VQEG Meeting Minutes

Berlin March 2019

Day 1 – Mon, March 4

Group Reports

**Audiovisual HD**: 1st part of project is adaptive streaming metric validation, a joint with ITU-T Study Group 12, in validation phase, models submitted, designing databases. Status update with more details on Tuesday. Regular conference calls, each Thursday. 2nd part of project is investigation of advanced subjective methods. Not much activity in that area.

**PsyPhyQA**: Psycho-physiological quality assessment, test plan mostly ready, soon to conduct experiments. Seeking other participants for 3 to 6 month range, to conduct experiments.Test plan available on Google Drive.

**eLetter**:  Next VQEG eLetter will be based on PsyPhyQA research; there is space for more articles.

**SAM**: Statistical analysis methods, looking into improved ways to process subjective data, including a standard method to store subjective data. JSON format selected, with a database back end. Important results: effect of discretization, and effect of saturation. Regular audio calls occurring (bi-weekly).

**CGI**: computer generated content, most of interest now from gaming. Most activity from Deutsche Telekom, Kingston University, Technische Universitat Berlin, and Simula, but many others interested. Subjective testing method, modeling gaming QoE, gaming video quality metrics (including machine learning), and gaming datasets created. One already out, two others available soon. New dataset will have 4K video and high frame rates.

**NORM**: no-reference metrics. Application focus: (1) entertainment production networks (e.g., broadcast center like Sky and YouTube). (2) camera capture & network optimization. Group does not have critical mass for regular audio calls; may explore joining with other group.

**JEG-Hybrid**: no one here to give update

**QACoViA**: Metrics targeted for machines instead of humans. New group with not many contributions yet. AGH will soon start a new project to build a model for QACoViA.

**5GKPI**: framework collaboration with the 5G world, who want information from use cases like video. No formal plan yet. Seeking interested parties to contribute.

**IMG**: Quality assessment for immersive media, 3D, VR, AR and 360 degree video. Started work on a test plan, especially for 360 degree video. Some discussions on this test plan at last meeting, and audio calls conducted to refine these proposals.

**Independent Lab Group (ILG) and Joint Effort Group (JEG)** are historical support groups, not much needed any more.

**HFVE and Joint-Qualinet-VQEG team** are liaison groups

**Tools and Subjective Lab Setup** keeps a list of open source resources provided by different organizations.

**VQEG Administrative and Web Support**: email reflectors continue to be problematic

**MPEG:** US delegation initiative to merge MPEG and JPEG, to form a unified group.

Presentation #120, “Input to ICDM's IDMS ver 2 chapt 4 - visual assessment” by Kjell Brunnström.

500+ pages, display measurement standard, working on version 2.0. ICDM chair attended fall VQEG meeting requesting input, Scott Daly (Dolby) and Kjell Brunnström (RISE) wrote input on DSIS and PC methods. They have draft text and would welcome other people contributing to this chapter. Contact Kjell if interested.

Presentation #105, “Are people pixel-peeping 360° videos?” by Stephan Fremerey

Presentation #106, “Impacts of internal HMD Playback Processing on Subjective Quality Perception” by Stephan Fremerey

Use 90fps if possible; use whirligig or virtual desktop or another 360 degree; avoid 25 fps playout. Large increase in quality changing from 25fps, 30fps or 60fps to motion interpolated 90fps.

Presentation #107, “Subjective Quality Evaluation of Tile-based Streaming for Omnidirectional Videos” by Ashutosh Singla

Up to 47 ms, found minimal quality impact of delay

Presentation #108, “Assessing Media QoE, Simulator Sickness and Presence for Omnidirectional Videos with Different Test Protocols” by Alexander Raake

Comparing short and long questionnaires for simulator sickness and presence. Short simulator sickness questionnaire worked well as long as the subjects had not seen the long questionnaire, to distinguish low vs high simulator sickness. But not to evaluate the impact of different factors. Short presence questionnaire has less detail but could replace the long presence questionnaire.

Presentation #123, “Quality metrics for immersive 360VR content” by Narciso García

Presentation #128, “Complexity measurement and characterization of 360-degree content” by Jesús Gutiérrez

Presentation #127, “Quality assessment of FTV” by Patrick Le Callet

Interpolate new (virtual) viewpoints from free viewpoint cameras; these photos contain new artifacts and the quality cannot be predicted by traditional metrics (e.g., PSNR, SSIM, VIF). Mostly geometric distortions. CSC-NRM metric developed for this application.

Some video coders produce similar artifacts (e.g., VP9, AV). Databases focused on H.264 may miss such artifacts.

Presentation #104, “Analyze And Predict the Perceptibility of UHD Video Contents” by Steve Göring

Comparison of striped/temporally changing test methods for finding difference between HD and UHD, some contents have a high classification rate (> 80%), others not (only slightly above 50%).

Discussion about the usefulness or applicability of time-switching method since it may not be realistic. Depends on the situation though. Questions about what would be a more realistic test (e.g. UHD@1.5H vs. HD@3H), or how the labeling itself may impact the results.

Presentation #110, “AvrateNG demo and short overview” by Steve Göring

<https://github.com/Telecommunication-Telemedia-Assessment/avrateNG>

Presentation #111, “AVTrack360 demo and short overview” by Stephan Fremerey

<https://github.com/Telecommunication-Telemedia-Assessment/AVTrack360>

Day 2 – Tuesday, March 5

Audiovisual HD & P.NATS Phase 2 Overview

Working on three model types:  Parametric, Pixel & Hybrid models.  Developing long and short databases for training and validation of methods.  Short sequences for video quality training and long sequences for delivery KPIs.

Work split into four task forces.  Current status:  35 models have been created.

Topics for breakout sessions:  Review and sharing of sources.  Design of validation test matrix.  HRCs for long-sequences.  Possibility of opening P.NATS2 DB for a larger group.  Verification of processing chain for CRF encoding.

Looking for participation, particularly 4K/UHD footage that can be used in subjective tests!   Conference calls: Thursdays 13h - 14h CET.

Registration information:  <https://www.its.bldrdoc.gov/vqeg/projects/audiovisual-hd.aspx>

NORM Presentations

Presentation #124, “Final Version of Lip Sync Quality Indicator” by AGH

Developing a no-reference method for lip-sync detection using comparison of the audio and the lip motion detected using facial recognition.

Slide 11 where the audio was shifted had several questions around the similarity of the plots.

Presentation #121, “NR Metrics: Alternate Training Data” by ITS

Introduction to the content.  Discussion on Thursday morning.

USC JND database: “VideoSet” - <https://ieee-dataport.org/documents/videoset>

5G-KPI Presentations

Presentation #112, “Use cases for 5G networks: video in mobility scenarios” by Nokia

Reviewed applications for video in 5G including entertainment in a mobile environment, connected autonomous cars and gaming (VR).

Presentation #119, “Use cases for 5G networks: video in industrial scenarios” by RISE

Reviewed applications for video in 5G for industrial applications.  Looking at the QoE requirement for VR video for controlling machinery.  KPIs and QoS for ML.  Low level KPIs (network metrics).

Interest in this topic:  Opticom, Nokia, RISE, Spirent, UWS, HHI, Keystone.

QACoViA session

**Co-chair: Mikolaj and Pablo**

No activity is going on however further information will be available in next meeting.

Is ground truth available or what is the method will be used for evaluation

Call for expression/join group is open

There is a good discussion on the different aspect of project.

Margaret shows interest to be part of this project.

Session Closed

Presentation: Infrastructure for High-Attendance, Simple Psychophysical Experiments by Mikolaj

Call to run subjective evaluation experiment in AGH

VQEG eLetter

Call for special Issue paper will be

on the website.

Each chapter will be appeared separately along with consolidated issue to get indexed by google scholar etc.

eLetter will include articles on regular basis in addition to special issue

Associate Editor list will be refreshed.

Tea Break

PsyPhyQA Session

**Co-chairs: Sebastian Bosse, Naeem Ramzan**

**1st Presentation:** Towards objective metrics via electrophysiology in visual media quality assessment by JP Tauscher from Technical University of Braunschweig

**2nd Presentation:** Linking subjective affect and neurophysiology in 3D movies and immersive VR by Michael Gaebler from MPI CBS Leipzig

End Session

ADM/email Reflector

Individual has to check with their IT department if there is particular spam filter

The other options are to use Slack/Google groups

Chris can check if he can setup reflector in his company

Day 3 – Wednesday, March 6

**G.QoE-VR**

draft available, original plan was to seek consent at the May meeting. After discussions, we will  seek review and comment from VQEG; identification of VR service use case, categorization factors that influence VR QoE. Halted to first consider following the idea of the collaborative “Survey Paper”, see below.

**P.360-VR**

Draft in TD-GEN-06-09, plan is to consent November 2019 meeting but may need to reschedule, Subjective evaluation of 360 degree video with head-mounted displays, enabling 3-DoF interaction exploration of content using head rotation.

**Survey Paper Proposal**

Discussed ideas to encourage academic participation in standardization efforts of Study Group 12. Idea is to start with a joint paper (between Study Group 12 and VQEG) that conducts a survey of the topic (e.g., QoE for VR), and then use this as a basis for a draft new Recommend. SG12 is also discussing Study Group 9 approach of identifying contributors to Recommendations. IEEE Access would be a suitable venue, or Springer Quality of Experience Journal.

IMG

Decided in July call: 360 degree video for joint work

Quality assessment for short 360 videos

* SRC: review dataset and get more content (at least 10 SRCs)
* At least 4K res, at least 30s, uncompressed/high quality content
* Possible options: 3 videos from Nokia, 3 videos from TU Ilmenau, Vsense, ImmersiaTV

-     Testing interfaces: same tool to run tests in all labs; possible options: AVTrack360, tool from Nokia,UdN

-     HRCs: coding (uniform / non-uniform e.g. tile based)

-     Other aspects: effect of duration of sequences, methodology (ACR, DCR), display, effect of audio.

Labs: UdN, Nokia, UPM, TU Ilmenau, CWI, Roma3, RISE, Ghent; other volunteers are welcome. Small common set and the whole test set will be split among labs.

It is expected that tests will be run in May 2019.

Discussion:

Tiling was mentioned as an issue - it is clarified that static tiling will be considered. There is interest in 180 content from industry.

Simulator sickness

The test plan is described, including division of sequences in active and rest periods, with assessment before and after active periods.

Each time the user provides a score on 1-5 scale on a single question (a couple of them will be selected and randomly assigned to different subjects). One of the aims is to study the correlation of SSQ results with single-question results.

User monitoring (via gaze position, eye tracking, other physiological measurements) is optional.

IMG workplan for long sequences

2-10 minutes sequences are considered (full content item).

The test methodology will consider in-sequence (SSCQE, SSDQE) and post-sequence evaluation (ACR, presence, simulator sickness).The latter will be done with and without HMD.

The same test session structure will be used for all users and labs - only methodologies may vary. Four types of questionnaires are considered for post-assessment (Slater-Usoh-Steed - but very long questions, MEC spatial presence, Spatial-Presence Experience Scale, ad-hoc questions). A small pre-test was run. Lessons learned  from pre-tests: the SSCQE slider can be problematic; some “pristine sequences” may be needed for in-sequence testing, although not engaging.

Labs currently involved: Nokia Bell labs. Others volunteers are welcome.

During the meeting TU Ilmenau and Univ. Nantes volunteered; Univ of Madrid, AGH and RISE are potentially interested.

Also in this case there is industry interest in 180 content.The potential to standardise this will be checked with ITU.

Future Meetings

Decision: Fall 2019 meeting will be October 14-18 in China

Decision: Spring 2020 VQEG meeting will be in Seattle, USA, hosted by Amazon

CGI

Investigation of perceptual video quality of dimensions (Falk)

Coordinate representing quality of feature values. Test methodology: PC with MDS. SD with PCA. Resulting in 5-D space (fragmentation, unclearness, discontinuity, noisiness, suboptimal lumenance). Rate scale: 7 point (extremely bad, bad, … excellent, ideal). HRC: (blockines, blurring, jerkiness, bitrate, ….packet loss 0.5%, 1.5%) Material: H+S video, others (including CG).

Observation from results: MOS scores vs. conditions on FRA, NOI, DIC, UCL, LUM scales. Quality prediction: linear combination of different dimensions: correlation pearson 0.90 with MOS.

Contribution of different dimensions is quite stable contributing to the final rating.

Conclusion: perceptual video space consists of 5 dimensions. Can assess directly via corresponding quality dimensions. Independent of video content. Possible to model overall video quality model from dimension models.

Towards the impact of spatial and temporal video quality on gaming quality of experience (Steven Schmidt)

Cloud gaming: game server, video streamed to user.

Working on: guidelines on acceptance…

Game QoE: hedonic … Pragmatic. First layer: aesthetics, novelty, interaction quality, playing quality. Second layer:

ITU recommendations: ITU-T SG-12. G.1032 (G.QoE-gaming), P809 (P.GAME). Future: G.OMG (this talk)

Study design: 85 conditions (5 bitrate x 3 res x 4 frame rate, 3 GoP x 5 frame loss rate, 5 delays.. Curse of dimensionality! -> Reducing stimulus & reducing rating time. Do we need to assess them all in interactive tests?

Method: (subselection of studies). Study 1: mix of bitrate-resolution pairs.

Assessed quality aspects: overall quality, 5 dimensions (like last talk), 7-points continuous scale.

Observations from results: strong correlation between video quality and gaming quality. Most gamer can’t distinguish 30 and 60 fps.

Conclusions: main effect of framerate on overall quality; main effect of framerate on input quality. Spatial quality...

Jeremy: market perspective: how is doing “cloud gaming”. Narciso: The market is still quite small, but working with companies in Spain. Jeremy: have question on setup of bitrate.

Jeremy: issues - bandwidth, latency?

On the Use of CNN for Quality Assessment of CGI (Saman Zadtootaghaj, Markus Utke)

Introduction to CNN.

VQ assessment using CNN: challenges: which level - video or frame level? How to make it work on video level?

Using motion flow together with image data.

Using image level information

Tried: use VMAF as quality indicator of each frame (similar to DeViQ[7])

Idea: retrain some part of the network and retrain some layers based on the subjective results

“Transfer learning”: train network for some goal, use the trained network for something else. Freeze layers, train the last layers. How many layers?

Work in progress. Use VMAF, no subjective test. Current results:

DenseNet121: SRCC 0.918

Next question: find a subset of frames that can be used for training of the last few layers of CNN

Data: 24 videos, 400 distorted sequences, 30 seconds each.

Transfer learning: using last 2 layers: SRCC: 0.914; more layers (?): SRCC: 0.949; … SRCC: 0.950

Summary:

* Plan is to make no-reference quality metric using CNN or gaming
* Main Aim: not only to predict quality but also measure the type of distortion
* Gaming content has special characteristic which make it suitable content for machine learning methods

Ioannis: complexity compared to VMAF itself?

* Speaker: cloud gaming scenario don’t have reference, so has to be no reference.

Zhi: should test on monotonicity behavior with encoding resolution/QP

* Speaker: tried on different bitrates 1 - 2 Mbps. Need more results

Margaret: similarity of across different content in terms of speed

SAM

AccAnn: a new QoE measurement methodology and its application on video streaming (Jing Li)

For streaming service provider: more important: below which the VQ is acceptable? Is satisfactory?

State of the art: two-step (is video acceptable? Y/N is video annoying? Y/N)

2015 D. Khaustova: one-step (not annoying, annoying but acceptable, not annoying)

Originally applied to 3DTV, this work: apply in general

User profile -> user’s expectation

Experiment instruction: imagine you are in basic (6 euro) or premium (12 euro), rate according to this setup

Clarify by Lukas: mentioned two profiles in the instruction, how much each pay.

Analysis of AccAnn data: do not assign a MOS score, but consider 3 categories

Define threshold: annoyance threshold: 50% subject voted annoying; acceptable threshold: 50% subject voted acceptable.

Application I: benchmarking state-of-the-art metrics: VMAF performed the best.

VMAF: threshold of acceptability: 66; threshold of annoyance: 80.

Premium users are more picky than basic users (users are presumed to be in a premium/basic)

Influence of device using Eliminated by Aspect (EBA) model. Observation: higher the quality, less influence from device; more influence on tablet than on devices.

Conclusions:

…

Important information for service providers:

* How much difference could be made on video encodes for difference devices?

Train objective quality:

* Based on recovered device-neutral AccAnn score + adapt to different devices

Reference: TMM’19 and PCS’18

Jeremy: device, acceptance - has been studied since the 80’s.

Florence: end-to-end, how long (~1 year)

Lucjan: did you also ask the MOS besides Acc/Ann? -> No

Day 4 – Thursday, March 7

SAM

Tables for planning the number of subjects (Kjell Brunnström)

Presentation and work continue on the work described in the journal article:

“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.” Fulltext: <http://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Ari%3Adiva-35233>

Based on a number of expected t-tests / within or between experiments / expected MOS difference and standard deviation of results → get the number of subjects needed. The calculation has to be done before the experiment.

Tables with typical numbers and the code will be proposed for amending/correcting ITU-T recs (P.913, P.910) and ITU-R Rec. BT.500.

Margaret offers to help draft the text.

Question about whether it’s also possible to have as input “number of SRC/HRC” which is what people usually have when planning an experiment

Margaret proposed to consider differences among subjects.

Online repository for code: <https://github.com/VQEG/number-of-subjects>

Demo app to calculate numbers interactively: <https://slhck.shinyapps.io/number-of-subjects/>

Problem with the Continuous Model (Lucjan Janowski)

The presentation is here.

Presentation and discussion (Margaret, Narciso).

Bayesian Subject Model - Demo Session (Krzysztof Rusek)

Live, an online demonstration using Google Colab. Link: <https://colab.research.google.com/github/krzysztofrusek/tfp_subject_model/blob/master/notebook/Subject_model.ipynb>. Discussion with Zhi, Lucjan, Ioannis and Jing.

suJSON - A Uniform JSON-based Subjective Data Format (Jakub Nawała)

Presentation of suJSON. Diagram Enhanced Entity-Relationship (EER) available at <http://bit.ly/2NImZy2>. Followed by a live demonstration.

Questions:

* **Do you agree that having a standardised subjective data format is crucial? Anyone objecting?** Pablo raising a question about format complexity. Kjell seconding. Werner as well. The bottom line is that perhaps the format should be more limited (covering the major part of subjective experiments, but not necessarily all of them). Ioannis proposing to cover just video experiments. Zhi pointing to the simplified diagram (slide #5)
* **Where should we stop with improving the format? Maybe now?** Discussed together with the previous question.
* **Imagine the case when we want to ask a tester to order 10 PVSs. Assume that he/she can change the ordering at any time. Do we just store the final ordering or do we want to monitor all events in between?** Ioannis: intermediate steps are just extra scores. Perhaps this can be introduced as optional data in the structure. Pablo: to add a simple CSV to be easy to be parsed. Zhi: extending the diagram to include pair comparison. Kjell: to propose support (by providing some insight into his subjective testing software).
* Werner to share data formats used in ITU. (Note from Werner: the format is similar to the one assumed in [this utility script](https://github.com/Netflix/sureal/blob/master/resource/util/mos_to_sureal.py), basically CSV with headers “PVS\_ID”, “S1”, “S2”)

NORM

Alternate Training Data (Margaret Pinson)

Discussion.

QP deltas of 6 yields discernable quality differences, between 22 to 42 QP for VP9, AV1, AVC, and HEVC; works for most video professional content. 28 to 38 perhaps for user-generated content.

User-generated content: measure degradation or absolute score / MOS? What is the baseline?

We need an inventory analysis.

Camera noise, varry camera noise power low to high, e.g., Netflix movies creatively add camera noise for artistic intent. Adobe Premiere Pro, for example, an option. This may be considered “post-production noise”

Regular noise, perceptual noise, unrelated to camera, such as white gaussian noise, double for each step. Other types of noise are relevant.

Enhancement and over-enhancement: metric should be able to detect optimal quality through enhancement, and over-enhancement. Sharpening, color enhancement, contrast enhancement, de-noising, deblurring, upscaling, downscaling. There is a dataset for sharpening.

Apply scaling to reduce contrast: multiply the range will yield inferior performance. For example, 0.8, 0.6, 0.4.

Increase exposure until saturation, a typical defect that burns sky, automatically increase linearly contrast or use tool. Histogram should show clipping as problems appear.

Desaturation is trickier, since black & white can be desirable.

Color temperature, up to some point (white balance). This is also tricky.

JEG-Hybrid

JEG-Hybrid (Enrico) gave an update on their project status.

Florence informed the group on her efforts to get source content for JEG-Hybrid, commercial applications and objective metric validation.

NORM

Florence presented professional produced content pipeline and were potentially objective quality metrics should be used in it.

Agreement was reached to write a liaison encouraging standard metadata to hold quality assessments added to MPEG. Ioannis will draft this liaison.

The above effort will be added to the NORM group’s topics under discussion. This will enable work on reduced reference (RR) metrics that pass information (e.g., NR metric) down the chain, as the basis for an RR metric or workflow quality evaluation.

Decided to have monthly audio calls, at the same time as SAM but on alternate weeks (e.g., 8am Pacific time (PST), starting March 18).

Several organizations will seek source material to enable this research, for example representative of user generated content, by Facebook; and Sky to seek professionally produced live content; and NTIA to contribute footage representative of public safety content.

ITU-T SG12 “Q44” Interim Meeting

ITU-T Study Group 12 had a three day breakout session. See the ITU Report for details.

Q13 discussed VR documents mentioned above.

Q44 had remote participation throughout.

P.NATS Phase 2 / AVHD-AS: Input to validation database gathered. Footage issue discussed; several organizations may be able to contribute footage, so this is on a good path. Good progress was made, including Python scripts for HRCs, actions to be taken when organizations miss deadlines, and discussions on whether to make some subjective test results available. The task is on track.

Q17 focuses on IP packet performance (e.g., how to measure the internet). Capacity measurements are under discussion. The old methods for access measurement are failing due to technology advances. Comparison against calibrated networks. Goal is to change and standardize the whole speed measurement system.