

Video quality of video professionals for Video Assisted Referee (VAR)

Kjell Brunnström^{a,b}, Anders Djupsjöbacka^a, Johsan Billingham^c, Katharina Wistef, Börje Andrén^a, Oskars Ozoliņš^{a,d}, Nicolas Evans^c
^aRISE Research Institutes of Sweden, Kista, Sweden,
^bMid Sweden University, Sundsvall, Sweden,
^cFédération Internationale de Football Association (FIFA), Zürich, Switzerland
^dRoyal Institute of Technology (KTH), Stockholm, Sweden

Abstract

Changes in the footballing world's approach to technology and innovation contributed to the decision by the International Football Association Board to introduce Video Assistant Referees (VAR). The change meant that under strict protocols referees could use video replays to review decisions in the event of a "clear and obvious error" or a "serious missed incident". This led to the need by Fédération Internationale de Football Association (FIFA) to develop methods for quality control of the VAR-systems, which was done in collaboration with RISE Research Institutes of Sweden AB. One of the important aspects is the video quality. The novelty of this study is that it has performed a user study specifically targeting video experts i.e., to measure the perceived quality of video professionals working with video production as their main occupation. An experiment was performed involving 25 video experts. In addition, six video quality models have been benchmarked against the user data and evaluated to show which of the models could provide the best predictions of perceived quality for this application. Video Quality Metric for variable frame delay (VQM_VFD) had the best performance for both formats, followed by Video Multimethod Assessment Fusion (VMAF) and VQM General model.

Introduction

TV broadcast consists of multiple quality affecting steps from the moment of filming until the video or TV program is aired on TV. The International Telecommunication Union (ITU) identifies three distinct phases within the production and distribution process of TV broadcasting [1]:

- "Contribution – Carriage of signals to production centers where post-production processing may take place.
- Primary distribution – Use of a transmission channel for transferring audio and/or video information to one or several destination points without a view to further post-processing on reception (e.g., from a continuity studio to a transmitter network).
- Secondary distribution – Use of a transmission channel for distribution of programs to viewers at large (by over-the-air broadcasting or by cable television, including retransmission, such as by broadcast repeaters, by satellite master antenna television (SMATV), and by community based-network, e.g., community antenna television (CATV)."

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 is based on standardized procedures [8-17].

The novelty of this research is that it has conducted a user study specifically targeting video experts, as the majority of the research conducted have targeted end or naïve users. Using professionals whose main occupation is video production. The study measured how they perceived the quality of the shown videos. In a second step, six video quality models were benchmarked against the data that was collected in the first phase of the research. With this, it was possible to identify those video quality models that are providing quality predictions with a high degree of confidence in relation to the perceived video quality.

Method

Video quality user study

To measure the users' opinion of the video quality the Absolute Category Rating (ACR) method, with hidden references was used [8-10]. This method uses single stimulus procedure. One video is presented at a time to the user, and they are asked to provide their rating for each video after the video stops. The ratings were provided in this study via a voting interface on the screen, asking the user to "judge the video quality of the video?". The rating scale used was the five graded ACR quality scale:

- 5 Excellent
- 4 Good
- 3 Fair
- 2 Poor
- 1 Bad

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

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

The order in which the different video formats were played to the subjects, as well as the order of the single video sequences within each video format were randomized for each subject. For the video playback and randomization, the VQEGPlayer [32] was used. The time required by the subjects to watch the videos and provide the rating for all three sessions was in total approx. 45 minutes, 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 hours.

There were 60 so called 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. Each video was 10 seconds and with an average estimated voting time of 5 seconds, a trial was about 15 seconds.

Instructions were written out for the subject 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.

To create a controlled and uniform environment for the subjects the test room was set-up to comply with the requirements of the ITU-R Rec. BT.500-13 [8]. A high-end consumer-grade 65" 4K TV (Ultra HD, LG OLED65E7V) was used for the experiments, having a resolution of 3840 x 2160 pixels. 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. The interlaced 1080i video was deinterlaced in software and the deinterlacing of the TV was not used. Viewing distance was 3H i.e., 120 cm.

In the experiment, 25 Swedish-speaking video experts participated as subjects.

All viewers were tested prior for the following:

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

In total 25 video experts participated: 23 males and 2 females. The average age of the test users was 37.8 years, with a standard deviation of 10 years. All subjects 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 color vision.

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:

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

All SRC were obtained as uncompressed videos during live football broadcast productions, as well as from the Swedish Television (SVT) production Fairytale that was produced for research and standardization purposes[18]. From all collected videos, video clips of the length of 14 seconds were extracted.

There were 10 different video processing per video format (including the reference) and each video processing was 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 video processing is the following:

- 1080p: H.264 (80 Mbit/s – 10 Mbit/s) and Motion JPEG (80 Mbit/s – 20 Mbit/s).
- 1080i: H.264 (50 Mbit/s – 10 Mbit/s), Motion JPEG (80 Mbit/s – 20 Mbit/s) and bad deinterlacing.
- 540i: H.264 (50 Mbit/s – 10 Mbit/s) and different scaling algorithms (lanczos, bilinear and neighbor).

Objective video quality assessment methods

Objective video quality models were evaluated for their performance on the video format 1080p and 1080i. The methods considered were:

- Video Multimethod Assessment Fusion (VMAF)[3]
- Video Quality Metric (VQM) – General model (ITU-T Rec. J.144)[5]
- Video Quality Metric (VQM) – (VQM_VFD)[2]
- Peak Signal to Noise Ratio (PSNR) ITU-T Rec J.340[19] [20]
- Structural Similarity Index (SSIM) [21] [20]
- Visual Information Fidelity (VIF) [22] [20]

Results

Video quality user study

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

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

where μ_{ij} is the score of the user i for PVS j . N is the number of users 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)[23, 24].

In Figure 1 the different video processing schemes and bitrates that were applied to the SRCs for 1080p are shown. 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.

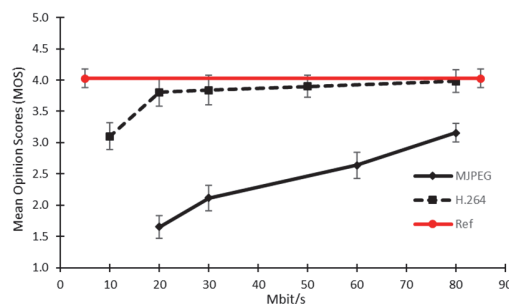


Figure 1. The mean quality for 1080p (y-axis) of the degradations taken over all source video clips (SRCs) and users, divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line without tying it to the bitrate to not make the x-axis too long.

A breakdown of the different processing schemes and bitrates applied to the SRCs for 1080i is shown in Figure 2. 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. One error condition was a simple deinterlacing applied directly to the uncompressed video and its MOS has been drawn in a similar

way as the reference, as a yellow line across the graph. This error condition was not liked very much by the users 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 ($p = 0.03 < 0.05$).

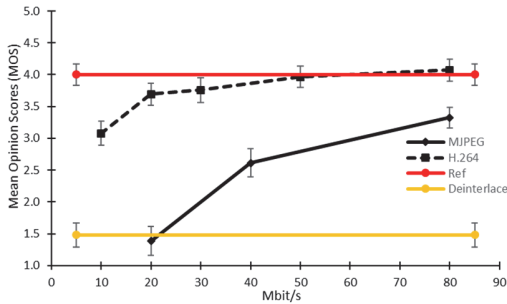


Figure 2. The mean quality for 1080i (y-axis) of the degradations taken over all source video clips (SRCs) and users, divided into the different codecs used (MJPEG in solid black curve and H.264 dashed black curve) and plotted against the bitrate (x-axis). The MOS of the reference is marked as red line 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.

Objective video quality models evaluation

In the evaluation, we have studied the overall performance given by Pearson Correlation Coefficient (PCC)[25] and the Root Mean Square Error (RMSE)[25], between the scores of the objective model and the Difference Mean Opinion Scores (DMOS). 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: difference score = 5 - (reference score - distorted score). The PCC measures the linear relationship between the model scores and the DMOS. 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[25]. This usually improves the PCC somewhat, but it also enables the calculation of the RMSE. A statistical hypothesis test was also applied to the RMSE values. The null hypothesis, H_0 , is 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 is the number of points in the RMSE calculation, minus 4 due to the 3rd order monotonic polynomial fit i.e. $54 - 4 = 50$ [25]. The Spearman Correlation Coefficient (SCC) was also calculated.

The p-values of the statistical significance tests are shown in Table 1 and Table 2. 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.

Table 1: P-values of statistical test on the difference in RMSE based on ITU-T Rec. P.1401[25] for 1080p. Significant values are marked with *, based on an alpha of 0.05 and the method of Holm 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 2: P-values of statistical test on the difference in RMSE based on ITU-T Rec. P.1401[25] for 1080i. Significant values are marked with *, based on an alpha of 0.05 and the method of Holm 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

Conclusions

The performance of six different video quality models has been evaluated for 1080p and 1080i. VQM_VFD had the best performance for both formats, followed by VMAF and VQM General models. SSIM, PSNR, and VIF have similar performance that is lower than the evaluated video models. SSIM has particularly low performance for 1080i, mostly due to the low-quality deinterlacing method, but from the scatter plots it is evident that also PSNR and VIF had similar problems.

References

- [1]. ITU-T. (2021). *Transmission and delivery control of television and sound programme signal for contribution, primary distribution and secondary distribution*. Available from: <https://www.itu.int/en/ITU-T/studygroups/2017-2020/09/Pages/q1.aspx>, Access Date: 16 April 2021.
- [2]. Wolf, S. and M. Pinson. (2011). *Video Quality Model for Variable Frame Delay (VQM_VFD)* (NTIA Technical Memorandum TM-11-482). National Telecommunications and Information Administration (NTIA), Boulder, CO, USA.
- [3]. Li, Z., A. Aaron, I. Katsavounidis, A.K. Moorthy, and M. Manohara (2016). *Toward A Practical Perceptual Video Quality Metric*. Netflix Technology Blog. Available from: <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., S. Möller, and A. Perkis. (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/CoE_whitepaper_v1.2.pdf)), Lausanne, Switzerland.
- [8]. ITU-R. (2019). *Methodology for the subjective assessment of the quality of television pictures* (ITU-R Rec. BT.500-14). 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. (2014). *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., H. Choi, E. Lee, S. Lee, and J. Choe. (2006). *Comparison of various subjective video quality assessment methods. in Image Quality and System Performance III*. Bellingham, WA: SPIE.
- [12]. Huynh-Thu, Q. and M. Ghanbari. (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. p. 70-76.
- [13]. Tominaga, T., T. Hayashi, J. Okamoto, and A. Takahashi. (2010). *Performance Comparisons of Subjective Quality Assessment Methods for Mobile Video. in Second International Workshop on Quality of Multimedia Experience (QoMEX 2010)*. Trondheim, Norway. p. 82-87, DOI: 10.1109/QoMEX.2010.5517948.
- [14]. Berger, K., Y. Koudota, M. Barkowsky, and P.L. Callet. (2015). *Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains. in 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. 2015.
- [15]. Pitrey, Y., M. Barkowsky, P. Le Callet, and R. Pépion. (2010). *Subjective Quality Evaluation of H.264 High-Definition Video Coding versus Spatial Up-Scaling and Interlacing. in Euro ITV*. 2010. Tampere, Finland.
- [16]. Barkowsky, M., S. N., L. Janowski, Y. Koudota, M. Leszczuk, M. Urvoy, P. Hummelbrunner, I. Sedano, and K. Brunnström. (2012). *Subjective experiment dataset for joint development of hybrid video quality measurement algorithms*.
- [17]. Choe, J.-H., T.-U. Jeong, H. Choi, E.-J. Lee, S.-W. Lee, and C.-H. Lee, (2007). *Subjective Video Quality Assessment Methods for Multimedia Applications*. Journal of Broadcast Engineering. **12**(DOI: 10.5909/JBE.2007.12.2.177).
- [18]. Haglund, L. (2006). *The SVT High Definition Multi Format Test Set*. Sveriges Television AB (SVT), Stockholm, Sweden.
- [19]. ITU-T. (2010). *Reference algorithm for computing peak signal 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.
- [20]. Hanhart, P. (2013). *VQMT: Video Quality Measurement Tool*. Available from: <https://www.epfl.ch/labs/mmspg/downloads/vqmt/>, Access Date: 7 April 2021.
- [21]. Wang, Z., A.C. Bovik, H.R. Sheikh, and E.P. Simonelli, (2004). *Image quality assessment: From error visibility to structural similarity*. IEEE Transactions on Image Processing. **13**(4): p. 600-612.
- [22]. Sheikh, H.R. and A.C. Bovik, (2006). *Image information and visual quality*. IEEE Transactions on Image Processing. **15**(2): p. 430-444, DOI: 10.1109/TIP.2005.859378.
- [23]. Maxwell, S.E. and H.D. Delaney, (2003). *Designing experiments and analyzing data : a model comparison perspective*. 2nd ed. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, Inc.,
- [24]. Brunnström, K. and M. Barkowsky, (2018). *Statistical quality of experience analysis: on planning the sample size and statistical significance testing*. Journal of Electronic Imaging. **27**(5): p. 11, DOI: 10.1117/1.JEI.27.5.053013.
- [25]. ITU-T. (2020). *Statistical analysis, evaluation and reporting guidelines of quality measurements* (ITU-T P.1401). International Telecommunication Union, Telecommunication standardization sector, Geneva, Switzerland.

Acknowledgement

This work was mainly funded by Fédération Internationale de Football Association (FIFA) and partly supported by the Sweden's Innovation Agency (VINNOVA, dnr. 2021-02107) through the Celtic-Next project IMMINENCE (C2020/2-2) as well as RISE internal funding.

Author Biography

Kjell Brunnström is a Senior Scientist at RISE Research Institutes of Sweden AB and Adjunct Professor at Mid Sweden University. He is leading development for video quality assessment as Co-chair of the Video Quality Experts Group (VQEG). His research interests are in Quality of Experience for visual media especially immersive media. He is area editor of the Elsevier Journal of Signal Processing: Image Communication and has co-authored > 100 peer-reviewed scientific articles including conference papers.

Anders Djupsjöbacka was born in Solna, Sweden in 1958. He received a M.Sc. degree in 1982, and a Ph.D. degree in 1995. From 1982 to 2002, he was at Ericsson Telecom AB where he worked with high-speed optical transmission. In 2002 he joined Acreo AB (that later become RISE) continuing in the same area. In 2015 he joined the Visual Media Quality group where he performed display-quality measurements, and assessments of video quality.

Johsan Billingham received a BEng in Sports material from Swansea University in 2014 and a MSc in Sports Engineering from the Sheffield Hallam University in 2015. He joined FIFA in 2015 and is now working as a Research Manager with a role to harness applied research in the areas of computer vision, material engineering, advanced modelling, biomechanics, data analytics, as well as many others to better understand key challenges and opportunities in football.

Katharina Wistel holds a diploma in Sports and Event Management of the European School of Higher Education. In 2011 Wistel joined FIFA and is now Group Leader of the FIFA Quality Programme. In this role she is driving the development and implementation of new quality standards and is in close exchange with the football industry about new technologies. Besides, she is currently studying Business Psychology (BSc) at Kalaidos University of Applied Sciences Switzerland.

Oskars Ozoliņš is a Senior Scientist and Technical Lead at the Kista High-speed Transmission Lab (Kista HST-Lab), RISE Research Institutes of Sweden. He is also an Associate professor at the Department of Applied Physics, KTH Royal Institute of Technology. His research interests are in the areas of digital and photonic-assisted signal processing techniques, high-speed short-reach communications and devices, optical and photonic-wireless interconnects, and machine learning for optical network monitoring and Quality of Experience prediction.

Börje Andrén has a MSc from the Royal Institute of Technology (KTH). He is a Senior Scientist at RISE Research Institutes of Sweden AB and has worked with optical research, image quality and colour issues and visual ergonomics for both 2D and 3D for almost 43 years. He has participated in

the development of the visual ergonomic part of the TCO Certified since 1995 and developed requirements and test methods.

Nicolas Evans is the Head of Football Research & Standards within FIFA's Technology Innovation Sub-Division. He has been at the heart of standards creation, innovation management and the new data ecosystem at

FIFA for more than a decade, leading research & validation efforts for new technologies. He is part of a multi-disciplinary team consisting of industry experts, engineers and data scientists that works with more than 150 stakeholders (industry, academia, football clubs/federation) on a daily basis.