Day 1, March 19th

Working Group Summaries\*

**AVHD** parametric & full reference models for audiovisual quality of adaptive bit-rate services. Validation test of models in cooperation with ITU-T SG12 Q14 is busy creating training databases for models, to be followed by validation databases. Each proponent creates 2 training datasets, total of 17 training datasets. Goal is Spring 2019 selection of winning models. Topic for discussion: planning for what to do after AVHD/P.NATS Phase II completes. What should future models be like?

**PsyPhyQA** Have identified some sequences and will start soon testing. Assessing video quality through physiological measurements like heart rate, EEG, etc. Test plan ready, and labs have agreed to run the test. Preliminary testing results to be presented after lunch.

**SAM** Processing subjective data to extract more data than is traditionally reported. Can we extract more information from the data that yields more information about subject opinions? Just started work in Krakow. Goal this meeting is to establish regular audio calls.

**NORM** New endeavor, seeks open collaboration for no reference metrics, open source. Have identified two use cases: broadcast and first responder.

**JEG-Hybrid** Collaborative endeavor, have large database of HEVC encoded videos, seek one Hybrid video quality metric. Looking for interesting patterns in big dataset to improve metric. Bi-weekly audio calls, announced to the ML. Anybody welcome to join.

**QART** Quality assessment for recognition and task based assessment; see web page for description of scope.

**IMG** (slides) Goals; baseline quality assessment for today’s systems using repurposed traditional content for virtual reality; new content captured specifically for virtual reality; subjective test methods, presentation requirements, QoE guidelines; virtual reality gaming. Have created a document describing issues to be addressed for 360 degree video. Seeking contributions for any of these issues. Main sessions start Tuesday, with SG12 Q13 documents, seeking feedback to improve two draft ITU Recs.

**UltraHD / HDR / WCG**  to be reorganized and merged into one group

**eLetter** publication of work, collecting knowledge on specific topics, not as rigorous as a journal publication but effort to make this a good collection of short articles around a specific topic. Last was immersive media.

**Email reflectors** some people are having distribution problems. Send information of problems to Margaret Pinson, particularly detailed rejection emails that can be used to diagnose these distribution problems.

**ITU** summaries presented. Q10 conferencing systems, especially audio of interest; Q9 speech quality, just finalized / updated two Recs, PESC for wideband and POLQA.

UHD, HDR, and WCG

Presentation #25: Zhu Li (Netflix):  VMAF metric for 4K video

The significant difference between UHD and upscaled HD (VMAF scores) was studied.

The model will be released as open source. New HDR model developed as well.

Discussion on group reorganization

Overlap with groups mentioned. Proposal to add high frame rate as fourth dimension and keep the topic.

4K and HDR is contemporary display technology. There are four axes to future display technologies: higher frame rate, higher resolution, higher dynamic range, wide color gamut.

Open issues mentioned: for instance which are the the main quality  influencing factors (tone-mapping in the display or source, etc.). Proposal to start with “ideal display” discussed - “setting the bar high” in terms of display used for the tests. Proposal to close the effort and reopen it once there is renewed interest. It is finally concluded that interest is still present, but the topics can be afforded in other groups.

Activity of the group (Patrick Le Callet)

Methodologies for QoE assessment, datasets, objective measures, extension of VQM to 10 bits.

Ongoing: selection of content for WCG, physiological measurements, HDR/WCG for 360.

VQEG is interested in future display technologies, including UltraHD, WCG and HDR. However, these endeavors are more of interest within the context of other working groups. Therefore, these working groups will be closed, and the topics absorbed into other working groups.

**Agreement was reached:** to close the UltraHD and HDR / WCG working groups

**Website update** to reflect that we seek presentations on research related to video quality, but not necessarily within the scope of any active VQEG work. (For example, UltraHD, HDR and WCG work falls within the scope of several other working groups.)

**Task**: Chairs of each working group need to update their website descriptions.

Presentation #17: Pablo Perez (Nokia Bell Labs): The role of video QoE in the Future X Network

Team presented: Distributed reality for human communication

Vision on the future of the network: disappearance of clear separation between mobile and fixed network, higher flexibility,etc. Data in the network: higher bandwidth requirements and delay requirements; VR content, data for autonomous driving (requiring very low latency -> towards 1ms).

Impact on Video Quality:

* Video information still relevant;
* Entertainment not only domain: training, design, communication expected to become more important);
* Latency critical;
* Quality as a Service (QoE parameters as part of service definition).

Projects:

* MOVISE - Monitoring segmented video - Architecture to monitor QoE - no reference metrics;
* Terminal-oriented video transcoding (pres. **#**28);
* Video adaptation for 360VR;
* Distributed storage for video (highly scalable);
* 5GSTB - implementation of client device for 5G mobility environments, adaptive streaming policies. Possible use metric developed in  AVHD;
* VINEDO: end to end value chain of professional 360VR VoD systems.

PsyPhyQA

Presentation #8: Sebastian Bosse: SSVEP  - based image quality assessment

Prediction of MOS from EEG data.  3 SRCs with block based image compression are considered. 6 stimulation frequencies (alternate presentation of original and compressed image), from 2 Hz to 10 Hz. Stimulation frequency has an influence on SNR.

Good SNR on second harmonic at 3 Hz. In general good SNR in the region 5 Hz - 7.5 Hz. This can be justified by considering the measurement  of noise.

Questions on possibility to extend it towards a no-reference metric and on using this for assessment of quality value rather than  checking difference with original.

AVHD

Presentation #20: Steve Göring, “QoE Evaluation of HTTP Adaptive Streaming – Open Dataset and Software for ITU-T P.1203”

Dataset and software for research purposes; P.1203 proponents raised concerns about licensing.

NORM

**Margaret Pinson:**

**First responder use case**

Audience: consumers & public safety practitioners

Consumer grade cameras in general.

Use cases:

* Camera capture, with compression not a major issue - impairments from camera
* Transmission & storage optimisation
* Artificial intelligence (video analytics)

Datasets its4s available and its4s2 in preparation. 1488 photos from Flickr, 41% public safety, 59% entertainment.

4-minute training session at the beginning and end (to check changes  after long test session).

Subject bias associated to content.

* Camera capture

Medical emergency considered but privacy issues.

Firefighters scenario with location based services.

**New Issue for use case**: body-worn cameras (police, etc.) may have a different perspective / do not give the same representation of the scene as appearing to the user. For instance an object may appear as a gun to an officer, who can respond, while the camera can clearly show it is not the case (or vice versa).

* Transmission and storage optimization
* First responder, Artificial Intelligence systems

Point raised on relevance of scenario to cope with attempts to  fool AI systems performing recognition. Changing a single pixel may prevent the recognition. Comment on useful reference information (Naeem - connected remotely  -  will send a useful link about it)

**Broadcast scenario**

KPIs:  macro blockiness, blurriness, ringing artifacts, packet losses, motion artifacts etc.

Two separate use cases for detection of errors and MOS were proposed in online meeting. Is it required / useful to separate compression errors from transmission ones?

More hierarchical approach proposed: processing errors before, then the other layers.

Player to be considered as well according to earlier discussion.

What is missing in the industry is a metric able to predict the QoE, for instance to compare encoders. Example: content with artificial blur was considered with same quality as with no blur according to an existing tool assessing quality NR. Comment: we do not need NR for the encoding part in this scenario. In a linear workflow, latency is a constraint. Comment on adaptation over time. Topic partially  covered by AVHD (adaptation aspects). Difference with AVHD can also be seen in the joint effect of compression artefacts and packet losses here. AVHD results in terms of  NR metrics: P.NATS Phase 1 did not propose a NR metric. The hybrid models proposed have a reasonable performance. Here focus on pixel only, not bitstream.

Video Content: Mezzanine feeds. Call for contributions on content.

Day 2, March 20th

ITU-T Q13 Session

Presentation #18, Rachel Huangyihong (Huawei), “Output from ITU-T SG12 Q13 Interim meeting”

Both G.QoE-VR and G.360-VR were progressed since the Krakow VQEG meeting. New baseline document for G.QoE-VR established (see VQEG meeting documents). G.360-VR updates include comparison of modified absolute category rating (M-ACR) (i.e., view twice before rating) and double stimulus impairment scale (DSIS) (reference video, then test video) for 360 degree video and QoE evaluation.

Rachel presented G.QoE-VR (document WD14-Restructured version of G.QoE-VR baseline) and G.360-VR (WD11-Proposed baseline for G.360-VR-v2).

VQEG reviewed the G.360-VR document, and provided these comments, which will be sent to SG12 in a liaison:

* Clause 6, Figure 1 should recognize that the video content impacts immersion.
* Consider whether, Figure 1 of G.360-VR is better suited to G.360-VR or G.QoE-VR .
* Clause 6 should note that people need to think about all of the factors identified in Figure 1, even if some of these factors will not be exercised by the experiment.
* Lower blue box of Figure 1, replace “audio sampling rate” with “audio bandwidth”
* Lower blue box of Figure 1, the audio and video bitrate only makes sense if you mention the codec being used.
* Figure 1, top level “presence” should seems instead be representing  “audiovisual coding quality for VR.” The document can then sub-specified that “presence” is an important factor..
* Figure 1 refers to “visual fidelity” and “acoustic fidelity” are terms that have been used in other fields with alternative definitions (e.g., comparing two images). It would be better to choose new terms that do not have other definitions, and move these terms (with their typical definitions) to a lower level of this diagram.
* The definition of “visual fidelity” implies that this Rec. only applies to natural content. Revise the definition to include synthetic content (computer graphic generated). “Visual naturalness” may be a better choice of term, but it would likewise imply that the document only applies to natural content.
* Difference between presence and immersion is well stated in the text. However, it is not clear why immersion is not mentioned in Figure 1.
* Perhaps another Figure should be created that describes QoE, to compliment Figure 1, that mentions “naturalness” and other factors that impact QoE.
* (Clause 7) Examples of “sports, drama, film, speech, music” seem less likely to be impactful than training, education, etc., but this can be left as-is since the point of the rest of that sentence is valid.
* Clause 7, There should be some stimuli that stress the system. Look for a recent description of the content influence on subjective experiments, to be sure the stimuli cover all important factors. For example, stimuli with a clear main interest in the center will not be used as 360-video, because people will not look around. Interesting cases will encourage people to look around. In the future, new results may indicate types of content that must be included in a test, so that the results are not limited.
* Clause 7, How to ensure appropriate recording (capture) quality is not trivial. This should be mentioned in the Rec., with advice for ensuring appropriate 360 degree video capture (e.g., type of content, resolution, whether there is a clear focus or no obvious direction to look).
* Clause 7, omit phrase “The quality of the SRCs should also be as equal as possible.” This will prohibit people from conducting experiments that compare different camera recording systems. Or at least modify this requirement, to allow such experiments. For example, the lens setups produce different artifacts that are relevant and interesting to compare.
* Clause 7.5 specifies an extremely wide range (10 sec to 5 min). It may be difficult to compare results for such different tests. It may be nice to advise people on preferred stimuli durations, so that experiments conducted by different labs are more likely to choose similar durations. Guidelines would be helpful on why different durations would be chosen (e.g., longer durations aid immersion).
* Have other labs conducted the M-ACR method? This is a new method. The traditional methods (like ACR, DSIS) have been conducted in multiple labs and compared between labs. It seems inappropriate to standardize a method that has not been checked against other labs. Independent validation of the M-ACR method is advisable before including it in the Rec.
* The suggestion to use longer sequences for subjects in the test for the first time seems difficult to implement. Pragmatically, how would this be accomplished? Does this mean naive subjects cannot be used, if the test focuses on 10 sec sequences?
* Repeating SRC with different HRCs will train people on the SRC. Subjects are unlikely to need as long to explore scenes they have seen before, and 10 sec is likely to be sufficient. Otherwise, 10 sec seems to short. 20 sec or 30 sec is more realistic, particularly for novel stimuli, so that subjects can explore the whole scene. It would be nice to insert advice on appropriate and inappropriate uses of 10 sec sequences (e.g., appropriate for evaluating coding artifacts).
* Mention other test methods in Clause 8. For example, it may be appropriate to use SAMVIQ, but this method has not been tried and validated for 360 degree video.
* Remove the mandated times from Figure 2 and Figure 3.  For example, the “max 20 seconds” voting time artificially limits the experiment. All voting time limits should be removed from the Rec. and left to the discretion of the person designing the experiment.
* It seems odd that the voting time for DSIS (6 sec in Figure 3) is shorter than the suggested voting time for M-ACR (20 sec in Figure 2). The 6 sec voting time is very short, perhaps too short, and regardless any suggested voting time should be the same.
* Clause 8, verbal voting should be allowed by the method. Voting time may need to be increased, and different subjects may need more time for verbal voting.
* [later comment] If there is a reason to believe that people might become ill (e.g., due to VR delay), then subjects should not be left alone. The test moderator should remain with subjects, due to concerns of sickness, and ensure subjects will halt the test when feeling sick, despite not finishing the session or the current interactive task (for task based VR tests).
* The above comments will be reviewed and be as as support for writing the Liaison statement back to ITU

IMG may schedule an offline audio call to further discuss these documents. If IMG has free time today, these documents may be discussed more later.

Action: Jesus will lead the team to draft the reply liaison to SG12.

Immersive Media Group (IMG)

Presentation #2: “Modeling Gaming QoE” by Saman Zadtootaghaj

Note that gaming includes people watching others play video games, which is increasingly popular. Delay is the main bottleneck.

Issues discussed: Can we mix videos from multiple games in a single test? The encoding is different, due to parts of the video that do not move at all. Are there any issues? Nearly static games (like card game), may quality metrics don’t work because the visual changes are small. There is not a “gaming only” codec. Usage rights on gaming videos is unclear.

Presentation #19: “Quality of Experience for a Virtual Reality simulator” by Kjell Brunnström

Truck driver uses VR system to load logging truck. This reduces crane cost, because don’t need to build a separate crane control cabin. VR simulator used to analyze system and training purposes. Added Joystick delays were not hard for subjects; trucks generally have such delays in the mechanical and hydraulic systems, and people could compensate for this. Added display delays were problematic; 10 of 39 subjects stopped early due to longer display delays. For other subjects, added display delays were not problematic; and their data masks the problems experienced by the people who became sick and dropped out of the test. Basically, there are strong differences among subjects in terms of simulator sickness response. Overall, low display delays are important for VR/AR systems (to avoid sickness).

Presentation #26: “Proposals for IMG work plan” by Pablo Pérez

Propose two phase approach:

1. Audiovisual quality of 360 VR video → establish methodology based on existing standards
2. Full QoE of immersive experience → based on the findings of phase 1

Action: Jesus to specify changes to VQEG IMG website; Margaret to update IMG website to link to the IMG documents.

**Discussion:**

Crucial point is use case of 360 degree videos. It probably won’t be watching movies, where VR may be more annoying than helpful. This will strongly influence everything else. Response: “movie” may be a simple content to work through subjective test issues.

NTIA/ITS may make some interactive VR content available, on first responder subject matter.

Perhaps add M-ACR to the list of methods of interest (modified ACR, displaying each video twice before rating on the ACR scale).

Content may be available from Google and YouTube, but such discussions have not yet produced content.

Potential use case for 360 degree video is remote viewing: unmanned ground vehicles, unmanned aerial vehicles, remote driving, automated warehouses, videoconferencing, …

Interactive applications may be more interesting.

**Interest in participation:**

Kjell (Acreo), Lucjan (AGH), Patrick (Nantes), Francesca (CWI), Pablo (Nokia), Narciso (UPM)

Action: send reply liaison to SG12, expressing VQEG’s interest in conducting tests that explore issues around VR subjective test method.

Presentation #28: “Spherical Structural Similarity Index for Objective Omnidirectional Video Quality Assessment” by Zhenzhong Chen, Grace Zhang

Presented a mapping function to transform image that stretches the VR projection onto a sphere. This correction factor modifies a 2D objective model for the spatial distortions that occur when converting 360 degree video to a VR / AR headset display. Mapping function was applied to SSIM. This framework can be easily generalized for various projection methods.

Presentation #30: “Salient360: Visual attention for 360 content” by Jesús Gutiérrez, Patrick Le Callet

This presentation was an update of research presented in the last two face-to-face meeting. The presentation has many links to work available elsewhere (datasets, papers, saliency map software, …)

Presentation #5: “Quality of Experience for a Virtual Reality simulator” by Shahid Mahmood Satti

Presentation: “Subjective Test Encoding” by Shahid Mahmood Satti

Problem: The subjective quality of two differently encoded videos should look the same, but is it actually the same? Pair comparison test (A/B test) with “same” option; compared to split screen (left is better / right is better / quality is the same).

Presentation #3: “Broadcast and Creating Use Case for Non-Reference” by Jeff Webb (remote)

Sky UK: why moving video workflow trom appliances to containers; what are the challenges of modern broadcast environment; why industry needs NR video quality tools. Change in workflow to ad-hoc streaming channels, started and stopped quickly.

Broadcast use cases for NR models

* IP contribution (input video source, SDI and mezzanine feeds, detection problem more than quality problem / high quality expected, quality is “acceptable vs excellent”)
* Distribution feeds (adaptive bitrate ladder, what customer receives)
* Streaming clients (final validation of QoE as seen by customer, client side video quality linked back to broadcaster’s internal system, validation, maybe partial-reference metric)

See slides for more guidance on these broadcast NR use cases.

Question: why use NR for distribution feeds? FR models (e.g., Netflix uses) need comparable signals; sometimes these distribution feeds cannot be compared (e.g., may not have reference video, because feed was already compressed; streaming a live event so don’t have time to do FR comparisons, cannot afford the extra latency; may be 60 sec off from live, and FR metric latency may be objectionable; incoming quality from IP contribution may be too low for a FR metric).

IP contribution, less control over quality, may want to check / ensure quality. May receive lower quality feeds, for example from overseas.

Discussion on whether FR metric can accommodate distribution feed quality evaluation? Accuracy of FR metric is higher. Some broadcast workflow does not accommodate FR metric.

Question: why is the extra latency so objectionable?

3rd use case (streaming client) may also be suited for a FR, RR, bitstream, or hybrid-bitstream model.  However, may want to model problems in cheap customer player problems, like a cheap android smartphone with CPU limitations.

Immersive Media Group Discussion on Collaboration

There is interest in the ITU efforts to establish Recommendations for 360 degree video, and in particular contributing to those draft Recs. Continue offline discussions on the documents not addressed at this face-to-face meeting. Also interest in researching how to modify the methods for 360 degree video. Reference content (e.g., from existing databases) of interest, to create common databases in the future.

Qualinet VR has not progressed very much. There is a list of interested organizations that may have data to contribute; we could reach out to them. Christian Timmerer has a list of relevant documents.

Use case for VR interaction with the real world is the most likely use case (e.g., training, operating remote machinery). That should be brought into the discussion of IMG. Interactive VR testing may attract more participation.

Action: write use case documents for 360 degree video (similar to NORM use case documents). Pablo

Delay / interactivity and compression is more interest generally than some other issues, like projection. Must rethink quality factors to evaluate (e.g., depth perception likely not as interesting as ability to complete the task). Expected outputs, want to identify sequences and create subjective methods that can be contributed to ITU Recs. as a starting point. Subjective assessment more interesting than objective metrics right now. There are too many open questions on subjective assessment; objective metrics difficult to develop until those are answered.

SG12 reply liaison author list:  Pablo, Jesus, Francesca, Kjell, Narciso.

Day 3, March 21st

SAM

Goal is to extract "more" from subjective test results.

Presentation #24: Status update and Introduction by Ioannis Katsavounidis

* Email reflector was setup
* Interest from 20 people
* No regular conference calls yet, plan to establish this now
* Need to formally elect chairs

Repetition of the Concept of MLE (Maximum Likelihood Estimation of Subjective Parameters)

 A presentation with an introduction to the method:

ftp://vqeg.its.bldrdoc.gov/Documents/VQEG\_Krakow\_Nov17/MeetingFiles/VQEG\_General\_2017\_205\_\_Zhi\_Li.pdf

A paper with details:<https://arxiv.org/pdf/1611.01715.pdf>

* Method which tries to take the inconsistency and bias between subjects, as well as errors due to content into account when analyzing votes from subjective experiments. This is achieved by modelling each individual vote as a combination of OS (Opinion Score) plus Gaussian distributions of per subject inconsistency and content related ambiguities.
* Results allow subtraction of per subject bias before adding MOS values.
* Same for subject inconsistency and content ambiguity
* This result is pool of "average" subjects with no bias between them. The constant bias is taken into account later.
* It is assumed that the constant  contribution of the subject inconsistency and the content ambiguity on the final OS is equal. Alternatively, the minimum content ambiguity is required to be very small (almost 0). The latter may be required for references contained in the subjective experiment.

Algorithms

Requires partial second order derivatives (Hessian Matrix), which requires taking care of potential singularities!

Different methods to solve the linear equations are possible, either using full Hessian Matrix or making assumptions on diagonality of the Hessian.

Alternatively first order iterative methods can be used or projection based iterations.

Discussions Around Behavior of MLE

* with non-Gaussian distributions of votes
* biases which are inaccurate due to MOS scale limitations (e.g. truncation at MOS=1 and MOS=5)

Questions raised about assumed linearity of the scale (MOS scale saturates at the bottom and end), and how can the algorithm be refined to take into account this?

Further questions about how tests should be designed differently to deal with these limitations (e.g. extending the scale).

Presentation of some results calculated by Univ. Nantes

* The number of subjects can be reduced by up to 30% (which is still debated)
* Full matrix of votes is no required (not all subjects need to see all videos)

Presentation #27: Device Characterization for Conditional Encoding

By Juan Casal, Narciso Garcia, Nokia and UPM

Goal is to save bitrate since this is an important cost factor in networks

Nokia working title: "Conditional encoding"

Technical Challenges: Per title or per scene encoding for ABR systems and extraction of video complexity features by IA techniques.

Result: Optimized encoding saves up to 30% bitrate

Research challenges: Which resolution should be targeted for which device / screen size

Content selection for subjective testing was based on filling the space between temporal and spatial complexity.

Devices used: LG-P720, Samsung A3, Samsung Galaxy S7, iPad Air 2

Tests were conducted including audio (using ear buds) and scores were for audio-visual quality

Users could decide on comfortable viewing distance, most users preferred device on the table, some preferred had-held.

50 naive observers (avg. 22 years)

Result charts -> Presentation

Curiosity: Female observers consistently scored lower than male observers.

Viewer comments:

* Knowing best/worst quality beforehand was very helpful
* Low spatial complexity content was difficult to rate
* …

Future work:

Random split in two subsets

Random order of devices

Note: Viewers knew the device type and this could have impacted their expectation and therefore scoring!

**Demo:** Presentation of realtime conditional encoding

Presentation #23: "Quantifying VMAF Model ~~Uncertainty~~ Variability using Bootstrapping"

By Zhi Li, Ioannis Katsavounidis, Netflix

Uncertainties for model training are stemming from various sources (mostly subjective data and SVM predictor)

Presentation focuses on uncertainty caused by SVM

Bootstrapping - "Resampling with Replacement"

Bootstrapping is in principle an optimized training method which is based on systematically replacing data from the training set (“bagging”)and iteratively repeating the training, while trying to minimize the variability seen within a "population" of models which are using slightly different parameters. The main goal is to calculate confidence intervals for e.g. perceptual models.

Applying this method to perceptual metrics helps to identify the uncertainty of model predictions

Organisational stuff

Parties interested to continue this topic:

IST, RISE, Univ. Nantes, AGH, OPTICOM, UPM, Kingston Univ., HHI Fraunhofer

Proposal for an Objective of SAM

Ioannis presents some proposal for modification of ITU-R BT.500, where MLE is added as an Annex. This is by no means meant to be final and could even be started from scratch or moved to a different Rec. To be discussed later.

The proposed text would also nicely fit into P.1401 (statistical analysis methods for the performance of objective models).

Organization

Decision: We wait to see if November works for Google to host the meeting. The final host and possible meeting dates will be decided quickly; and a survey conducted on the main VQEG reflector.

Decision: Pablo Perez was chosen to be vice-chair of IMG group.

Proposal: Patrick Le Callet proposed to setup a new group Human Factors for Visual Experiences HFVE. This activity will be with cooperation with IEEE Standards / Computer Society. At the next VQEG meeting we will have a session dedicated to setup the connections and synchronized.<https://standards.ieee.org/develop/wg/HFVE.html>

New support group: led by Maria of Kingston University, try to coordinate a remote session of IEEE HFVE within the next VQEG meeting, which would take 2-3 hours.

Decision: HFVE will be be created with Maria Martini as the chair.

Proposal: Patrick Le Callet proposed to create CGI Computer Generated Graphic Contents. It is not clear if animation and computer games should be considered separately.

Proposal: ask that new groups are proposed with a clear objective and a plan of how to reach that objective.  E.g., CGI = quality assessment of CGI content

For CGI group the starting topic could be establish subjective method for video game quality assessment.

Decision: CGI (Computer Generated Imagery) group will be created, the chair is Saman Zadtootaghaj (Deutsche Telekom)

Decision: replace Christian Schmidmer with Shahid Satti as AVHD Chair.

Tentative decision: When suggesting a new groups it should provided a name of the group, a chair person and a problem description.

[Note: while approving minutes on day 4, the above decision was discussed again, the wording slightly modified, and the decision confirmed.]

NORM

Presentation #14 European Research Project on Video Subjective Quality on massive deployments

Antonio Cuadra & Jaime Ruiz

Question: How do you know which metric is the most important to QoE.

Answer: There is no general rule. For each specific use case the scope has to be defined.

Comment: How can I use the knowledge about MOS of my users? Linking to revenue is much more important.

Comment: How to link MOS to user behavior is very tricky and difficult.

Comment: The work from Akamai showed link between QoE and watching time.

Comment: If bitrate is increased people watch it longer.

Question: Does this tool work with encrypted content?

Answer: The metric is at the end node so it can use webbrowser data.

Question: Do you analyze pictures?

Answer: No, we are analyzing only html file.

Question: You have plugin. How you do it for devices like TV?

Answer: It works the same as long as we can access the application.

Question: How do you classify particular session?

Answer: We are users, it is not running on a user device, those are external probes just to measure it.

Question: Can the tool be available for others?

Answer: Some tools are open source, the whole system cannot be shared.

Comments: The tool will be added to the NORM list of tools.

Presentation #16 BRISQUE and NIQE Blind Quality Metrics

Todd Goodall, Alan C. Bovik

Question: Your definition of naturalness, how does it work on artificial images?

Answer: For animation it works very well.

Comment: On gaming it works well since it was validated.

Comment: We are creating list of available resources. Could you please send the tool.

Question: Can this method be extended to video?

Answer: Yes! You have to think about temporal aggregation.

Question: What is the computational complexity?

Answer: 20-30 fps in a python script. It should be easy to make it faster.

Presentation #15 VIDSPECT: A System for Evaluating Artifacts in Digital Masters

Todd Goodall, Alan C. Bovik

Question: Have you tried sequences with numerous distortions at once?

Answer: No.

Comment: The problem is important to many use cases including first responders.

Comment: Extending it to a different dataset is very interesting idea.

Comment: Youtube can help with providing some data, but you will not have reference and the sequences will be with bad quality mostly.

Comment: All the tools provided by University of Texas at Austin are available to use, free of charge.

Presentation #4 UGC Video Quality Challenges

Balu Adsumilli

Question: Is it common to make a preprocessing?

Answer: We do not make preprocessing, just understanding the format and codec. This information is used to better transcode. The noicing can be used but not for all videos.

Comment: Can we change FR metrics such that they will work with the difficult problem of comparing compressed sequence with bad quality source.

Question: How do you decide if a video have to be denoise?

Answer: Device dependent noise coming from information about device.

 Decision: SAM chair: Lucjan Janovski vice-chair: Zhi Li, Ioannis Katsavounidis, Patrick La Callet

[Note: while approving minutes on day 4, the above decision was discussed and the roles slightly changed, as discussed immediately after day 3 closed.]

Day 4, March 22nd

Thursday 22.03.2018 VQEG Madrid
09:30AM: minutes of the meeting approved
09:30AM: Lucjan started QART session

Talk 1: Lucjan "Video summary evaluation"
-> Conclusion: The results for algorithm marking the frames for summary and human expert marking the frames are not well correlated, something like 30%, just a bit better than the
random chance.

09:50AM: QART session closed

09:50AM: Lucjan started NORM session

Talk 2: Lucjan "Lip-sync detection"

-> Conclusions/discussions:
1. overview of available publication. in general there is no agreement on the how much audio/video delay is acceptable and beyond which threshold it is annoying.
2. lip-sync issue continues to be problem, for the broadcasters they receive video and audio on different cast.
3. In general it is referred to as "AV sync" instead of "lip-sync".
4. The algorithm forms a signal from lip movement after lip detection and correlates it with the speech signal. The results are still to be processed.

Talk 3: Jakub Nawala "How to compress like industry does", presented by Lucjan
-> why the industry captured sequences are blurry and lab coded sequences are blocky. The answer is to change the encoded resolution.

Talk 4: "A logo based approach for visual quality evaluation in telemedicine applications" presented by Maria
-> Concerto EU project overview
-> A logo which represent the type of video (ultra sound etc) would be used as a Reduced reference feature to evaluate the quality of the video. On the receiver side
the quality is measured as a FR metric as logo is known. The video quality of the logo correlates highly with the quality of the actual video using MSSIM and PSNR.

Decision: Enrico will replace Lucjan as JEG-Hybrid group vice-chair

Talk 5: Enrico Masala "Large Scale Database: new Content and Pooling Strategies"
-> 3 new sources of sports content contributed by Sky
-> Sigmodial fitting from different metrics to 1-5 scale
-> it is to be checked if the sky golf sequences is in slow motion
-> sports content does not behave too differently, however, confetti sequence deviates significantly due to difficulty of encoding.
-> SSIM is sensitive to noise of confetti like objects in the video
-> harmonic mean yields quite variable results. Zhi Li: This could be due to the nature of harmonic mean as it gives more pessimistic results even if a few scores are low.
-> it was checked and the golf video should not be playing slow motion.

Talk 6: Steve Göring "DeViQ - A deep no reference video quality model"
-> The video image is divided into subimages and a pretrained DNN is used to make classification and create the feature set, feature selection and random forest is used to
create the per frame quality score. For the random forest training the VMAF scores are used. Comparison with BQRISQU and NIQE show good quality of prediction.

Thursday 22.03.2018 VQEG Madrid  afternoon

Administrative

**Decision**: at least have 1 vice-chair for each group, at least 1 vice-chair from an organization different from the one of the chair

Patrick: VQEG could take position about quality for computer vision applications. Possibly QART group.

Lucjan: agrees. Important to have a methodology/framework for systematic test.

Margaret: agrees.

Pablo: interested.

**Decision**: QART interim vice-chair: Pablo Perez

Lucjan: NR metrics for computer vision would probably be used differently. Therefore it is better to keep this activity in QART and not NORM.

Narciso: interested in the group.

Presentation #13  Objective and subjective quality assessment for gaming videos

Nabajeet Barman

Q: settings VP9? Lack of documentation. deadline=realtime quality=realtime

A comparative Quality Assessment Study for Gaming and non-Gaming Videos

Nabajeet Barman

SSIM for gaming videos quite bad

VMAF pretty good also for gaming videos

An evaluation of video quality assessment metrics for gaming video streaming

Nabajeet Barman

Q: quality 480p with reference at 1080? Yes, upscaled at 1080p

Database will be made available

Progress report

By Nabajeet Barman

Decided to have at previous meeting

Would be nice if presenters could provide an abstract for the next report and maybe a list of publications. Maybe using google doc.

Kjell: it is probably good enough and it does not need refinement. Put effort in next meeting.

Presentation #1 Content-aware PSNR and its applications in adaptive video coding for video streaming

Benhur Ortiz Jaramillo

Proposes CPSNR method (published in journal paper, available in Matlab code), used to tune coder parameters (e.g. bitrate), 10-37% bitrate savings with same PVQ

<https://telin.ugent.be/~bortiz/>

Presentation NIST - Information Technology Laboratory - Video Analytics Research

Jim Horan (remote)

Video Quality Metrics for Public Safety - substantial needs in video analytics

Existing video quality metrics not suitable for video analytics

Other business

**Decision**: Silvio as vice-chair of AVHD group

Mikolaj (remote): QART name was already changed in the past to include computer vision applications.

**Current:**

Name: Quality Assessment for Recognition and Task-based multimedia applications (QART)

Mission: To study the quality of video used for recognition tasks and task-based multimedia applications.

Lucjan: focus also on “automatic systems”

Patrick: suggests to include “Computer vision” in the title: Quality Assessment for Computer Vision Applications (QACoViA)

Kjell: Quality Assessment for Human and Computer Vision Applications (QAHCoViA)

Patrick and Mikolaj agree “human” is not needed, it is already included

Decision: Change the name of QART into Quality Assessment for Computer Vision Applications (QACoViA)

Action: Mikolaj and Patrick: will elaborate a description for QACoViA (former QART)

**SAM**

Decision: SAM Bi-weekly meetings (calls) on Monday 5pm CET/CEST (Paris time)

First meeting 16th April 2018, 5pm Paris time, by Hangout, 1 hour duration.

Keep schedule fixed, but cancel the meeting in case there is no significant update. People needs to enter prospective talks in shared google doc, a decision will be taken before each meeting.

Day 5

A reply liaison was written from VQEG IMG to ITU-T Study Group 12. This liaison will be sent later, by the VQEG co-chairs.

Close of meeting.