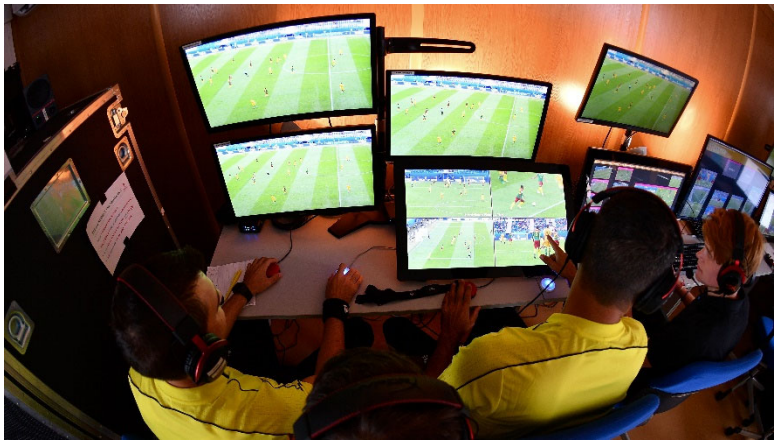
Video quality based on a user study with video professionals for Video Assisted Refereeing

Kjell Brunnström, Anders Djupsjöbacka, Börje Andrén

RISE Research Institutes of Sweden AB

10 March 2021 (ver. 1.3)



RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Table of Contents

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Editorial History

Version	Date	Nature of the modification
0.1	24 June 2019	Initial Draft, edited by Kjell Brunnström, based on VQEG HDTV test plan [1]
1.0	8 July 2019	Version delivered to FIFA for approval
1.2	21 Feb 2021	Changing the title for making it a RISE report. Exchanging subjective test to user study and subjects to test persons
1.3	10 March 2021	Adding RISE report number, ISBN and DOI

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Abstract

This document describes a user experiment with the purpose of finding a baseline quality that is suitable for system and a database for training and evaluating the objective quality measurement methods suitable for assessing the video quality of VAR systems.

A user experiment was performed involving 25 Swedish video experts. Three different video formats were incorporated 1080p, 1080i and 540i. The degradations were in most cased done using encoding with Motion JPEG (MJPEG) and H.264 in the bitrate range from 80 Mbit/s down to 10 Mbit/s.

MJPEG loses quality very fast and already at 80 Mbit/s it has significantly lower quality than the uncompressed reference and then for even lower bitrates the quality falls quickly to bad. On the other hand, H.264 was not found to be significant different from the uncompressed reference until the bitrate had dropped to 10 Mbit/s for 1080p. For 1080i 20 Mbit/s was also weakly and for 540i 20 Mbit/s was significantly lower for some of the scaling methods. For 1080i the deinterlacing requires careful consideration, since the deinterlacing scheme introduced received very low quality scores. For the scaling scheme lanczos was the best and bilinear the worst.

Requirement levels on bitrate and the encoders MJPEG and H.264 based on this experiment

- MJPEG require more than 120 Mbit/s
- H.264 require more than 50 Mbit/s

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

1. Introduction

FIFA has expressed the need to provide technical guidelines (minimum requirements) for any Video assistant refereeing (VAR) system provider that should be approved for the game of football. The concern that FIFA have seen in the current experiments is around the processing of images and the various challenges linked to coding, decoding, synchronizing and re-formatting broadcast feeds. FIFA is therefore seeking experts who would be able to help establish objective test methods that could be used to ensure that a system can provide an adequate solution, see Figure 1. Three measurement points (MP) are defined. MP 0 is where the camera signals enter the Video Operating Room (VOR)[2], MP 1 is just after the video server in the VOR and MP 2 is where the video are sent back to the Outside Broadcast (OB)-van or to broadcast provider.

Their main concerns are:

- Measurement of time synchronicity of broadcast images (immensely important for offside decisions) at MP 1
- Conversion and integration of different formats (1080, ultra-motion cameras, varying frequencies & formats) and image sources into a single system: quality of the resulting output at MP 1
- Measurement of absolute latency of processed images vs. “live” feed at MP 1
- Most importantly: measuring the output image quality from a VAR system back to the broadcaster for transmission on air at MP 2

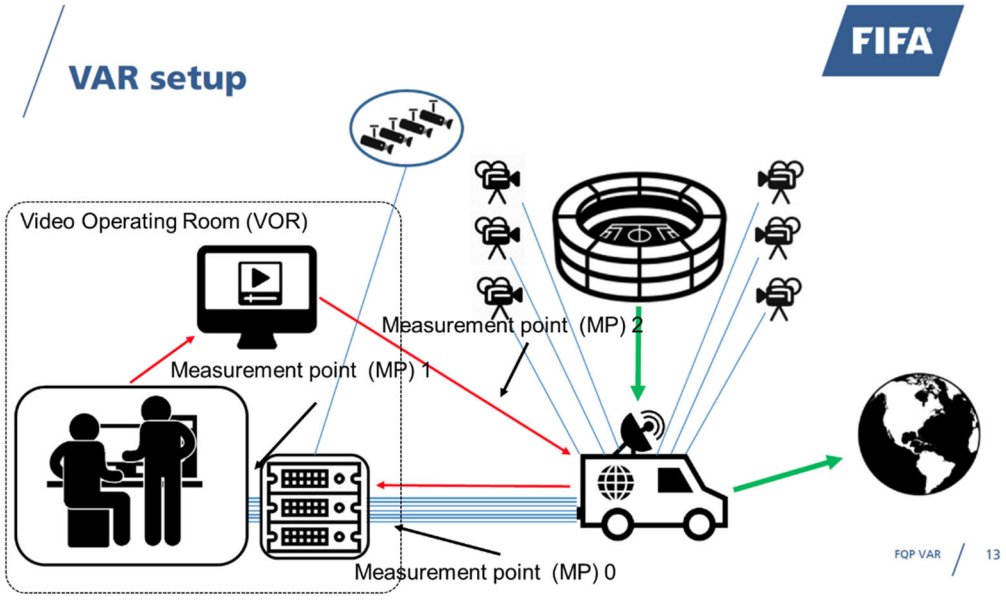


Figure 1: Schematic overview of the VAR setup. Three measurement points (MP) are indicated MP0, MP1 and MP2, for the evaluation of latency, synchronization and video quality.

This document describes a user experiment with purpose of finding a baseline quality that is suitable for VAR systems as well as a database for training and evaluating the objective quality measurement methods. This targets the concern of video quality at MP1 and MP2.

1.1. Changes in implementation of user experiment in relation to test plan

During phase 3, based on the investigations performed and discussions within the project and with FIFA, some changes in the implementation of the user experiment compared to the delivered test plan after phase 2 have been done. One part of the plan was to get distorted video from the current VAR system

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

providers, but that turned out to be quite difficult and time consuming. Although, they were willing to collaborate, the response to return any video material was slow or non-existing. In addition, after obtaining a couple of examples from two VAR system providers, it was also realized that the range of qualities would become very narrow. It was therefore decided to produce a broader range based on standard encoders such as e.g. H.264[3] and Motion JPEG. That also enabled the usage of a single stimulus methodology Absolute Category Rating (ACR) as compared to the double stimulus Double Stimulus Impairment Scale (DSIS), that was suggested in the plan [4, 5], effectively doubling the number of stimuli that could be presented to the test persons. The advantage of this is that a larger variety of content and degradations could be included in the test.

Increased ambition on the number of test persons, with a target of 25

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

2. User video quality tests

The user experiment that was performed assessed the quality of videos as judged by a panel of expert observers.

2.1. User Test Conditions

2.1.1. Viewing and Lab Conditions,

The test room conformed to ITU-R Rec. BT.500-13[4] requirements.

The test persons were seated facing the center of the video display at the specified viewing distance. That means that the test person's eyes were positioned opposite to the video display's center, centered both vertically and horizontally.

A high-end consumer grade 65" 4K TV (Ultra HD, LG OLED65E7V) was used for the experiments, having a resolution 3840 x 2160 pixels. The video was played using a modified version software player VQEGPlayer[6]. As the videos used in the experiment had a lower resolution (1920x1080 and 960x540) than the screen the video was displayed pixel matched in the center of the screen with a grey surround, see Figures 2 and 3. The interlaced 1080i video was deinterlaced in software and the deinterlacing of the TV was not used. The TV was characterized for e.g. luminance, contrast, colours and colour temp and verified to conform to ITU-R BT.709[7].

See Appendix 1 for more detail of the lab conditions

Figure 2: The video was presented in the centre of the screen with a grey surround. Here the 1920x1080 pixels sized video is shown.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Figure 3: The video was presented in the centre of the screen with a grey surround. Here the 960x540 pixels sized video is shown.

2.1.2. Viewing Distance

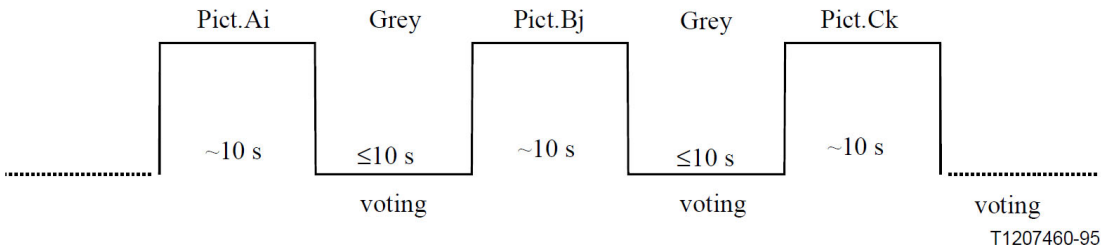
Different video formats and display resolution have been optimized to be viewed from a specific viewing distance, roughly corresponding to an angular pixel pitch of 1 minute of arc. For HD (1920x1080) this is 3H, where H = Picture Height (height of the video window, not the physical display), corresponding to 120 cm in our case. The test persons 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.

2.2. User Test Method

The user test method to be used is named Absolute Category Rating (ACR) method[4, 5], with hidden references. This method is single stimulus i.e. one video is presented at the time and then rated directly afterwards, see Figure 4. High quality or pristine reference videos were mixed in among all the other videos and rated as the others. A voting screen was presented after each video clip had been played, see Figure 5 and the selection was performed with a mouse. The rating was done based on the question "judge the video quality of the video?" (In Swedish: "Bedöm videokvaliteten hos videon?"). The rating scale used was the five graded ACR quality scale (Swedish language translation in parentheses):

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

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68



Ai Sequence A under test condition i
Bj Sequence B under test condition j
Ck Sequence C under test condition k

Figure 4: Stimulus presentation in the ACR method (Figure borrowed from ITU-T P.910[5].

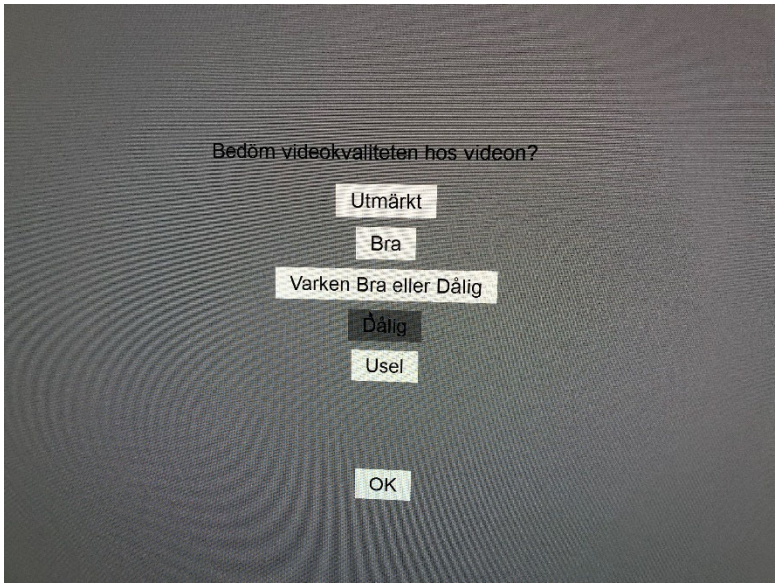


Figure 5: The voting screen used in the experiment.

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 pressing the OK button. This is a modification from the standard procedure, where a fixed time for voting is prescribed. Each test person was given a different randomized order of video sequences.

Three sessions were performed by each test person, but the order was randomized.

1) Full size 1920x1080 video based on progressive source (1080p).
2) Full size 1920x1080 video based on interlaced source (1080i).
3) Quarter size 960x540 video based on interlaced source (540i).

2.3. Length of Sessions

The time of actively viewing videos and voting was about 45 minutes per test person, divided into three sessions with breaks in between. Total session time, including instructions, visual testing, training, pre- and post-questionnaire, was about 1.5 hour.

2.4. Number of test conditions

The time for a trial i.e. time to show a degraded video (process video sequence PVS) including the voting time was about 15 seconds. For 15 minutes sessions, this gave 60 videos to be evaluated per session. The test design was based on 6 source video clips (SRC) crossed with 10 error conditions (HRC), giving 60 PVSs or 6 SRC x 10 HRC.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

2.5. Test persons and User Test Control

In the experiment 25 Swedish speaking video experts participated as test persons.

The term expert is 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 former Agile Solutions AB), which has been used previously by RISE for recruitment of participants in Usability testing. Feedbackfrog took all the contacts with the test persons, scheduled the time to do the experiment and finally also expedited the compensation after the completion of the experiment. The anonymity of the data collected was therefore be ensured as no contact information of the test persons was digitally stored at RISE. The test persons 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).

2.6. Instructions for test persons

Instructions were written out for the test person 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.

2.7. Randomization

The order of the video sequences was randomized for each test person. A randomization process was used in the VQEGPlayer to generate the randomized orders. The same content was not shown directly after each other. In addition, the different sessions were given in different order.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

3. Source Video Sequences

3.1. Selection of Source Sequences (SRC)

The following video formats were used as source sequences in the experiments:

- 1920x1080 progressive 50 frames-per-seconds (1080p)
- 1920x1080 interlaced 50 fields-per-seconds (1080i)

The source sequences were obtained from grabbing video uncompressed from OB-vans at two different locations and times, as well as taken from the Swedish Television (SVT) production Fairytale that was produced for research and standardization purposes[8]. Different video clips were extracted from the longer video sets. The length of each of these were 14 s.

The grabbing of uncompressed video took place in Madrid, more specifically at the Butarque football stadium 30 April 2019, at the same day when FIFA had offside detection testing event. Then both 1080p and 1080i video were grabbed and stored. The other event was at the Tele2 football stadium in Stockholm 14 May 2019 at a football match in the Swedish national football Ligue. Only 1080i video was obtained at this time. The grabbing equipment was in both cases a Blackmagic Design UltraStudio HD Mini grabbing box combined with a laptop computer Lenovo Carbon X1 6th gen (CPU: Intel Core i7-8550U, RAM: 16 GB) and an external 500 GB SSD harddrive (Samsung X5) with Thunderbolt connection. The program used for grabbing was Blackmagic Design Media Express Version 3.5.7 and the format used for grabbing was uncompressed AVI 16 bit 4.2.2 uyvy.

The video clips from SVT are cut-outs from a 6.5-minute-long video Fairytale[8], that was professionally filmed and produced on 65mm analogue film in 50 fps (slow motion up to 100 fps) and then scanned frame by frame while colour correcting and applying film grain noise reduction, to produce the 4K (3840x2160 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 (3840x2164p/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 resolution to 1920. For more details on the production see [8].

3.2. Content

The content was primarily football but was mixed with some other contents to make the videos more diverse, see Table 1 and Table 2. A common thread is that they contained overall a lot of motion to be representative for football.

3.3. Scene Duration

Final source sequences were cut to become10 seconds, but the source scenes used for PVS creation were 14 seconds to have some extra content at the beginning and end to accommodate any transient behaviour of the encoder.

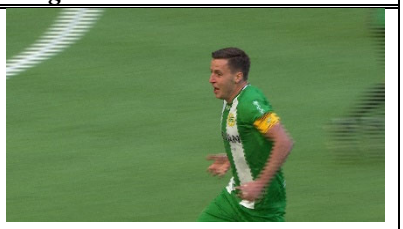
RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Table 1: Description of the source video sequence of the 1080p format

Name	***Description***	***Origin***	***Image***
Butarque_follow_ball_2 (SRC1)	Camera follows the football in an almost empty stadium. Very few people. Motion mostly camera pan	Grabbed at Butarque football stadium	
Butarque_follow_player_2 (SRC2)	Camera follows one football player in an almost empty stadium. Very few people. Motion mostly camera pan	Grabbed at Butarque football stadium	
Butarque_passing_1 (SRC3)	Three players are passing the football. Some camera pan and zoom at the same time as players are moving around	Grabbed at Butarque football stadium	
SVT_CrowdRun (SRC4)	Many people running at the same time. A lot of spatial details and motion. Almost static camera. This sequence is defined in [8] and classified as “difficult” on coding difficulty.	SVT Fairytale[8].	
SVT_RunRun (SRC5)	Many people running at the same time. Slow motion, scene change. This sequence is not defined in [8].	SVT Fairytale[8]	
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 [8] and classified as “difficult” on coding difficulty	SVT Fairytale[8]	
SVT_Searching (used only in the training session)	Some people moving back and forth. Lot of leaves and flowers. Colourful. This sequence is not defined in [8].	SVT Fairytale[8]	

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Table 2: Description of the source video sequence of the 1080i format

Name	***Description***	***Origin***	***Image***
Football_celebrating_1 (SRC1)	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 football stadium	
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 football stadium	
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 football stadium	
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 football stadium.	
SVT_CrowdRun (SRC5)	Many people running at the same time. A lot of spatial details and motion. Almost static camera. This sequence is defined in [8] and classified as "difficult" on coding difficulty.	SVT Fairytale[8]	
SVT_ParkJoy (SRC5)	A few people running at a distance, the camera is moving with them and trees as passing in the foreground. High contrasts. This sequence is defined in [8] and classified as "difficult" on coding difficulty	SVT Fairytale[8]	
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 football stadium.	

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

4. Processed video sequence generation

4.1. Sequence Processing for generating degradations

A Hypothetical Reference Circuit (HRC) is a processing performed by a system or a software that can introduce degradations to the video e.g. video encoding, scaling, transmission etc. 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 were applied to each SRC for each of the formats, making 60 processed video sequences (PVS) per format. All PVSs were 10 seconds long.

A summary of the HRCs are the following:

- 1080p: H.264 (80 Mbit/s – 10 Mbit/s) and Motion JPEG (80 Mbit/s – 20 Mbit/s), see also Table 3.

1080i: H.264 (50 Mbit/s – 10 Mbit/s), Motion JPEG (80 Mbit/s – 20 Mbit/s) and bad deinterlacing, see also

- Table 4.

540i: H.264 (50 Mbit/s – 10 Mbit/s) and different scaling algorithms, see also Table 5.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Table 3: 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 80M
3	Motion JPEG	60 Mbit/s	FFMPEG -c:v mjpeg -b:v 60M
4	H.264	80 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 80M
5	H.264	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50M
6	H.264	30 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 30M
7	H.264	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20M
8	Motion JPEG	40 Mbit/s	FFMPEG -c:v mjpeg -b:v 40M
9	Motion JPEG	20 Mbit/s	FFMPEG -c:v mjpeg -b:v 20M
10	H.264	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10M

Table 4: 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 80M
3	H.264	80 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 80M
4	H.264	50 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 50M
5	H.264	30 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 30M
6	H.264	20 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 20M
7	Motion JPEG	40 Mbit/s	FFMPEG -c:v mjpeg -b:v 40M
8	Motion JPEG	20 Mbit/s	FFMPEG -c:v mjpeg -b:v 20M
9	H.264	10 Mbit/s	FFMPEG -c:v libx264 -profile:v high422 -b:v 10M
10	Deinterlacing	Uncompressed	Blend fields, double the frame rate, top-field-first, VirtualDub (version 1.10.4)

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Table 5: The HRCs or processing applied to the 540i format

HRC	*Processing*	*Bitrates*	*Actual processing*
1	Scaling (lanczos)	Uncompressed	FFMPEG -vcodec rawvideo -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 50M FFMPEG -vcodec rawvideo -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 50M FFMPEG -vcodec rawvideo -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 50M FFMPEG -vcodec rawvideo -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 20M FFMPEG -vcodec rawvideo -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 20M FFMPEG -vcodec rawvideo -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 20M FFMPEG -vcodec rawvideo -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 10M FFMPEG -vcodec rawvideo -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 10M FFMPEG -vcodec rawvideo -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 10M FFMPEG -vcodec rawvideo -pix_fmt uyvy422 -vf scale=960:-1 -sws_flags neighbor

4.2. Sequence Processing for preparing the PVSs for displaying

The experimental program VQEGPlayer[6], takes the raw format BGR and not AVI as a file format and a colour space conversion were therefore applied to PVS:s as a last step. In addition, the interlaced video clips (1080i and 540i) were deinterlaced in software.

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

- FFMPEG -vcodec rawvideo -pix_fmt uyvy422 -vf yadif=1:0:0,mcdeint=3:0:1

The last step in 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.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

For 1080p and 1080i the following command was used

- iconvert.exe uyvy422:$outfile1@1920x1080,frames=100-600 -fyuv444_to_rgb888@mode=709

For 540i the following command was used

- iconvert.exe uyvy422:$outfile1@960x540,frames=100-600 -fyuv444_to_rgb888@mode=709

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

5. Results

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

A characterization of the quality of the video clips is the Mean Opinion Scores (MOS) which is the mean over the ratings given by the test persons

$$MOS_{pvs} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad (1)$$

where μ_{ij} is the score of test person *i* for PVS *j*. *N* is the number of test persons and *M* is the number of PVSs.

The statistical analysis that has been performed is by first applying a repeated measures Analysis of Variance (ANOVA) and then performing a post-hoc analysis based on Tukey Honestly Significant Difference (HSD)[9, 10]. This is an analysis that takes several aspects into consideration at the same time. The MOS comes out as side product, but also if there are significant differences between the MOS.

5.1. 1080p

The mean quality of the source video clips (SRCs) for 1080p taken over all degradations (HRCs) and test persons is shown in Figure 6 left graph (see also Table 1). The MOSs are close to 3 and slightly above indicating that they have about the same amount of high-quality degradations as low degradation with slightly more high quality than low quality. SCR2 (Butarque_follow_player_2) had the overall highest quality, indicating that this video was the least challenging to encode. SCR4 (SVT_CrowdRun) had the overall lowest quality, indicating that this video was the most challenging to encode, followed by SCR6 (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 Figure 6 right graph (see also Table 3) the mean quality of the degradations (HRCs) taken over all source video clips (SRCs) and test persons are shown. It can be noted that there are some HRCs (4-7) that have the almost the same quality (MOS) as the reference (HRC1). These have not been found to be statistically significantly different from the reference. The other HRCs, were statistically significantly different from the HRC1 with p = 0.00001 < 0.05.

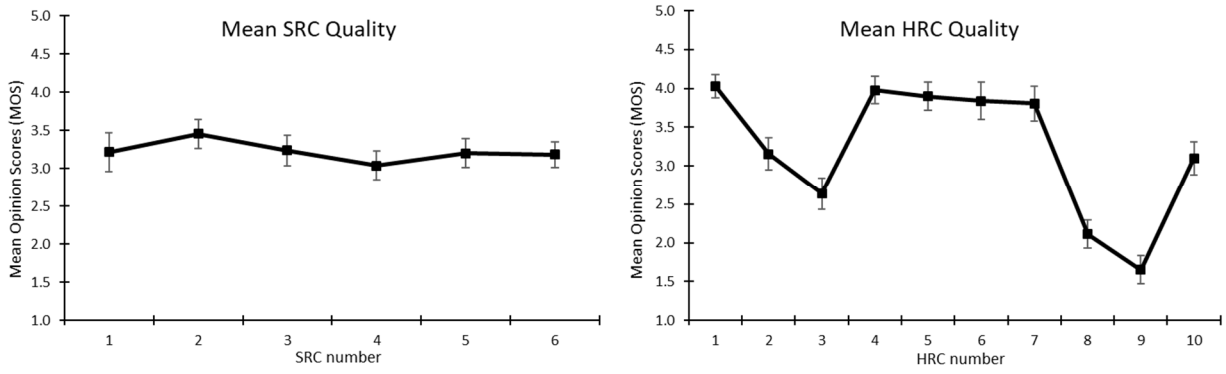


Figure 6: (left) Mean quality of the source video (y-axis) clips (SRCs, x-axis) for 1080p taken over all degradations (HRCs) and test persons, see also Table 1. (right) Mean quality (y-axis) of the degradations (HRCs, x-axis) taken over all source video clips (SRCs) and test persons, see also Table 3.

A more interesting view can be created, if we break down the HCRs into the different processing schemes and bitrates that have been applied to the SRCs, as shown in Figure 7. 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.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

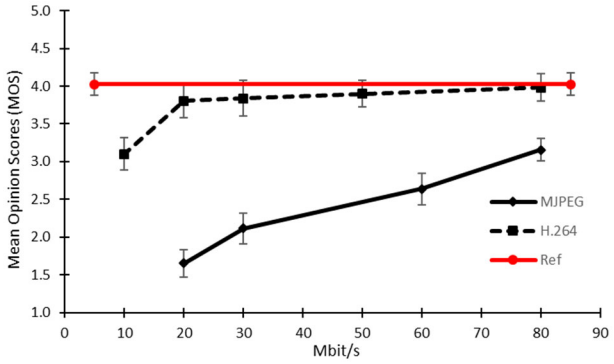


Figure 7: The mean quality (y-axis) of the degradations(HRCs) taken over all source video clips (SRCs) and test persons, divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

We can further break the analysis down to the individual source video clips (SRCs). The Figures 8 – 10 show for each SRC the MOS for the different error conditions in the same way as in Figure 7. It can be noted that there are differences due to the content of the video clips, making them more or less hard to encode. This is usually connected to the amount of fine details and the amount of motion in the video clip. SCR1-3 (Figures 8 and 9 (left)) were not judged to be statistically worse than the reference for any of the H.264 bitrates and not even 10 Mbit/s, although MJPEG still was experienced badly for the lower bitrates. These three videos come from the same recording, showing an empty stadium with few people moving on the pitch, see Table 1, although there are camera pan and individual motion among objects and people, they were less challenging for H.264. Another observation for SRC1 and SRC2 is that the references were scored relatively low i.e. MOS < 4. This could be that these scenes contain some motion blur due to fast camera pan and that could have been hard to distinguish from blur caused by an encoder. SCR4-6 (Figures 9 (right) – 10) were more challenging with a clear drop in quality for H.264 10 Mbit/s and low quality for all the MJPEG bitrates. An interesting observation for SRC5 (Figure 10 (left)) is that the reference was not scored lower than some of the high bitrate error conditions. This is not statistically significant and can be due to just random chance but could also be due to a slow-motion scene (shot in higher frame rates) that has been mistaken as a degradation.

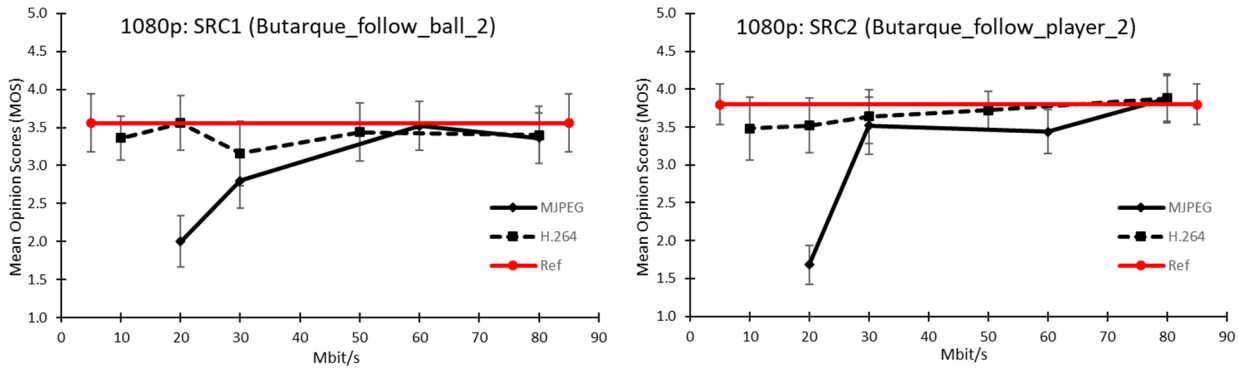


Figure 8: The mean quality (y-axis) of the degradations (HRCs) for SRC1 (left) and SRC2 (right), divided into the different co-decs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

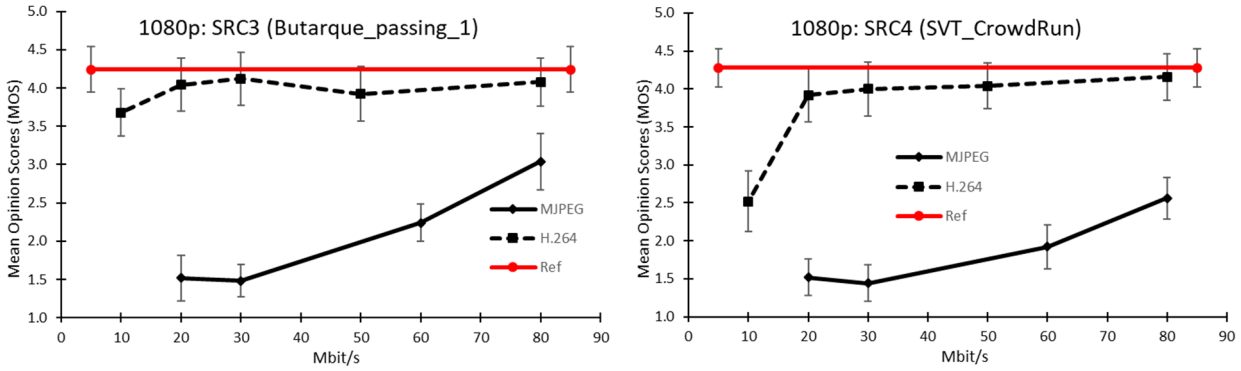


Figure 9: The mean quality (y-axis) of the degradations (HRCs) for SRC3 (left) and SRC4 (right), divided into the different co-decs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

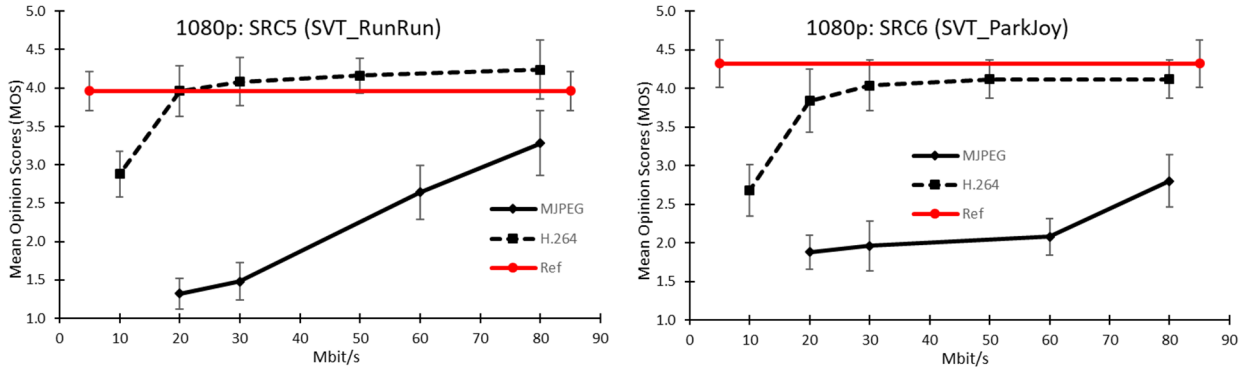


Figure 10: The mean quality (y-axis) of the degradations (HRCs) for SRC5 (left) and SRC6 (right), divided into the different co-decs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

5.2. 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 test persons is shown in Figure 11 left graph (see also Table 2). The MOSs 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 SCR4 (Football_greetings_1, $p = 0.0007$) and SCR2 (Football_goal_1) have lower MOS than SRC1 ($p = 0.005$) and SRC3 ($p = 0.02$).

In Figure 11 right graph (see also Table 4) the mean quality of the degradations (HRCs) taken over all source video clips (SRCs) and test persons are shown. Also, for 1080i there are some HRCs (3-5) that have the almost the same quality (MOS) as the reference (HRC1). These have not been found to be statistically significantly different from the reference. However, HRC6 is statistically significantly lower, even this can be hard to see ($p = 0.03$). The other HRCs, were also statistically significantly different from the HRC1 with $p = 0.00001 < 0.05$.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

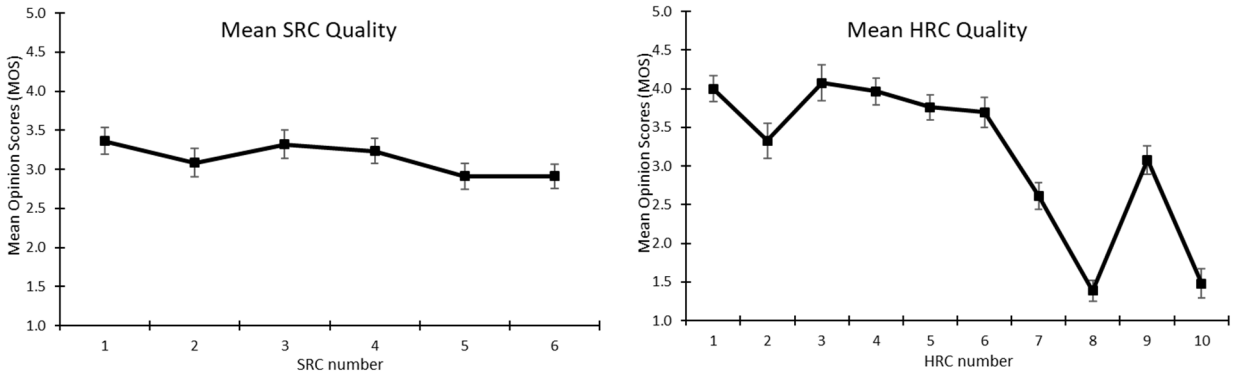


Figure 11: (left) Mean quality (y-axis) of the source video clips (SRCs, x-axis) for 1080i taken over all degradations (HRCs) and test persons, see also Table 2. (right) Mean quality (y-axis) of the degradations (HRCs, x-axis) taken over all source video clips (SRCs) and subjects, see also Table 4.

A breakdown of the HCRs into the different processing schemes and bitrates applied to the SRCs is shown in Figure 12. 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 similar way as the reference HRC1, as a yellow line across the graph. This error condition was not liked very much by the test persons and 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 ($p = 0.03 < 0.05$), although with not very much. This may just be a coincidence due differences in the video material involved.

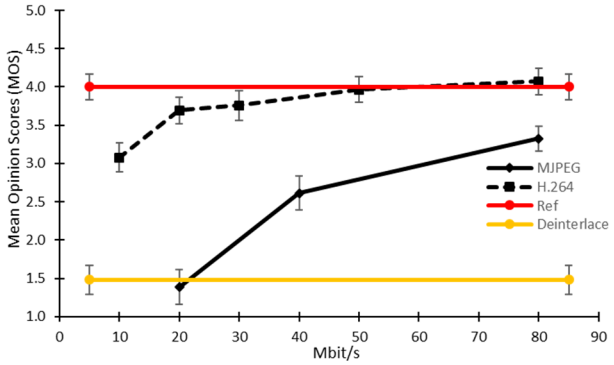


Figure 12: The mean quality (y-axis) of the degradations(HRCs) taken over all source video clips (SRCs) and test persons, divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red 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 is shown as yellow line.

The Figures 13 – 15 shows for each SRC the MOS for the different error conditions in the same way as in Figure 12. SCR2-4 (Figure 13 (right) and 14) were not judged to be statistically worse than the reference for any of the H.264 bitrates and not even 10 Mbit/s, although MJPEG still was experienced badly for the lower bitrates. These three videos come from the same recording, showing footage from a regular football match, see Table 2, as also SRC1 is. SRC2 and SRC 4 are very similar showing an overview of the penalty area where one team is attacking the goal. Although there are camera pan and several players are moving, the camera is almost still at the end of the video clips, making less over all motion in the image. SRC3 is quite different from these other football scenes in that it is showing people moving quite slowly (just a couple moving faster in the background) with a static camera, making the overall motion quite low. SRC1 (Figure 13 (left)) has a more mixed content with fast motion, close-up and overview of the spectators cheering while waving flags, which can explain that it was more challenging than the other three. SRC2 was the only reference that was scored lower than MOS four. Similar to the 1080p case, SCR5-6 (Figure 15) were more challenging with a clear drop in quality for H.264 10 Mbit/s and low quality for all the MJPEG bitrates.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

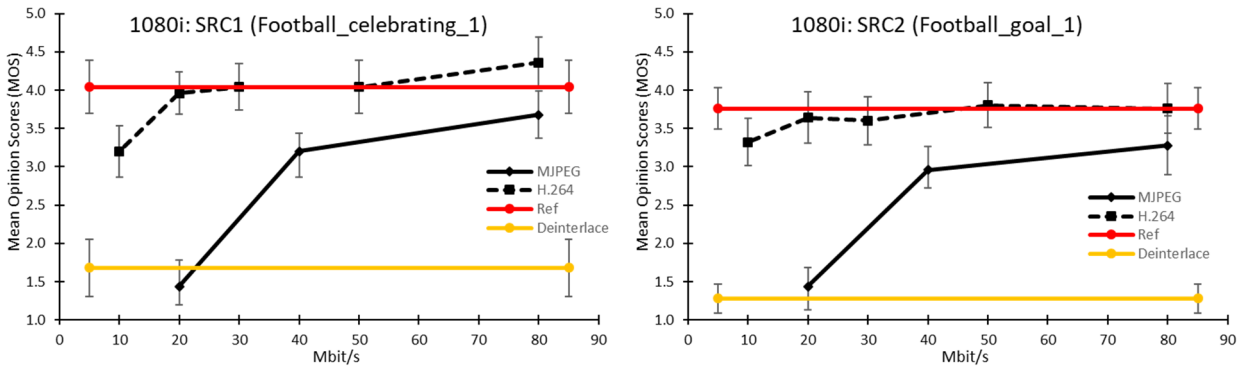


Figure 13: The mean quality (y-axis) of the degradations (HRCs) for SRC1 (left) and SRC2 (right), divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long. Similarly, the error conditions based on simple deinterlacing on an otherwise uncompressed video is shown as yellow line.

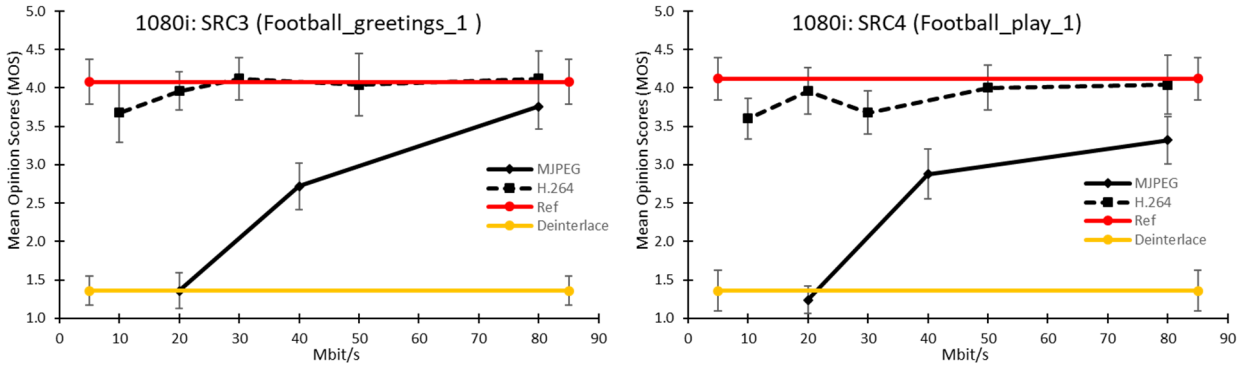


Figure 14: The mean quality (y-axis) of the degradations (HRCs) for SRC3 (left) and SRC4 (right), divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long. Similarly, the error conditions based on simple deinterlacing on an otherwise uncompressed video is shown as yellow line.

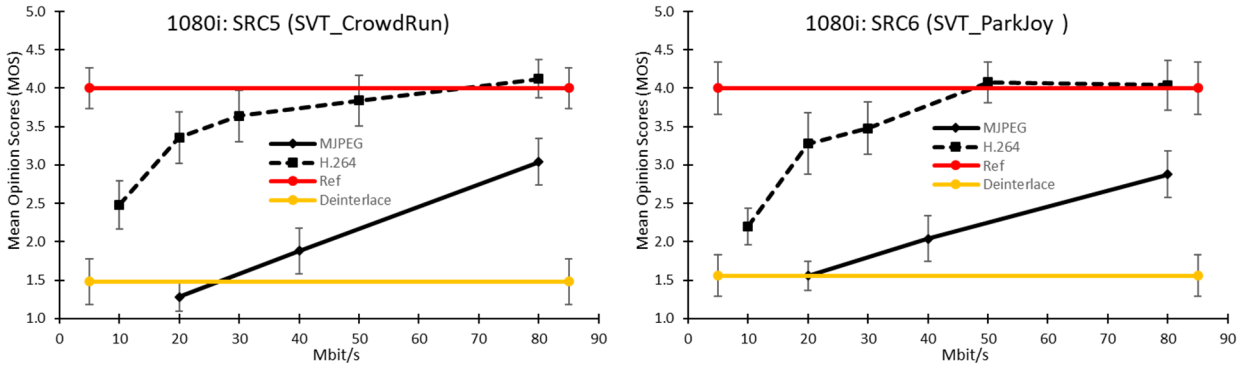


Figure 15: The mean quality (y-axis) of the degradations (HRCs) for SRC5 (left) and SRC6 (right), divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long. Similarly, the error conditions based on simple deinterlacing on an otherwise uncompressed video is shown as yellow line.

5.3. 540i

The mean quality of the source video clips (SRCs) for 540i taken over all degradations (HRCs) and test persons is shown in Figure 16 left graph (see also Table 2). The average quality i.e. the MOSs are here more above 3 than what could be observed for 1080p (Figure 6) and 1080i (Figure 11). 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 <

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

0.05). The MOS of SRC6 (SVT_ParkJoy) was statistically significantly lower ($p < 0.05$) than SRC5 (SVT_CrowdRun, $p = 0.0003$) and SCR4 (Football_play_1, $p = 0.0002$) as well.

In Figure 16 right graph (see also Table 5) the mean quality of the degradations(HRCs) taken over all source video clips (SRCs) and test persons are shown. The different bitrates applied with H.264 comes 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 that 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 neighbour HRC7. HRC3 bilinear 50 Mbit/s was statistically significantly worse than the reference HRC1 and lanczos 50 Mbit/s HRC2. In fact, lanczos were statistically significantly better than bilinear for all the bitrates.

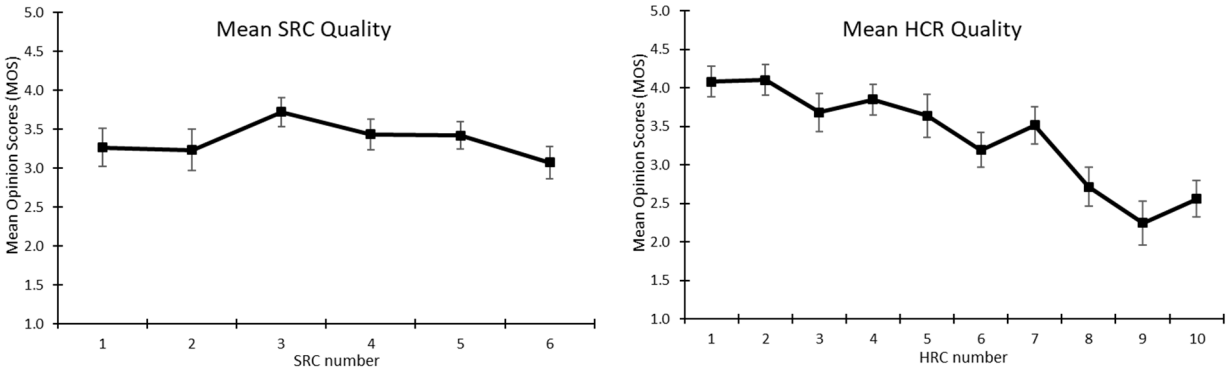


Figure 16: (left) Mean quality (y-axis) of the source video clips (SRCs, x-axis) for 540i taken over all degradations (HRCs) and test persons, see also Table 2. (right) Mean quality (y-axis) of the degradations (HRCs, x-axis) taken over all source video clips (SRCs) and test persons, see also Table 5

.

A breakdown of the HCRs into the different processing schemes and bitrates applied to the SRCs is shown in Figure 17. The different scaling methods are drawn as 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 and hardly any quality decrease for 50 Mbit/s.

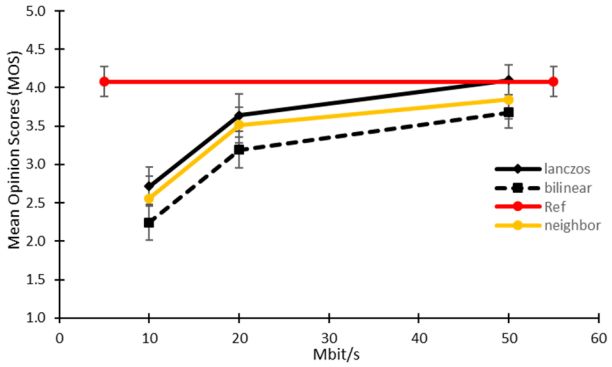


Figure 17: The mean quality (y-axis) of the degradations(HRCs) taken over all source video clips (SRCs) and test persons, divided into the different scaling methods (lanczos in solid black curve, bilinear in dashed black curve and nearest neighbour in yellow curve). The scaling was combined with H.264 encoding. The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

The Figures 18 – 20 show for each SRC the MOS for the different error conditions in the same way as in Figure 17. All of them show more or less the same general behaviour as for the aggregated results in Figure 17. The rank order in quality between the different scaling methods changing a bit between the different SRCs, but lanczos have in most cases the highest quality and bilinear the lowest.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

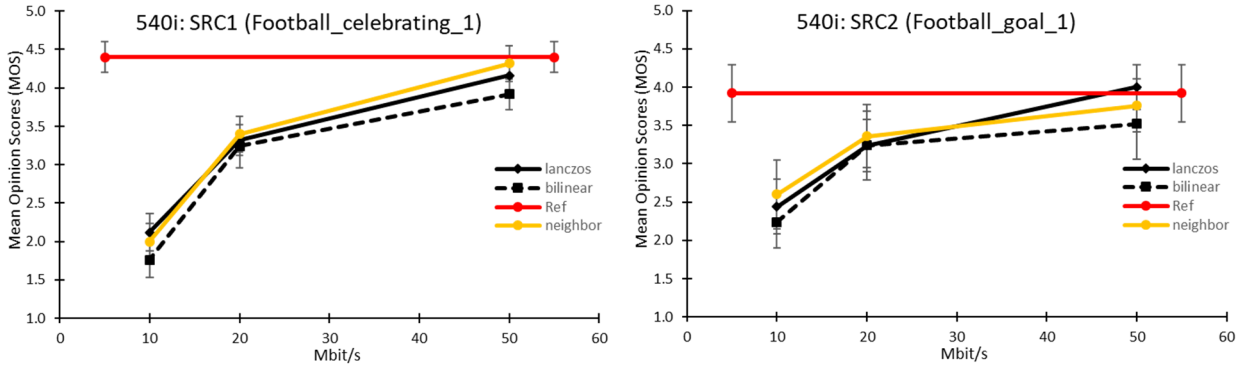


Figure 18: The mean quality (y-axis) of the degradations (HRCs) for SRC1(left) and SRC2(right), divided into the different scaling methods (lanczos in solid black curve, bilinear in dashed black curve and nearest neighbour in yellow curve). The scaling was combined with H.264 encoding. The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

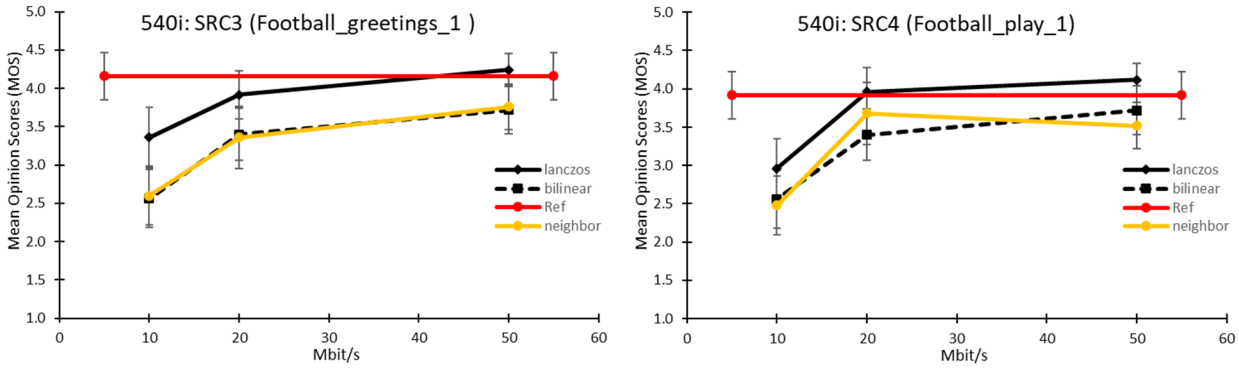


Figure 19: The mean quality (y-axis) of the degradations (HRCs) for SRC3(left) and SRC4(right), divided into the different scaling methods (lanczos in solid black curve, bilinear in dashed black curve and nearest neighbour in yellow curve). The scaling was combined with H.264 encoding. The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

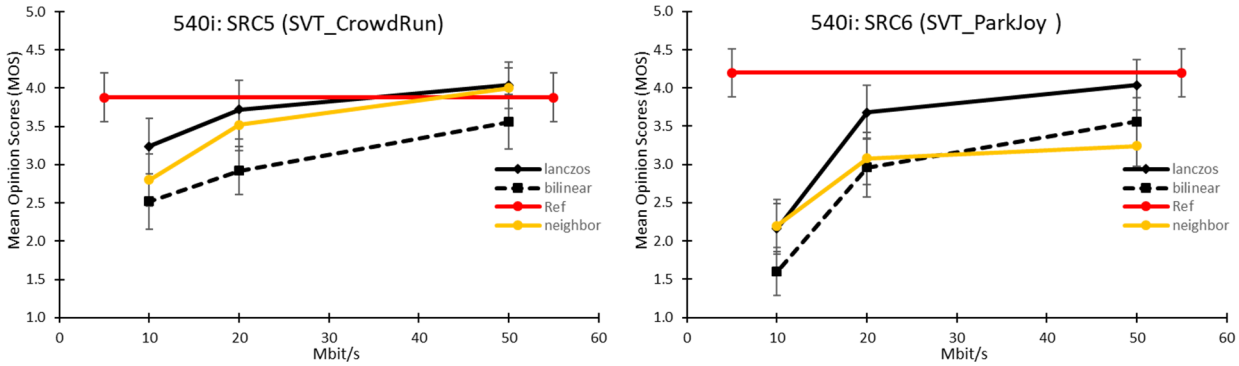


Figure 20: The mean quality (y-axis) of the degradations (HRCs) for SRC5(left) and SRC6(right), divided into the different scaling methods (lanczos in solid black curve, bilinear in dashed black curve and nearest neighbour in yellow curve). The scaling was combined with H.264 encoding. The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

5.4. Requirement levels

The experiment makes it possible to define lower bounds on the bitrates for MJPEG and H.264 to ensure that the quality is Good i.e. MOS = 4. In Figure 21 MJPEG performance has been extrapolated to higher bitrates to see approximately where it crosses 4. Now extrapolation is always very uncertain so the exact value should just be taken as an indication. Calculating the cross points for 1080p it becomes about 116 Mbit/s and for 1080i 121 Mbit/s. Then setting the lower bound for MJPEG at 120 Mbit/s would then give

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

safely good quality, as there is still a margin going slightly lower where viewer would not be able to see the difference.

For H.264 there is a range between 20 – 80 Mbit/s, where a difference could not be found on statistical grounds. However, the trend is still decreasing, so putting the level on 50 Mbit/s for both, would give safe margin, but even 30 Mbit/s would be acceptable.

In summary and based on this experiment:

- MJPEG require more than 120 Mbit/s
- H.264 require more than 50 Mbit/s

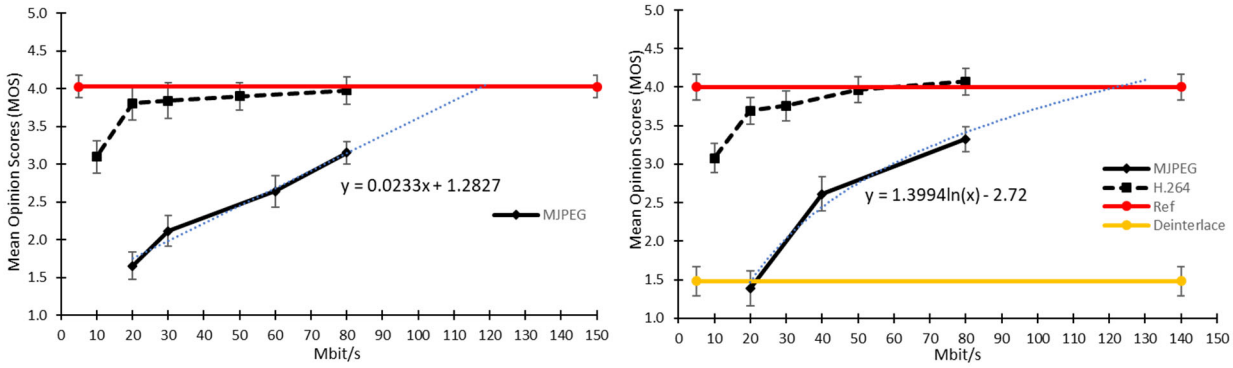


Figure 21: Extrapolating MJPEG for higher bitrates than 80 Mbit/s

5.5. Voting pattern

In Figure 22 histograms of the number of votes per quality category for the different formats are shown. Ideally these should be as even as possible i.e. equal number of votes per category. We can see that the category four is the highest for all formats i.e. Good. This is not so surprising since the experiment targeted basically very high quality video. All the levels have been used though, which means that the experiment spanned the whole range of qualities. Both 1080p and 1080i are quite similar, but 540i had a bit more unbalanced histogram

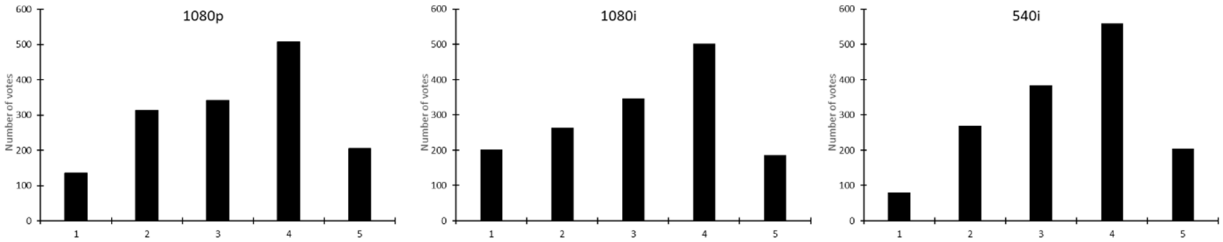


Figure 22: Histograms of the number of votes per quality category for the different formats: 1080p (left), 1080i (middle) and 540i (right)

The voting consistency between the test persons were quite high both for 1080p and 1080i. The mean correlation was 0.81 (standard deviation = 0.06, min = 0.7 and max = 0.9) for 1080p and 0.84 (standard deviation = 0.06, min = 0.7 and max = 0.9) for 1080i. However, for 540i the voting consistency was very low. The mean correlation was 0.71 (standard deviation = 0.07, min = 0.6 and max = 0.85). This indicates that this format was really hard to judge and there may not have been actually visible differences between many of the cases, which would give a more random voting pattern.

5.6. Data about test persons and answers on questionnaires

In total 25 test persons 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 persons 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

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

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 4.5 and standard deviation of 0.58 and only one test person put a 3. Only 2 of the 25 test persons scored below 4.

Very few had experience of being scientific video test persons. The average was as low as 1.48 (standard deviation 0.75) where 1 meant no experience and 5 a lot. Only one test person put a 4 and one put a 3. About 1/3 of the test persons 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 wear glasses or lenses. All had accurate colour vision.

The next 9 questions concern the test evaluation. For all questions it was possible to add a comment. These were often clarifications of the test persons’ answers.

Question 1: How easy was the test evaluation to do (1 = very easy, 5 = very difficult)? Average was 2.82, and standard deviation 0.81. Only one thought it was very easy, none very difficult. No additional comments.

Question 2: Do you think the video disturbances were typical for what you have experienced in your profession (1 = no, 5 = very much)? Average 4.0, standard deviation 1.02, max score 5 (10 test persons), min 1 (only one test person). It is safe to say that the video disturbances were typical. Some comments that explain their answer.

Question 3: Do you think the range of disturbances was typical for what you experience in you work (1 = not at all, 5 = very typical). Average is 3.42, and standard deviation 1.34. Only one test person put 1 and 9 test persons put 5. It could be interpreted as “quite typical”. Few comments.

Question 4: Did you use any specific part(s) of the images in your evaluations that you considered more important? The alternatives were 5; 1 = “faces”, 2 = “movements”, 3 = “image center”, 4 = “sharp edges” and 5 = “the whole image”. You could select more than one. The results are shown in the table below.

Table 6: The number of test persons selecting one of the alternatives in Question 4. They could give more than one alternative.

Alternative	Number of selections
1 = “faces”	10
2 = “movements”	25
3 = “image center”	4
4 = “edges”	18
5 = “whole image”	14

It is obvious that the test persons selected more than one, but all test persons selected “movements” as important for their evaluation of disturbances. A common combination was alternative 2, 4 and 5. Two persons selected all 5 alternatives. This is the question that had the most comments. Most often it concerned other parts or ways to evaluate the image quality e.g. grass, sky, details, artifacts, noise, gradients, panning.

Question 5: Did you get enough instructions before the experiment (1 = no, 5 = very clear)? Average 4.82, and standard deviation 0.37. No one selected alternative 1, 2, or 3, and 20 persons selected alternative 5. The conclusion is that the instructions were clear.

Question 6: Did you have difficulties to concentrate on your evaluation task at any time (alternative 1 = never, alternative 2 = at the end, alternative 3 = at the beginning, alternative 4 = sometimes and alternative 5 =all the time)? On average the results were 1.84 and the standard deviation 1.08. About 50% had no difficulties to concentrate on the task and no one had difficulties all the time. However, about 50% experienced concentration problems, which is not surprising when the task is very visually demanding and time consuming.

Question 7: Did you find your seating ok (alternative 1 bad to 5 excellent)? The average was 4.72 of 5 and the standard deviation 0.60. Only 2 persons gave 3 points and 23 4 or 5 points. The conclusion is that most people had a comfortable seat.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Question 8: Did you move your head much during the experiment (1 = often, 5 = never)? The average is 3.88 and the standard deviation 1.03. Only 1 person gave 1 point and 2 point each and 1/3 say that they did not move their head at all. The conclusion is that people have a tendency to move their heads during such hard and long experiment.

Question 9: Did you experience any disturbances during the experiment (1 = often, 5 = never)? The average answer is 4.40 of 5 and the standard deviation is 1.06. About 84% did not experience any disturbances. One person experienced small disturbances during the first part only.

It seems likely that one other person misunderstood the question 8 and 9 because he put 1 on both these questions without any comment, while he put a 5 on seating (question 7). He also put a 1 = never any problems on concentration. Maybe it had been better to have the “never” alternative as the number 1 alternative or the 5th alternative throughout the questionnaire.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

6. Summary and conclusions

A user experiment was performed involving 25 Swedish video experts in which they judged the perceived quality on a five graded quality scale having the categories, Excellent, Good, Fair, Poor and Bad. The content was selected to contain football and otherwise be relevant for football by containing a lot of motion and moving people. Three different video formats were incorporated 1080p, 1080i and 540i. The degradations were in most cased done using encoding with MJPEG and H.264 in the bitrate range from 80 Mbit/s down to 10 Mbit/s. 1080i also included an error condition that was a simple deinterlacing scheme. 540i was only encoded with H.264 and only in the range of 10 – 50 Mbit/s but had three different scaling schemes instead on top of the encoding.

MJPEG loses quality very fast and already at 80 Mbit/s it has significantly lower quality than the uncompressed reference and then for even lower bitrates the quality falls quickly to bad. On the other hand, H.264 was not found to be significant different from the uncompressed reference until the bitrate had dropped to 10 Mbit/s for 1080p. For 1080i 20 Mbit/s was also weakly significantly lower. For 540i 20 Mbit/s was also significantly lower for some of the scaling methods. For 1080i the deinterlacing requires careful consideration, since the deinterlacing scheme introduced received very low quality scores. For the scaling scheme lanczos was the best and bilinear the worst.

Requirement levels on bitrate and the encoders MJPEG and H.264 based on this experiment

- MJPEG require more than 120 Mbit/s
- H.264 require more than 50 Mbit/s

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

7. Future work

The statistical analysis of the user data can be further refined. Examples of analysis are to standardize the response per test person to take out the effect of the scale has not been used in the same way. It is also possible to try to model the user bias of there are any and in that getting narrower confidence intervals for the error conditions. The video clips should be objectively characterized i.e. to measure the spatial and temporal activity in them. Then some open and published objective models can be tested to see how well they can predict the user scores.

8. Acknowledgements

The contributions of Andreas Langell, for arranging the possibility to acquire footage at Tele 2 arena in Stockholm and providing input on where to find video experts candidates for the user experiment; Benny Norling for providing input on where to find video experts candidates for the user experiment; and Pär Johansson for proof reading, valuable discussions and general support in the project; is hereby gratefully acknowledged.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

9. References

[1]. VQEG. (2009). *HDTV group: Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content (ver 3.0)*. Video Quality Experts Group: www.vqeg.org.

[2]. IFAB. (2016). *Video assistant referees (VARs) experiment - Protocol (Summary)*. The International Football Association Board (IFAB): Münstergasse 9, 8001 Zurich, Switzerland (www.theifab.com).

[3]. MPEG. (2003). *Advanced Video Coding (MPEG 4, H.264)* (ISO/IEC 14496-10, ITU-T H.264). International Standardisation Organisation, International Telecommunication Union, Moving Pictures Experts Group.

[4]. ITU-R. (2012). *Methodology for the subjective assessment of the quality of television pictures* (ITU-R Rec. BT.500-13). International Telecommunication Union, Radiocommunication Sector.

[5]. ITU-T. (1999). *Subjective video quality assessment methods for multimedia applications* (ITU-T Rec. P.910). International Telecommunication Union, Telecommunication standardization sector.

[6]. Brunnström, K., R. Cousseau, J. Jonsson, Y. Koudota, V. Bagazov, and M. Barkowsky, *VQEGPlayer: open source software for subjective video quality experiments in Windows*. 2014, Video Quality Experts Group (VQEG): www.vqeg.org.

[7]. ITU-R. (2002). *Parameter values for the HDTV standards for production and international programme exchange* (Rec. ITU-R BT.709-5). International Telecommunication Union, Radiocommunication Sector.

[8]. Haglund, L. (2006). *The SVT High Definition Multi Format Test Set*. Sveriges Television AB (SVT): Stockholm, Sweden.

[9]. Maxwell, S.E. and H.D. Delaney, *Designing experiments and analyzing data : a model comparison perspective*. 2nd ed. 2003, Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc.

[10]. Brunnström, K. and M. Barkowsky, *Statistical quality of experience analysis: on planning the sample size and statistical significance testing.* Journal of Electronic Imaging, 2018. **27**(5): p. 11, DOI: 10.1117/1.JEI.27.5.053013.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

Appendix 1: A short description of the lab

The perception lab at RISE is designed according to guidelines from ITU-R BT.500-13[4]. The total lab area for this experiment is about 4.35 m × 3.2 m, but the curtains take away some space (see Figure 23).

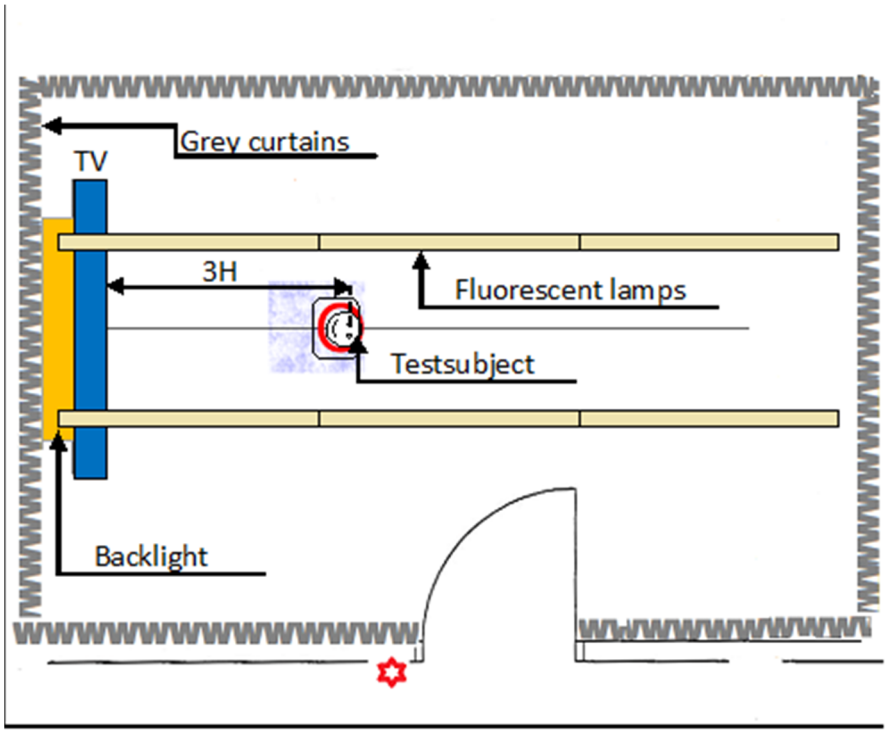


Figure 23: RISE Perception Lab room room set-up for the FIFA experiment

The 4 walls are covered with medium grey (≈50%) curtains and the table tops are covered with the same colour of medium grey cloth to get a neutral visual environment. The distance between the screen surface and the opposite curtain/wall is about 3.7 m. The subject is positioned on a chair at 3 image height (1.2 m) distance from the screen (eyes-screen). The door to the lab has extra sound-proofing to reduce any noise from external sources. The 6 fluorescent lamps in the ceiling are of high frequency type with interchangeable tubes and have a correlated colour temperature of about 6500K. The backlight behind the TV is located about 40 cm behind the TV and has a correlated colour temperature of 6433K (u′=0,1987; v′= 0,4687) and gives about 41 cd/m2, which gives the ratio between the background luminance and the white luminance of the screen of about 14%. The OLEDTV from LG is set to give a white luminance about 300 cd/m2. The exact values are shown in table 1 below. The chromaticity difference compared to D65 (u′= 0.1978 v′= 0.4684) is Δu′v′ about 0.003, which likely is an invisible difference. The edges of the TV are covered with black velvet in order to reduce any reflected light from the edges of the TV.

Table 7 Luminance (Lv), CCT (T), u′, v′ and the chromaticity difference compared to D65 (Δu′v′) and sRGB

Lv	T	u'	v'	**Δu′v′**
297,38	6674	0,1948	0,4691	0,003081

The gamma curve for the LG OLEDTV is presented in Figure 24. The gamma value is about 2.3 compared to sRGBs 2.2, which means that the OLEDTV is very close to the perfect match to sRGB and ITU-R BT.709 but perhaps a little bit more contrast. Black is very dark, about 0.05 cd/m2 which is partly due to the ceiling lighting which gives a vertical illuminance of about 20 lux in front of the screen.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68

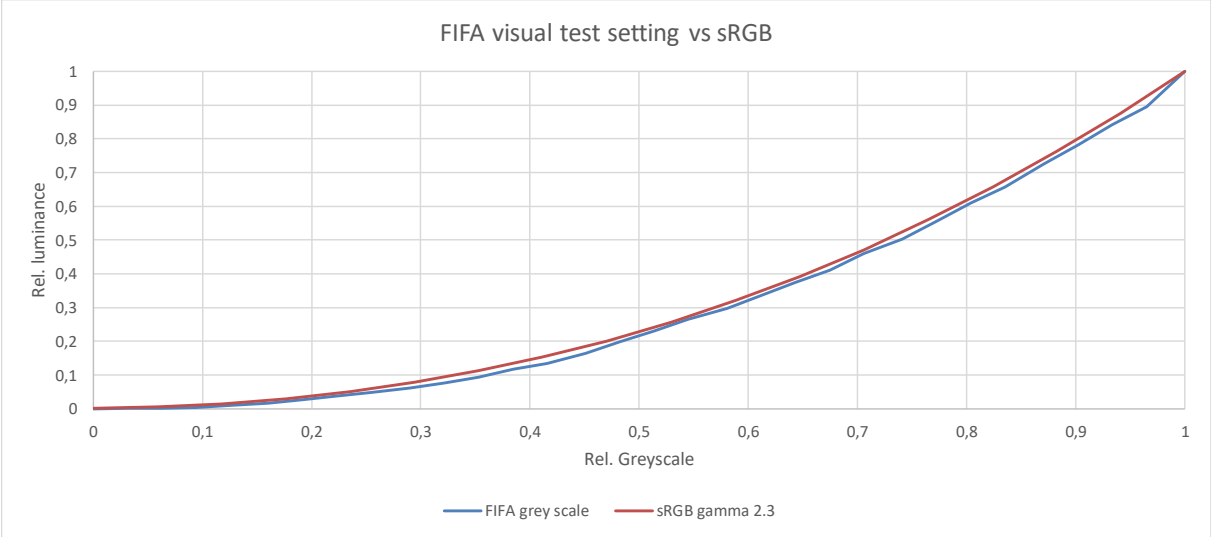


Figure 24 Gamma curve for the LG-OLEDTV with a gamma value of about 2.3.

Figure 25: The seating position of the subject in front of the TV.

RISE report 2021:29
ISBN 978-91-89385-14-6
DOI: 10.23699/79ja-gj68