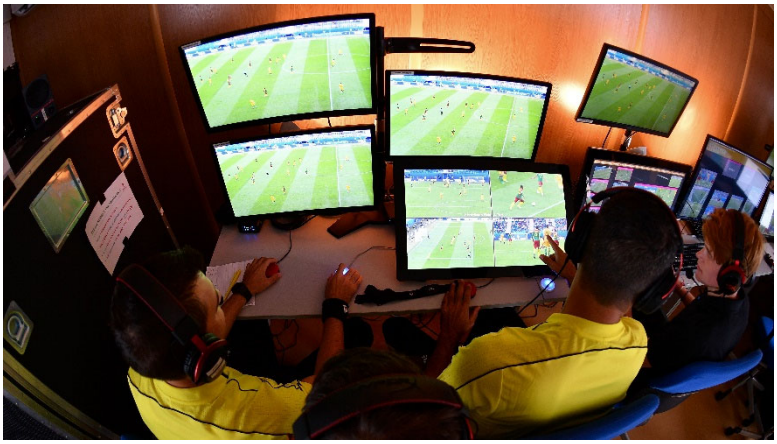
Objective video quality assessment methods for Video assistant refereeing (VAR) System

Kjell Brunnström, Anders Djupsjöbacka, Börje Andrén

RISE Research Institute of Sweden AB

19 Feb 2021 (ver 1.1)



RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

Table of Contents

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

Abstract

This report describes the work and conclusions drawn after phase 4 in the project “Assessment methods for Video assistant refereeing (VAR) System”.

The performance of six different video quality models have been evaluated, that were identified during phase 1, against the subjective video quality database that was created during phase 3. The results are slightly different for 1080p compared to 1080i. For 1080p the models VQM_VFD, SSIM and VMAF performs the best with Pearson Correlation Coefficients (PCC) above 0.9. For 1080i the PCC drops a bit overall and then VMAF and VQM_VFD are close in performance and performing the best. The overall performance for both formats VMAF an VQM_VFD stands out as the best models. In this comparison VQM_VFD has the added advantage to also be able to perform its own registration i.e. to fix any misalignment between the reference video and the distorted one.

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

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 **Error! Reference source not found.**.

Error! Reference source not found.. Three measurement points (MP) are defined. MP 0 is where the camera signals enter the Video Operating Room (VOR)[1], 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 MP1
- 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 MP1
- Measurement of absolute latency of processed images vs. "live" feed at MP1
- Most importantly: measuring the output image quality from a VAR system back to the broadcaster for transmission on air at MP2

This report describes the work and conclusions drawn after phase 4 in the project "Assessment methods for Video assistant refereeing (VAR) System" [2]. Based on the survey performed during phase 1 some video quality models were identified as promising and with the subjective data collected during phase 3 [3], an evaluation of these models has been performed in phase 4. The work was connected to the Tasks 4.6 in the project plan[2].

Project team for phase 4

Kjell Brunnström, RISE (project manager, responsible for video quality measurements)

Anders Djupsjöbacka, RISE (responsible for latency and synchronization)

Börje Andrén, RISE (technical assistance)

Pär Johanson, RISE (business and technical assistance)

Benny Norling, Boresight (consulting with long experience in broadcasting)

Andreas Langell, Mobilelink (consulting with long experience in broadcasting)

2. Method

For practical measurements, we have implemented a video extraction software, so when testing a VAR-system a known video is ingested into the system, the video is grabbed after being processed by the VAR-system. The video extraction software can then extract the video that was ingested even if it is embedded in other contents.

In the evaluation we have studied the overall performance given in by Pearson Correlation Coefficient (PCC)[4], between the scores of the objective model and the subjective difference mean opinion scores (DMOS). The DMOS is calculated by subtracting for each subject its rating of the reference from the rating of the distorted video. In reality to get the values on the same scale as the Mean Opinion Scores (MOS) i.e. 1-5, the following formula is used: *difference score = 5 – (reference score – distorted score)*. The PCC measure the linear relationship between the model scores and the DMOS.

Another criterion for a useful method is whether it can handle to register 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 small shift can therefore have a big impact on the score, which

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

is not reflected by a quality difference. The Video Quality Metric (VQM)[5] software contains also 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 performance of other algorithms. This registration method has been standardized by the ITU (ITU-T Rec J.244).

3. : Survey objective video quality models

Objective video quality models are mathematical models that approximate results from subjective quality assessment, in which human observers are asked to rate the quality of a video. Various models and approaches have been suggested in the literature[6]. They are very often subdivided into groups depending on the amount of reference information that the models are using.

- Full reference (FR) – access to a high quality or non-distorted version 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.
- No reference (NR) – no access to the reference

Although a huge number of models have been proposed in literature, the models that can be of interest for evaluating the VAR-system is a rather mature method preferable with a product on the market or at least a stable software, if interfacing to a regular computer can be satisfactory solved.

We will primarily target FR models, because they are more accurate then especially the NR models and the high-quality range that the FIFA use case is covering, makes it more likely to find an effective method based on FR. 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.

3.1. General requirements for methods

Ideally a hardware method that can take as input an SDI-signal at MP 0 as reference and then measure the video quality at MP 1 and MP 2 compared to the that reference. Alternatively, the video signal is stored by having an SDI-capture card in a computer. The quality is then computed by using software. The advantage with the first i.e. hardware solution is that it would be possible to just take the equipment out to the installation and measure the video quality. The advantage with latter is that the number of possible algorithms increases substantially. It would also be easier to upgrade if a new and better algorithm is developed.

A general problem is to identify reference video sequences through the VAR-setup, especially if integration of different feeds is taking place.

The primarily targeted picture format is the HD formats 1080p and 1080i, although it would most likely have to work for 4K.

The method should be optimized for a high-quality sports content i.e. bitrates on the order of 100 Mbit/s for 1080p or a bit lower for 1080i. Codecs used in VAR-system could be: MJPEG, XAVC-I, AVC-I, DNxHD and ProRes for HD/FHD

The degradations that can occur are (exaggerated examples below):

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

- blur

- blockiness

- double edges

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

- de-interlacing

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

3.2. Evaluated methods

A number of commercial methods were explored, and some were also evaluated. These results are not covered in this open report, see also the report[3].

3.2.1. Video Multimethod Assessment Fusion (VMAF)

Video Multimethod Assessment Fusion (VMAF)[7] is 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)

From the description of VMAF by Netflix[7].

"VMAF, that predicts subjective quality by combining multiple elementary quality metrics. The basic rationale is that each elementary metric may have its own strengths and weaknesses with respect to the source content characteristics, type of artifacts, and degree of distortion. By 'fusing' elementary metrics into a final metric using a machine-learning algorithm—in our case, a Support Vector Machine (SVM) regressor—which assigns weights to each elementary metric, the final metric could preserve all the strengths of the individual metrics and deliver a more accurate final score."

3.2.2. Video Quality Metric (VQM)

The Video Quality Metric (VQM) has been developed by 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)[8], but has shown very good performance, especially the updated version VQM for variable frame delay (VQM_VFD)[9]. It is available as open source software.

In an email correspondence with Margaret Pinson, she said the following. "The broadcasters and MPEG folk tell me that for contribution quality, PSNR is the best option. However, you cannot trust comparisons between two SRC. PSNR has a very strong SRC bias. The broadcasters I've talked to in the ITU all use PSNR for contribution quality video (e.g., > 20 Mbits/sec), and just live with this problem.

When the bitrates drop to distribution quality (e.g., < 12 Mbits/sec), then my models were trained on these bitrates and quality levels. VQM-VFD should be the best choice of mine. Netflix's evaluation put VMAF slightly ahead of VQM-VFD, and it should also work well for this range of quality. If these are the bitrates of interest, I recommend you ask the customer for a few sample videos. Run VQM-VFD, VMAF, and PSNR; then see what you think.

The problem for both of these models (and the other top models) is that distribution quality video spans a small range of quality. I am not at all sure that your customer will be happy with the prediction accuracy. This is why I recommend you look over some numbers as a sanity check. It sounds like your customer is interested in the higher bitrates, though, so PSNR is likely your best choice."

3.2.3. Peak Signal to Noise Ratio (PSNR)

Peak Signal to Noise Ratio (PSNR) is the most commonly used video quality method. It is actually standardized by ITU in ITU-T Rec J.340[10]. 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 system should fulfil.

An open source implementation is provided by ITS/NTIA (https://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx)

3.2.4. Structural Similarity Index (SSIM)

The Structural Similarity Index (SSIM) is an image quality methods that has been popular since it was first published[11]. It was developed to improve compared to PSNR but is still very simple. The idea is to

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

compare the structure between two images. It is defined for images and do not specify how it should be used for video. The paper[11] is one of the most cited papers in the image processing community.

4. Results

4.1. Video Multimethod Assessment Fusion (VMAF)

Video Multimethod Assessment Fusion (VMAF)[7] is 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). The performance using the default training, was a PCC of 0.91 for both1080p and 1080i, which is very good. This implementation of VMAF needs external registration software.

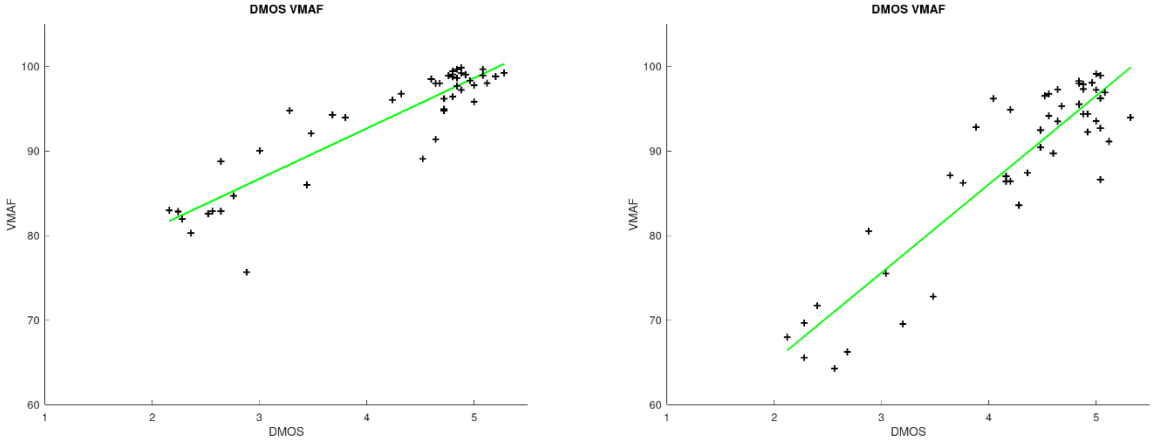


Figure 1: Scatterplots of VMAF performance. 1080p is to the left and 1080i to the right.

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

4.2. Video Quality Metric (VQM)

The Video Quality Metric (VQM) has been developed by 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)[8], but has shown very good performance, especially the updated version VQM for variable frame delay (VQM_VFD)[9]. It is available as open source software.

VQM General model (ITU-T Rec. J.144) was evaluated at RISE and the performance was good, with a PCC of -0.84 for 1080p and -0.85 for 1080i, see Figure 2. 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.

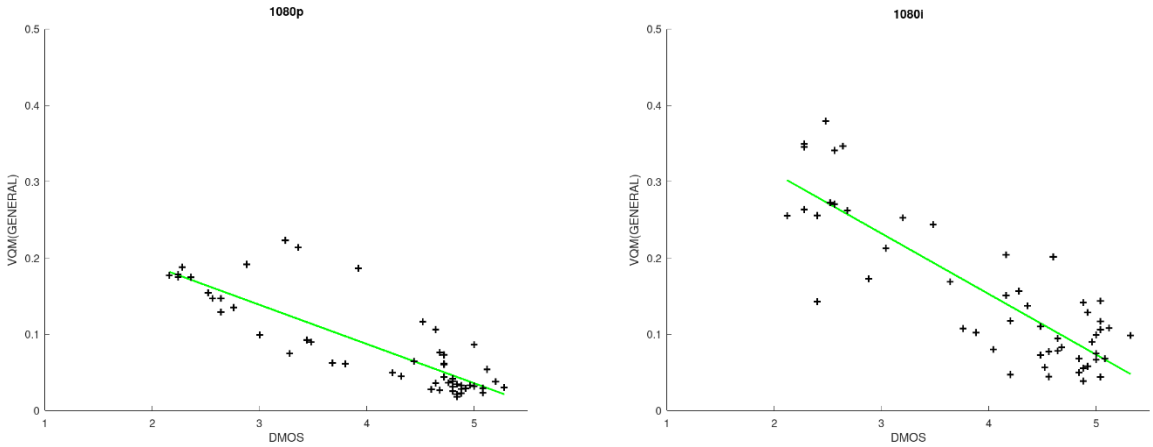


Figure 2: Scatterplots of VQM General performance. 1080p is to the left and 1080i to the right.

VQM has also an improved model, called VQM_VFD, where VFD stands for Variable Frame Delay. RISE evaluated this version too, which had really good performance, with a PCC of -0.96 for 1080p and -0.88 for 1080i, see Figure 3. VQM quality metric does not need an external registration software and the registration algorithm can be used on its own.

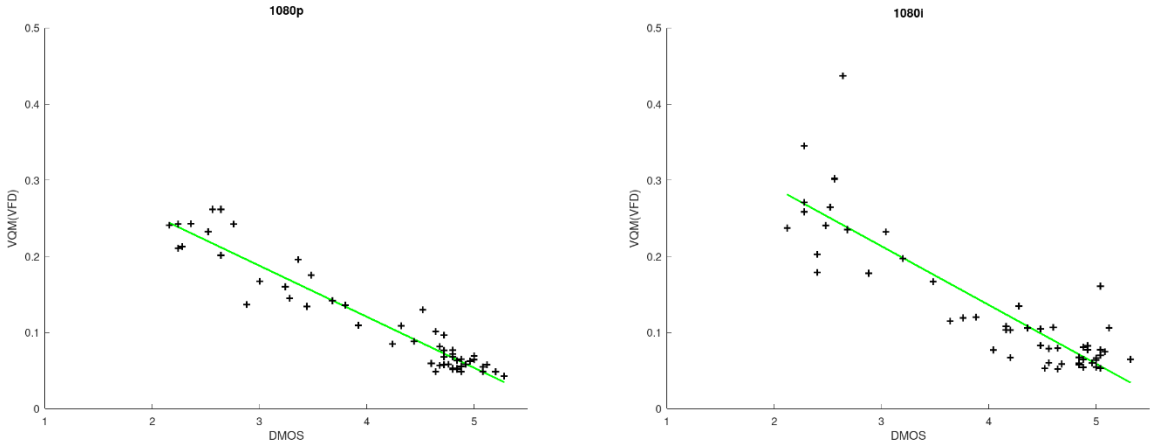


Figure 3: Scatterplots of VQM_VFD performance. 1080p is to the left and 1080i to the right.

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

4.3. Peak Signal to Noise Ratio (PSNR)

Peak Signal to Noise Ratio (PSNR) is the most commonly used video quality method. It is actually standardized by ITU in ITU-T Rec J.340[10].

An open source implementation is provided by ITS/NTIA (https://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx). The advantage with this is that it finds the temporal shift automatically, which another implementation may not do. A comparison between this implementation and another open source implementation (https://github.com/slhck/ffmpeg_quality_metric) was performed with very similar results, although in the latter case the temporal adjustments were performed manually. The latter gives the scores per frame and an average was calculated to get the overall score. The PCC is 0.99, see Figure 4. This implementation of PSNR needs external registration software, whereas the VQM implementation does not.

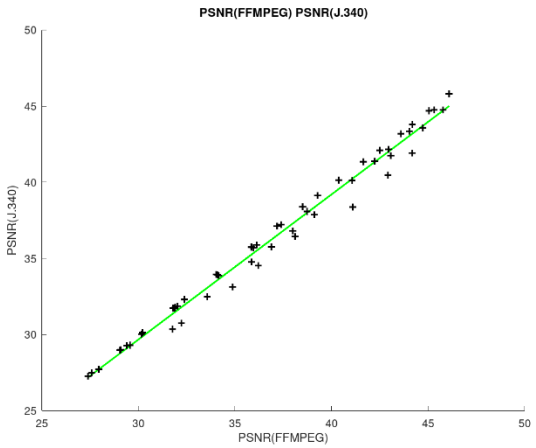


Figure 4:Comparison between two implementations of PSNR.

PSNR was evaluated at RISE and the PCC performance was for 1080p 0.71 and for 1080i 0.75, see Figure 5.

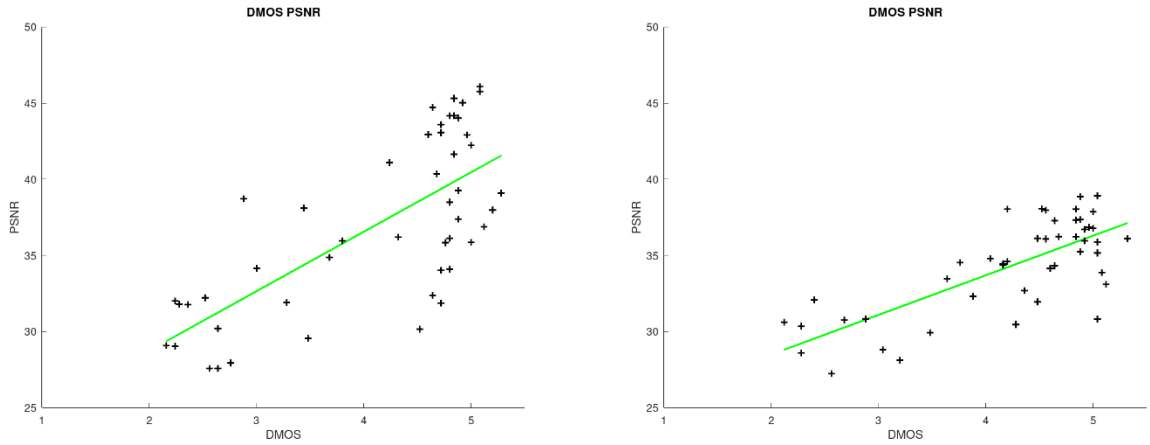


Figure 5: Scatterplots of PSNR performance. 1080p is to the left and 1080i to the right.

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

4.4. Structural Similarity Index (SSIM)

The Structural Similarity Index (SSIM) is an image quality methods that has been popular since it was first published[11]. 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 do not specify how it should be used for video. The paper[11] is one of the most cited papers in the image processing community.

SSIM was evaluated at RISE and the performance was not good for 1080p a PCC of 0.65 and good for 1080i 0.57, see Figure 6. It can be observed that the all SSIM values are very high and that they are in a very small range, except for a few outlier for 1080i. These values are for the HRC9 which is a very basic deinterlacing, which SSIM does not seem to handle well. This implementation of SSIM needs external registration software.

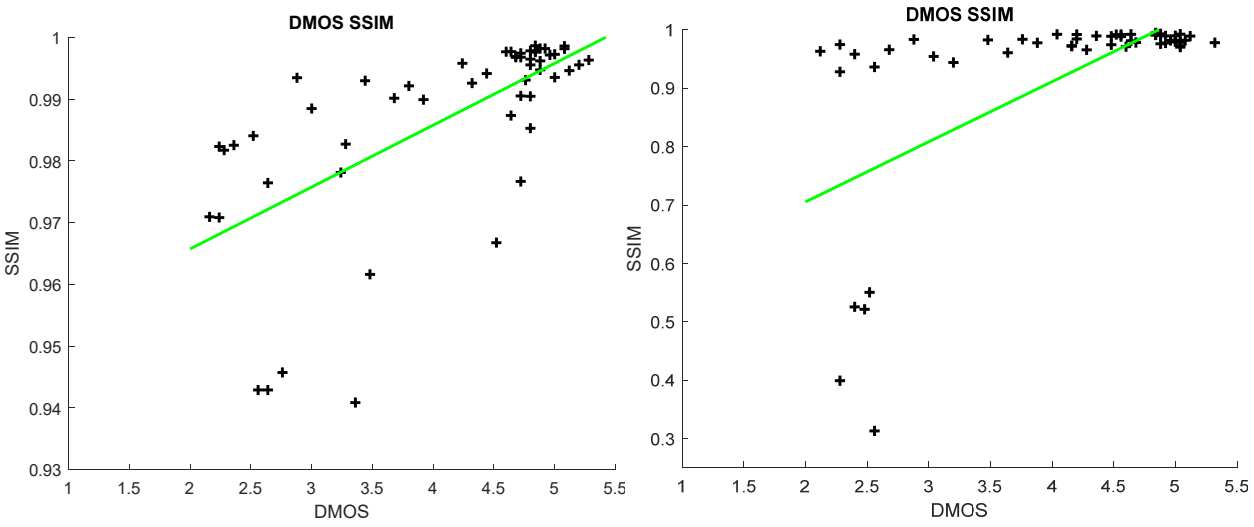


Figure 6: Scatterplots of SSIM performance. 1080p is to the left and 1080i to the right.

4.5. Conclusion

The performance of six different video quality models have been evaluated, that were identified during phase 1, against the subjective video quality database that was created during phase 3, see Table 1. The results are slightly different for 1080p compared to 1080i. For 1080p the models VQM_VFD and VMAF performs the best with Pearson Correlation Coefficients (PCC) above 0.9. For 1080i the PCC drops a bit overall and then VMAF and VQM_VFD are close in performance and performing the best. The overall performance for both formats VMAF and VQM_VFD stands out as the best models. In this comparison VQM_VFD has the added advantage to also be able to perform its own registration i.e. to fix any misalignment between the reference video and the distorted one.

Table 1: Summary of the Pearson Correlation Coefficients (PCC) for the different models that have been evaluated against the whole subjective video quality database. It shows the degree of linear relationships between the model scores and the differential mean opinion scores (DMOS). 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.

Model	1080p	1080i
VMAF	0.91	0.91
VQM General	-0.84	-0.85
VQM VFD	-0.96	-0.88
PSNR	0.71	0.75
SSIM	0.65	0.56

4.6. Plan for phase 5

The plan for phase 5 is to test ingesting known videos into VAR- or broadcast systems and grab these videos again. Extract the distorted ingested video automatically and then run a video quality analysis on the extracted video offline.

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

5. Glossary

Definition of some words used in this report. Some definition of words come from the International Football Association Board (IFAB)[1].

Absolute latency: The time difference between an event happens and when that event appears on a screen or in a video stream.

Assistant Video Assistant Referee (AVAR)[1] – pronounced A-V-A-R – usually a current or former referee appointed to assist the Video Assistant Referee especially to:

- watch the 'live' action when the VAR is undertaking a 'check' or a 'review'
- keep notes of incidents etc.
- communicate the outcome of a review to the broadcasters

Delay: When an event happens later than expected.

Frame synchronization: When frames arrive with perfect timing (i.e. with no time difference).

Full reference (FR) – access to a high quality or non-distorted version of the video to compare with the distorted video. Is put in front of objective video quality model or video quality model or used without if it could be understood from the context that it refers to these terms.

No reference (NR) – no access to the reference. Is put in front of objective video quality model or video quality model or used without if it could be understood from the context that it refers to these terms.

Objective video quality model: mathematical model that approximate results from subjective quality assessment, in which human observers are asked to rate the quality of a video.

Reduced reference (RR) – key quality parameters are computed from the reference and the distorted video, which are compared. Is put in front of objective video quality model or video quality model or used without if it could be understood from the context that it refers to these terms.

Relative latency: The time difference between the same event appears on two different screens or in two different video streams.

Replay operator (RO)[1] – person with technical knowledge who assists the VAR in the video operation room (VOR)

Time synchronization: When events appear with perfect timing (i.e. with no time difference).

Timing: The ability to make things happen at the right time. Or, the ability to make things happen at the same time.

Video Assistant Referee (VAR)[1] – pronounced V-A-R – a current or former referee appointed to assist the referee to correct a clear error in a match-changing situation (or if a serious incident is missed) by communicating information from replay footage

Video operation room (VOR) – the room/area where the VAR, AVAR and RO etc. view the match and have independent access to, and control of, the broadcaster's video replay footage. It may be in/near to the stadium or in a more central location (e.g. match centre)

[1]The VAR, AVAR, RO and RA must be neutral in respect to the competing clubs

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570

6. References

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

[2]. Brunnström, K. (2018). *Project plan: Assessment methods for Video assistant refereeing (VAR) System* (acr062142). RISE Research Institute of Sweden AB (Acreo), Kista, Sweden.

[3]. Brunnström, K., A. Djupsjöbacka, and B. Andrén. (2021). *Video quality based on a user study with video professionals for Video Assisted Refereeing (VAR) Systems* (RISE report 2021:29), DOI: 10.23699/79ja-gj68.

[4]. ITU-T. (2020). *Statistical analysis, evaluation and reporting guidelines of quality measurements* (ITU-T P.1401). International Telecommunication Union, Telecommunication standardization sector, Geneva, Switzerland.

[5]. Pinson, M. and S. Wolf, (2004). *A New Standardized Method for Objectively Measuring Video Quality.* IEEE Transactions on Broadcasting. **50**(3): p. 312-322.

[6]. Liu, T.-J., Y.-C. Lin, W. Lin, and C.C.J. Kuo, (2013). *Visual quality assessment: recent developments, coding applications and future trends.* APSIPA Transactions on Signal and Information Processing. **2**: p. e4, DOI: 10.1017/ATSIP.2013.5.

[7]. 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.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.

RISE report 2021:30
ISBN 978-91-89385-15-3
DOI: 10.23699/0x71-2570