. 25 and Table 11), based on the 3rd order monotonic polynomial fit (see Tables 14 and 15). This implementation of VIF needs external registration software. VIF has similar problems with HRC10 as PSNR and SSIM.

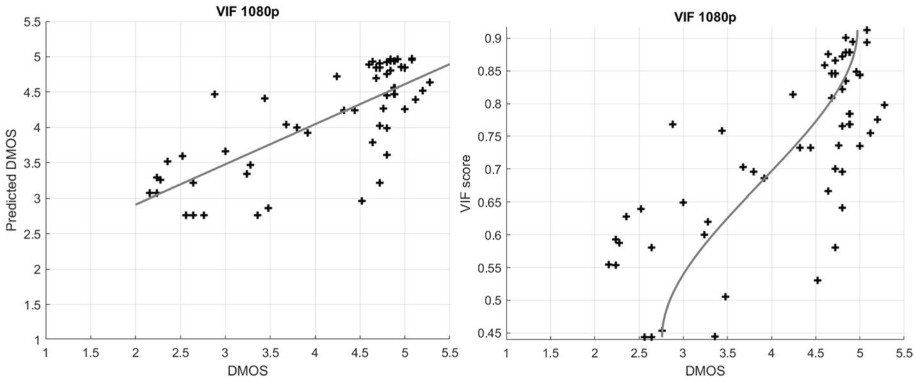

Fig. 24 Scatterplots of VIF performance for 1080p. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

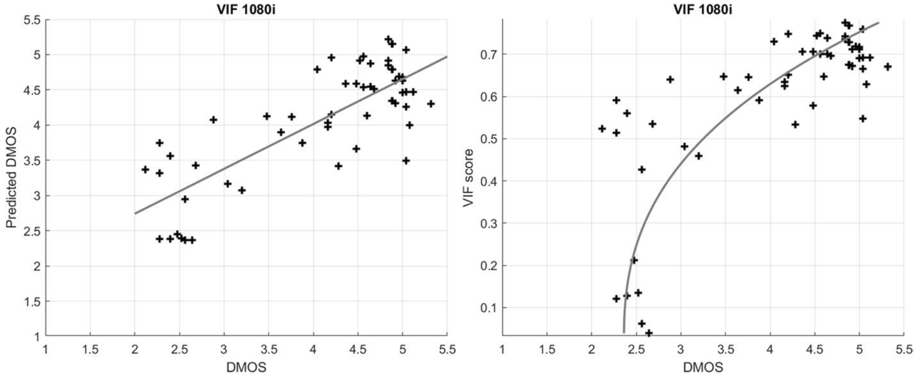

Fig. 25 Scatterplots of VIF performance for 1080i. The left graph shows the results after the 3rd order monotonic polynomial fit with a linear correlation line overlaid. The right graph shows the results before the 3rd order monotonic polynomial fit with the fitted function overlaid

4.2.6 Statistical significance testing

The results of statistical hypothesis testing, based on a pairwise comparison of the RMSE values, see Section 3.2, for a description of the procedure that is based on ITU-T Rec. P.1401 [37]. The p-values are shown in Tables 12 and 13. VQM_VFD is significantly better than all other models for 1080p and better than PSNR, SSIM, and VIF for 1080i. VMAF is significantly better than PSNR and VIF for 1080p. SSIM has a very low performance for 1080i and is significantly worse than all other models (Tables 14 and 15).

Table 12 P-values of statistical test on the difference in RMSE based on ITU-T Rec. P.1401 [37] for 1080p. Significant values are marked with *, based on an alpha of 0.05 and the method of Holm [54] for multiple comparisons of 15 comparisons

Model	VMAF	VQM_VFD	VQM General	SSIM	PSNR
VMAF					
VQM_VFD	0.00014 *				
VQM_General	0.22	< 0.0001 *			
SSIM	0.0067 *	< 0.0001 *	0.042 *		
PSNR	0.0034 *	< 0.0001 *	0.024 *	0.40	
VIF	0.0040 *	< 0.0001 *	0.028 *	0.43	0.48

Table 13 P-values of statistical test on the difference in RMSE based on ITU-T Rec. P.1401 [37] for 1080i. Significant values are marked with *, based on an alpha of 0.05 and the method of Holm [54] for multiple comparisons of 15 comparisons

Model	VMAF	VQM_VFD	VQM General	SSIM	PSNR
VMAF					
VQM_VFD	0.17				
VQM_General	0.29	0.066			
SSIM	0.00046 *	< 0.0001 *	0.0027 *		
PSNR	0.044	0.0042 *	0.12	0.049 *	
VIF	0.0343	0.0030 *	0.10	0.062	0.45

Table 14 Parameter values of the 3rd order monotonic polynomial fit ($p(x) = p_1x^3 + p_2x^2 + p_3x + p_4$) for each model for 1080p

Model	p_1	p_2	p_3	p_4
VMAF	-0.0001	0.0248	-1.944	51.5084
VQM_VFD	328.0566	-129.8252	0.5337	5.2293
VQM_General	213.1055	-38.4431	-14.6526	5.3536
SSIM	0.3668×10^5	-1.0566×10^5	1.0144×10^5	-0.3246×10^5
PSNR	-0.0002	0.0206	-0.4529	4.0492
VIF	-42.8238	87.0987	-51.9817	12.419

Table 15 Parameter values of the 3rd order monotonic polynomial fit ($p(x) = p_1x^3 + p_2x^2 + p_3x + p_4$) for each model for 1080i

Model	p_1	p_2	p_3	p_4
VMAF	0	0.0012	-0.0693	3.1374
VQM_VFD	-14.5705	33.7064	-21.1232	6.0813
VQM_General	135.1155	-84.7016	5.9277	4.7108
SSIM	15.0654	-20.8021	9.406	0.8727
PSNR	0.0002	-0.0078	0.1122	1.9218
VIF	3.9557	1.8947	-0.1668	2.3669

5 Discussions

This study was motivated by the need to find a video quality algorithm that is good in predicting quality for videos that are presented for assistant referees. Typically, the bandwidth used and thus the quality is higher than what is normally provided for end TV consumers. Therefore, video experts were invited to the experiment for collecting relevant data to differentiate the different video quality methods. We present the result here from six different well-known methods. A few commercially available methods were also considered, but since they did not perform better than the best performing methods in this investigation, we decided not to present these results.

Most evaluations for video quality uses naïve or non-expert viewers in their investigations. In ITU Recommendation (ITU-T Rec P.910) [9] it is recommended that the viewer should not be involved in video quality evaluations as part of their work. In comparisons between expert and non-expert viewers it has been shown that the general trends are similar, but the experts are more sensitive and give lower votes than the non-expert viewers, as reported by Speranza, et al. (2010) [66]. They also pointed out that this could pose a challenge for objective models, since they in general are based on non-expert viewer data.

The video experts used in this investigation considered themselves as real experts with mean 4.5 and a small standard deviation, i.e., 0.6. All of them with excellent vision which means that they had both the knowledge and the vision appropriate for the task. They also rated the distortions to be typical and to a larger extent in a quintessential range.

The task given was to rate the video quality in a single stimulus procedure [9] with a hidden reference, meaning that the test persons do not know when the reference is shown and it is rated in the same way as the other videos in the test. There are more sensitive methods when it comes to seeing and rating distortions as e.g., DSIS [8], but the experiment was designed to mimic the viewing situation by the referees as they will not have any references to compare with. Furthermore, the single-stimulus method allows for getting ratings on a larger number of different videos in the same time period of time as DSIS [10].

The results show that even for video experts it is not possible to observe any quality degradation for the highest bitrates of H.264 for 1080p, as shown in Fig. 6. According to the results, the quality degradations are apparent first below 20 Mbit/s. The Motion JPEG quality on the other hand drops very quickly with decreasing bitrate and even highest bitrate clearly lower quality than the reference. Similarly, for 1080i but here 20 Mbit/s was found to be significantly different as well as the 10 Mbit/s for H.264, see Fig. 9. Motion JPEG has the same behaviour as for 1080p. A simple deinterlacing was included as a degradation and this was found to have really low quality.

The 540i was a bit different as it tested different scaling algorithms along with different bitrates Fig. 12. The different scaling algorithms follow each other closely. Lanczos got consistently higher ratings per bitrates than bilinear and nearest neighbour, but it was only statistically significant for 20 Mbit/s and 10 Mbit/s and only between Lanczos and bilinear.

The estimation of the SOS parameter *a* for the different formats were 0.18 (1080i), 0.19 (1080p) and 0.22 (540i). The values were similar especially for 1080p and 1080i. They were also in the range expected based on the findings by Hossfeld, et al. (2011) [65]. The *a* value for 540i is a bit higher than expected, but the degradation were here of a different type. Also, the smaller picture size may have affected the results.

The evaluation of the objective video quality models followed the standardized procedure outlined in ITU-T Rec. P.1401 [37] and what the Video Quality Experts Group

(VQEG) has been following in their evaluations, see e.g. [20]. The RMSE has been the primary metric for the evaluation and the metric for which statistical significance has been calculated. There is often a non-linear relationship between the scores from the test persons and the objective scores [37], which was compensated partly with a monotonic non-linear fit between the MOS and the objective scores. Here we used a 3rd order polynomial function, but others can also be used e.g., a logistic function. Pearson and Spearman correlation were also calculated to give a richer description of the results. The statistical significance testing was based on the RMSE.

From a performance perspective, we can see that the VQM_VFD got the lowest RMSE and highest correlation score among all the models for both 1080p and 1080i, see Tables 10 and 11. It was statistically significantly better than all the other models for 1080p, but not significantly better than VMAF and VQM_General for 1080i, as shown in Tables 12 and 13. In Netflix's original evaluation VMAF and VQM_VFD have the same performance. They targeted H.264 also but with lower bitrates, with a mixtures of resolutions, with DSIS and naïve observers [3]. The second group is VMAF and VQM_General which have statistically the same performance i.e., they are not statistically significantly different from each, although VMAF shows slightly lower RMSE and higher correlation than VQM_General. The third group is SSIM, PSNR and VIF that show significantly worse performance.

Although VQM_VFD had the best performance, for practical reasons, the wide spread usage, implementation and general familiarity, along with the good performance, the method finally adopted was VMAF [1].

6 Conclusions

The research questions that this article was addressing were:

1) What levels of video quality are considered good or better for expert video professionals especially targeting sports content?
2) Which objective quality models that have been developed by using video quality data from naïve test persons and most likely a wider and lower quality range, can well predict the quality in a high-quality range as defined by expert professionals from the broadcast industry?

6.1 A user study of video quality

A user study was conducted, to answer research question 1, with the participation of 25 Swedish video experts who assessed perceived video quality on a five-point scale, encompassing categories ranging from Excellent to Bad. The video content selected for evaluation focused on football-related scenes with dynamic motion and moving subjects. Three video formats were included: 1080p, 1080i, and 540i. The degradation processes mostly involved MJPEG and H.264 encoding, with bitrates ranging from 80 Mbit/s to 10 Mbit/s. Additionally, 1080i involved a simple deinterlacing error condition, and 540i incorporated three different scaling schemes in addition to the H.264 encoding.

The results revealed that MJPEG experienced a rapid decline in quality, with a significant decrease observed at 80 Mbit/s compared to the uncompressed reference. For H.264, no significant quality difference was observed until the bitrate dropped to 10 Mbit/s for

1080p, and 20 Mbit/s for 1080i. In the case of 540i, some scaling methods showed a significant quality decrease at 20 Mbit/s. Regarding deinterlacing for 1080i, careful consideration is crucial, as a low-quality deinterlacing scheme led to poor quality scores. Among the scaling schemes, Lanczos performed the best, while bilinear yielded the worst results.

The SOS analysis confirmed that the study's video quality assessments fell within the expected range for such experiments. Both 1080i and 1080p exhibited similar results, with MOS values slightly skewed towards higher quality. Conversely, for 540i, the distribution was more symmetric.

6.2 Objective video quality model performance

To answer research question 2, six different video quality models were independently evaluated for their performance with 1080p and 1080i videos. VQM_VFD demonstrated the best performance for both formats, followed by VMAF and VQM General models. On the other hand, SSIM, PSNR, and VIF exhibited similar, lower performance compared to the evaluated video models. Notably, SSIM's performance was particularly low for 1080i, mainly due to the HRC with a low-quality deinterlacing method. Scatter plots indicated that PSNR and VIF also faced similar issues in their performance assessments.

6.3 Future work

Future work will include evaluation of new broadcast formats such as 4 K and 8 K as well as high dynamic range video. Different frame rates will also be valuable to investigate. New encoders are also developed and should be included in future studies. The impact of zooming was not covered by the current study, but it is important to study as zoom is used frequently for a closer view of situations. Is it possible to find a perception limit when the remaining quality after zooming is not good enough to be used for decisions? New objective methods are also developed especially using machine learning techniques. This will potentially also enable the usage of NR methods for live quality monitoring, but this is still not mature enough to be used in production [19]. This area is developing fast and new ways of processing the videos may prove to be very valuable [67].

6.4 Limitations

The uniqueness of this study lies in its focus on video quality opinions from video experts rather than end-users or consumers of broadcast TV. As such, its limitation is also a distinctive feature. The content primarily centres around sports, particularly Football, but additional content was included to broaden the relevance of the findings. The study encompassed video formats of 1080p, 1080i, and 540i, thus offering insights into these specific formats. However, the direct applicability of the results to other video formats may be limited.

It is essential to note that the study's scope is restricted to assessing the performance of the objective models considered within this investigation, which is contingent on the video content and codec employed in this study.

Acknowledgements The authors appreciate the work of Andreas Langell, Mobile Links Europe AB Sweden, who helped us with the possibility to acquire footage at Tele 2 arena in Stockholm and provided input on where to find video expert candidates for the user study. We also appreciate the work of Benny Norling, Boresight AB Sweden, who helped us video with related expertise, a study on monitors and video formats, and input on test person candidates for the subjective experiment. The authors acknowledge Pär Johansson (RISE) for proofreading, valuable discussions, and general support in the project as well as Jade Lumley for the thorough native language review.

This work was funded by Fédération Internationale de Football Association (FIFA) and Sweden´s Innovation Agency (VINNOVA, dnr. 2021-02107) through the Celtic-Next project IMMINENCE (C2020/2-2), which is hereby gratefully acknowledged.

Authors' contributions Kjell Brunnström: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization and Writing – original draft, Writing—review and editing.

Anders Djupsjöbacka: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization and Writing – original draft, Writing—review and editing.

Börje Andrén: Data curation, Validation, Investigation, Visualisation and Writing—review and editing.

Oskars Ozolins: Data curation, Validation, Investigation, Visualisation and Writing—review and editing.

Johsan Billingham: Data curation, Validation, Investigation, Visualisation and Writing—review and editing.

Katharina Wistel: Data curation, Validation, Investigation, Visualisation and Writing—review and editing.

Nicolas Evans: Conceptualization, Funding acquisition, Project administration, Writing—review and editing.

Funding Open access funding provided by RISE Research Institutes of Sweden. This project was funded by Fédération Internationale de Football Association (FIFA), Sweden's innovation agency VINNOVA (grant nr 2021–02107) and internal funding by RISE.

Data availability Not applicable.

Code availability The VQEGplayer is available at http://vqegjeg.intec.ugent.be/wiki/index.php/VQEGplayer-main.

Declarations

Ethics approval Not applicable.

Consent to participate All participating users signed a consent form before the participation in the study.

Consent for publication All authors have given their consent for publication.

Conflicts of interest/Competing interests RISE has business contracts with FIFA and is the organisation to perform certification measurement for Quality of VAR on behalf of FIFA.

References

1. Brunnström K, Djupsjöbacka A, Ozolins O, Billingham J, Wistel K, Evans N (2023) Quality measurement methods for video assisting refereeing systems. Sports Eng 26(1):17. https://doi.org/10.1007/s12283-023-00408-6
2. Wolf S, Pinson M (2011) Video Quality Model for Variable Frame Delay (VQM_VFD) (NTIA Technical Memorandum TM-11-482). National Telecommunications and Information Administration (NTIA), Boulder

3. Li Z, Aaron A, Katsavounidis I, Moorthy AK, Manohara M (2016). Toward a practical perceptual video quality metric. Netflix Technology Blog. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, Access Date: Oct 23, 2018
4. ITU-T (2016) Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference (ITU-T Rec. J.341). International Telecommunication Union, Telecommunication standardization sector
5. ITU-T (2004) Objective perceptual video quality measurement techniques for digital cable television in the presence of full reference (ITU-T Rec. J.144). International Telecommunication Union, Telecommunication standardization sector
6. ITU-T (2017) Vocabulary for performance, quality of service and quality of experience (ITU-T Rec. P.10/G.100). International Telecommunication Union (ITU), Place des Nations, CH-1211 Geneva 20
7. Le Callet P, Möller S, Perkis A (2012) Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003) (Version 1.2 (http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf)), Lausanne, Switzerland
8. ITU-R (2023) Methodology for the subjective assessment of the quality of television pictures (ITU-R Rec. BT.500-15). International Telecommunication Union (ITU)
9. ITU-T (2008) Subjective video quality assessment methods for multimedia applications (ITU-T Rec. P.910). International Telecommunication Union, Telecommunication standardization sector
10. ITU-T (2021) Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment (ITU-T Rec. P.913). International Telecommunication Union, Telecommunication standardization sector
11. Lee C, Choi H, Lee E, Lee S, Choe J (2006) Comparison of various subjective video quality assessment methods. In: Image Quality and System Performance III, San Jose, CA, United States, SPIE. https://doi.org/10.1117/12.651056
12. Huynh-Thu Q, Ghanbari M (2005) A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video. In: IASTED Int. Conf. on Signal Image Process. IASTED, pp 70–76
13. Tominaga T, Hayashi T, Okamoto J, Takahashi A (2010) Performance comparisons of subjective quality assessment methods for mobile video. In: Second International Workshop on Quality of Multimedia Experience (QoMEX 2010). Trondheim, Norway, pp 82–87. https://doi.org/10.1109/QOMEX.2010.5517948
14. Berger K, Koudota Y, Barkowsky M, Callet PL (2015). Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains. In: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), p 1-6. https://doi.org/10.1109/QoMEX.2015.7148114
15. Pitrey Y, Barkowsky M, Le Callet P, Pépion R (2010) Subjective quality evaluation of H.264 high-definition video coding versus spatial up-scaling and interlacing. In: Euro ITV. Tampere, Finland
16. Barkowsky M, Staelens N, Janowski L, Koudota Y, Leszczuk M, Urvoy M, Hummelbrunner P, Sedano I, Brunnström K (2012) Subjective experiment dataset for joint development of hybrid video quality measurement algorithms. In Proc of the Third Workshop on Quality of Experience for Multimedia Content Sharing (QoEMCS), EuroITV 2012, Berlin, Germany
17. Choe J-H, Jeong T-U, Choi H, Lee E-J, Lee S-W, Lee C-H (2007) Subjective video quality assessment methods for multimedia applications. J Broadcast Eng 12. https://doi.org/10.5909/JBE.2007.12.2.177
18. EBU (2011) Signal Quality in HDTV Production and Broadcast Services (Recommendation R132). European Broadcasting Union (EBU), Geneva
19. Pinson MH (2022) Why no reference metrics for image and video quality lack accuracy and reproducibility. IEEE Trans Broadcast: 1–21. https://doi.org/10.1109/TBC.2022.3191059
20. VQEG (2010) Report on the validation of video quality models for high definition video content. Video Quality Experts Group (VQEG). https://www.vqeg.org/media/4212/vqeg_hdtv_final_report_version_2.0.zip. Accessed 6 Dec 2023
21. Brunnström K, Hands D, Speranza F, Webster A (2009) VQEG validation and ITU standardisation of objective perceptual video quality metrics. IEEE Signal Process Mag 26(3):96–101. https://doi.org/10.1109/MSP.2009.932162
22. VQEG (2009) Validation of reduced-reference and no-reference objective models for standard definition television, phase I. Video Quality Experts Group (VQEG). https://www.vqeg.org/media/66832/rrnr-tv_final_report_v1_9.pdf. Accessed 6 Dec 2023
23. VQEG (2008) Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, phase I (VQEG Final Report of MM Phase I Validation Test). Video Quality Experts Group (VQEG). https://www.vqeg.org/media/66834/vqeg_mm_report_final_v26.pdf. Accessed 6 Dec 2023
24. VQEG (2003) Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (VQEG Final Report of FR-TV Phase II Validation Test). Video

Quality Experts Group (VQEG). https://www.vqeg.org/media/4176/vqegii_final_report.doc. Accessed 6 Dec 2023

25. VQEG (2000) Final report from the video quality experts group on the validation of objective models of video quality assessment. Video Quality Experts Group (VQEG). https://www.vqeg.org/media/8212/frtv_phase1_final_report.doc. Accessed 6 Dec 2023
26. Winkler S, Mohandas P (2008) The evolution of video quality measurement: from PSNR to hybrid metrics. IEEE Trans Broadcast 54(3):660–668. https://doi.org/10.1109/TBC.2008.2000733
27. Shahid M, Rossholm A, Lövström B, Zepernick H-J (2014) No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP J Image Video Process 2014(1):40. https://doi.org/10.1186/1687-5281-2014-40
28. Liu T-J, Lin Y-C, Lin W, Kuo CCJ (2013) Visual quality assessment: recent developments, coding applications and future trends. APSIPA Trans Signal Inf Process 2:e4. https://doi.org/10.1017/ATSIP.2013.5
29. Raake A, Borer S, Satti SM, Gustafsson J, Rao RRR, Medagli S, List P, Göring S, Lindero D, Robitza W, Heikkilä G, Broom S, Schmidmer C, Feiten B, Wüstenhagen U, Wittmann T, Obermann M, Bitto R (2020) Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204. IEEE Access 8:193020–193049. https://doi.org/10.1109/ACCESS.2020.3032080
30. Shahid M (2014) Methods for objective and subjective video quality assessment and for speech enhancement. (15), Blekinge Institute of Technology Doctoral Dissertation Series, Blekinge Institute of Technology, Karlskrona, Doctoral thesis, comprehensive summary. http://bth.diva-portal.org/smash/get/diva2:833983/FULLTEXT01.pdf. Accessed 6 Dec 2023
31. Pinson M, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322. https://doi.org/10.1109/TBC.2004.834028
32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861
33. Lee C, Cho S, Choe J, Jeong T, Ahn W, Lee E (2006) Objective video quality assessment. Opt Eng 45(1):017004. https://doi.org/10.1117/1.2160515
34. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444. https://doi.org/10.1109/TIP.2005.859378
35. Cheon M, Lee J (2018) Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience. IEEE Trans Circuits Syst Video Technol 28(7):1467–1480. https://doi.org/10.1109/TCSVT.2017.2683504
36. Rao RRR, Göring S, Robitza W, Feiten B, Raake A (2019) AVT-VQDB-UHD-1: a large scale video quality database for UHD-1. In: 2019 IEEE International Symposium on Multimedia (ISM), p 17-177. https://doi.org/10.1109/ISM46123.2019.00012
37. ITU-T (2020) Statistical analysis, evaluation and reporting guidelines of quality measurements (ITU-T P.1401). International Telecommunication Union, Telecommunication standardization sector, Geneva
38. ITU-T (2010) Reference algorithm for computing peak signa to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset (ITU-T Rec. J.340). International Telecommunication Union (ITU), Telecommunication Standardization Sector
39. Wulf S, Zölzer U (2012) Full-reference video quality assessment on high-definition video content. In: 2012 6th International Conference on Signal Processing and Communication Systems, p 1–10. https://doi.org/10.1109/ICSPCS.2012.6507948
40. Simone FD, Naccari M, Tagliasacchi M, Dufaux F, Tubaro S, Ebrahimi T (2009) Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel. In: 2009 International Workshop on Quality of Multimedia Experience, p 204-209. https://doi.org/10.1109/QOMEX.2009.5246952
41. Simone FD, Tagliasacchi M, Naccari M, Tubaro S, Ebrahimi T (2010) A H.264/AVC video database for the evaluation of quality metrics. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, p 2430–2433. https://doi.org/10.1109/ICASSP.2010.5496296
42. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. IEEE Trans Image Process 19(6):1427–1441. https://doi.org/10.1109/TIP.2010.2042111
43. Sedano I, Brunnström K, Kihl M, Aurelius A (2014) Full-reference video quality metric assisted development of no-reference video quality metrics for real time network monitoring. EURASIP J Image Video Process 2014(4). https://doi.org/10.1186/1687-5281-2014-4

44. ITU-T (2021) H.264 : Advanced video coding for generic audiovisual services (ITU-T Rec. H.264). International Telecommunication Union
45. ITU-T (2016) High efficiency video coding (ITU-T Rec. H.265). International Telecommunication Union, Telecommunication standardization sector
46. Ling S, Baveye Y, Nandakumar D, Sethuraman S, Callet PL (2020) Towards better quality assessment of high-quality videos. In: Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications. Association for Computing Machinery, Seattle, pp 3–9. https://doi.org/10.1145/3423328.3423496
47. ITU-R (2015) Parameter values for the HDTV standards for production and international programme exchange (Rec. ITU-R BT.709-6). International Telecommunication Union, Radiocommunication Sector
48. Haglund L (2006) The SVT high definition multi format test set. Sveriges Television AB (SVT), Stockholm
49. Apple Inc. (2008) Shake: Advanced digital composition. https://web.archive.org/web/20080122073447/http://www.apple.com/shake, Access Date: 11 Oct 2023
50. Wikipedia (2023) Shake (software). https://en.wikipedia.org/wiki/Shake_(software), Access Date: 11 Oct 2023
51. Brunnström K, Cousseau R, Jonsson J, Koudota Y, Bagazov V, Barkowsky M (2014) VQEGPlayer: open source software for subjective video quality experiments in windows. Video Quality Experts Group (VQEG). www.vqeg.org, Available from: http://vqegjeg.intec.ugent.be/wiki/index.php/VQEGplayer-main. Accessed 19 Dec 2023
52. FFMPEG (2023) FFMPEG: complete, cross-platform solution to record, convert and stream audio and video. www.ffmpeg.org, Access Date: 7 Aug 2023
53. Matlab (2006) corr - Linear or rank correlation. https://se.mathworks.com/help/stats/corr.html#mw_ae4a6910-6565-47ce-a488-30ebfb787127, Access Date: 1 May 2023
54. Holm S (1979) A simple sequentially rejective multiple test procedure, Scand J Stat 6(2):65–70
55. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101. https://doi.org/10.2307/1412159
56. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018
57. Li S, Zhang F, Ma L, Ngan KN (2011) Image quality assessment by separately evaluating detail losses and additive impairments. IEEE Trans Multimedia 13(5):935–949. https://doi.org/10.1109/TMM.2011.2152382
58. Hanhart P (2013) VQMT: Video Quality Measurement Tool. https://www.epfl.ch/labs/mmspg/downloads/vqmt/, Access Date: 7 Apr 2021
59. Pinson M (2019) VQM - Video Quality Metric. https://github.com/NTIA/vqm, Access Date: 7 Apr 2021
60. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693. https://doi.org/10.1109/34.192463
61. Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics
62. Simoncelli EP, Freeman WT (1995) The steerable pyramid: a flexible architecture for multi-scale derivative computation. In: Proceedings. International Conference on Image Processing, vol. 3, p 444-447. https://doi.org/10.1109/ICIP.1995.537667
63. Maxwell SE, Delaney HD (2003) Designing experiments and analyzing data : a model comparison perspective, 2nd edn. Lawrence Erlbaum Associates Inc., Mahwah
64. Brunnström K, Barkowsky M (2018) Statistical quality of experience analysis: on planning the sample size and statistical significance testing. J Electron Imaging 27(5):11. https://doi.org/10.1117/1.JEI.27.5.053013
65. Hossfeld T, Schatz R, Egger S (2011) SOS: The MOS is not enough! In: 2011 Third International Workshop on Quality of Multimedia Experience (QoMEX). Mechelen, Belgium: IEEE Xplore, p 131-136. https://doi.org/10.1109/QoMEX.2011.6065690
66. Speranza F, Poulin F, Renaud R, Caron M, Dupras J (2010) Objective and subjective quality assessment with expert and non-expert viewers. In: 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX), p 46-51. https://doi.org/10.1109/QOMEX.2010.5518177
67. Rezaee K, Rezakhani SM, Khosravi MR, Moghimi MK (2021) A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Pers Ubiquit Comput. https://doi.org/10.1007/s00779-021-01586-5

Authors and Affiliations

Kjell Brunnström[1,2] · **Anders Djupsjöbacka**[1] · **Johsan Billingham**[3] · **Katharina Wistel**[3] · **Börje Andrén**[1] · **Oskars Ozolins**[1] · **Nicolas Evans**[3]

✉ Kjell Brunnström
kjell.brunnstrom@ri.se

[1] RISE Research Institutes of Sweden AB, Stockholm, Sweden

[2] Mid Sweden University, Sundsvall, Sweden

[3] Football Technology Innovation Subdivision, Fédération Internationale de Football Association (FIFA), Zürich, Switzerland