VQEG Meeting Minutes

13 to 17 December, 2021

Day 1, Dec. 13, 2021

Overview of VQEG Groups

AVHD

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

Recently completed a joint project between ITU-T SG12 and VQEG, 35 candidate models submitted, extensive validation testing. Result is P.1204 including bit-stream model, reduced reference model, and hybrid no-reference model. Intend to expand standard to include new codecs and situations.

PsyPhyQA

<https://www.its.bldrdoc.gov/vqeg/projects/psyphyqa/psyphyqa.aspx>

Subjective quality evaluation using physiological signals, such as EEG. Due to COVID, few tests could be conducted recently. Please reach out to the PsyPhyQA group if you are interested in conducting this type of research. Test plan is listed on the VQEG webpage.

QAH (see presentation 229 for details)

<https://www.its.bldrdoc.gov/vqeg/projects/quality-assessment-for-health-applications-qah.aspx>

1) Topical review [paper](https://pubmed.ncbi.nlm.nih.gov/34225264/) “Comparative study of the methodologies used for subjective medical image quality assessment” (Lévêque *et al*.) published in July in *Physics in Medicine & Biology*.
2) Topical review paper “Objective quality assessment of medical images and videos: Review and challenges” (Rodrigues *et al*.) submitted in December to *Medical Image Analysis* (see presentation 227 for details).

3) Invitation to contribute to our *ICIP 2022* special session: see [Quality Assessment for Medical Imaging Applications](https://docs.google.com/document/d/1phXIn-f7ys8FUnkIOa__XspCxVWkEUepwBZS_3y3L60/edit?usp=sharing). Reach ASAP! lucie.leveque@univ-nantes.fr

SAM

<https://www.its.bldrdoc.gov/vqeg/projects/statistical-analysis-methods-sam.aspx>

Monthly meetings conducted on statistical analysis methods for subjective tests.

CGI

<https://www.its.bldrdoc.gov/vqeg/projects/computer-generated-imagery-cgi.aspx>

Quality analysis of computer generated imagery, with a focus on gaming in particular. There are related ITU work items: P.809, questionnaire for interactive cloud gaming quality assessment, quality assessment using crowdsourcing (P.CROWDG) and ITU-TP.BBQCG: bit-stream based quality assessment of cloud gaming services; codec comparison for computer generated content. Reach out to the chairs for more information.

NORM

<https://www.its.bldrdoc.gov/vqeg/projects/no-reference-metrics-norm.aspx>

Meetings held every other week, multiple topics:

1. SI/TI metrics. Are still being used but have become ambiguous over time. Goal of providing reference software and updating P.910 recommendation.
2. Standard metadata for video quality assessment-related information.

For this meeting, various other topics in presentations on Tuesday.

IGVQM

<https://www.its.bldrdoc.gov/vqeg/projects/implementer-s-guide-to-video-quality-metrics-igvqm.aspx>

Implementer’s guide for video quality metrics. Goal is to create a report on the accuracy of video quality metrics. Have made a high level plan: 1) collect databases, 2) collect metrics and methods for assessment, and 3) statistical analysis.

JEG-Hybrid

<https://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx>

Goal is joint work to develop hybrid perceptual/bitstream metric. Creating a large database for training Hybrid metrics, using FR metrics instead of subjective tests. Focusing on modeling individual observers. Just concluded a project in collaboration with Sky (presented in Spring 2021).

QaCOViA

<https://www.its.bldrdoc.gov/vqeg/projects/qacovia.aspx>

Quality assessment for computer vision applications.

5G-KPI

​​<https://www.its.bldrdoc.gov/vqeg/projects/5g-key-performance-indicators.aspx>

Study relationship between key performance indicators of new communications networks (especially 5G) and QoE of video services on top of them. Define relevant use cases, etc. 5G-KPI is holding monthly audio calls on 2nd Thursday each month, 15:30 to 16:30 CET/CEST. To join, send email to Pablo Perez, (Nokia) <pablo.perez@nokia-bell-labs.com>

Immersive Media Group (IMG)

<https://www.its.bldrdoc.gov/vqeg/projects/immersive-media-group.aspx>

360-degree content, virtual reality, augmented reality, light field/plenoptic content, 3D content, etc. Goals is baseline quality assessment: 1) datasets, and 2) QoE guidelines, subjective methods objective methods, etc. IMG test plan on quality assessment is finalized. ITU-T P.919. Contributing to MPEG AHG on quality of immersive media and ITU-T SG12 (G.CMVTQS). Ongoing joint work analyzing data for 360 degree long videos.

Support and outreach

* E-letter

Human Factors for Visual Experience

Deep learning based assessment. Draft standard submitted. Quality assessment for light field content. Submitting content to standard at the end of the month. VQEG will hear an update on these efforts on Friday.

Qualinet

Works on quality of experience topics. Currently one meeting a year, in connection with QoMEX conference.

Tools and Subjective Labs Setup Group

Website on GitHub that lists links to tools related to quality assessment. These tools are sorted by category and purpose. No tools have been added since the last meeting.

Website Support

Provided by Margaret Pinson (NTIA/ITS). Anonymous FTP is nearing end-of-life. Files can be accessed from old browsers or via an FTP app. The files will be embedded into the VQEG website over the next 6 months or so.

IRG-AVQA

No official meeting, but activities will be summarized on Wednesday, Session 1.

JEG-Hybrid

#214. “Updates on recent JEG-Hybrid activity” by Enrico Masala/Lohic Fotio Tiotsop, Politecnico di Torino

Overview of activities. In particular:

* Results from the Hodor project with Sky Group and affiliates have been finally published: L. Fotio Tiotsop, F.  Agboma, G. Van Wallendael,  A.  Aldahdooh, S. Bosse, L. Janowski, M. Barkowsky, E. Masala, "[On the Link between Subjective Score Prediction and Disagreement of Video Quality Metrics](https://doi.org/10.1109/ACCESS.2021.3127395)",  In: IEEE Access, 2021 (Open Access)
	+ Key outcome:  A measure that could allow to automatically identify PVSs for which VQMs are likely to deliver inaccurate MOS estimation

#215. “AI observers for video quality evaluation” by Lohic Fotio Tiotsop

Presenting:

* How to reuse existing subjective datasets to deal with new challenges?
	+ Potential extension to AV1

* A "stamp" in scientific publications for reporting  the training process and benchmarking of DNN-based VQMs
	+ Pablo Perez and David Lindero would like to be kept in the loop
	+ Pablo: if you want to go for standardization, start the process as early as possible

* AI Observers: bias and inconsistency
	+ Lohic’s question: how much inconsistency is acceptable from an artificial observer (compared to a human one)?
	+ In answer to Lohic’s question, see this paper:  Margaret H. Pinson, "[Confidence Intervals for Subjective Tests and Objective Metrics That Assess Image, Video, Speech, or Audiovisual Quality](https://www.its.bldrdoc.gov/publications/3253.aspx)," NTIA Technical Report TR-21-550, October 2020. <https://www.its.bldrdoc.gov/research-topics/video-quality-research/documents-dynamic.aspx>

Statistical Analysis Methods (SAM)

#213. “A tale of two datasets” by Zhi Li & Lukas Krasula, Netflix

Demonstrate changes in the spread and bias of media, influenced by the distribution of quality in the test. Use of an objective metric’s analysis in pre-testing could help avoid this problem.

NORM presentation #205 will analyze the CAMBI datasets.

#216. “FOWR methodology: update and follow-up” by Pablo Pérez, Nokia Bell Labs

Pablo Perez; Lucjan Janowski; Narciso Garcia; Margaret H. Pinson, "[Subjective Assessment Experiments That Recruit Few Observers With Repetitions (FOWR)](https://www.its.bldrdoc.gov/publications/3273.aspx) ," *IEEE Transactions on Multimedia (Early Access)*, 21 July 2021. [doi: 10.1109/TMM.2021.3098450](http://dx.doi.org/10.1109/TMM.2021.3098450)

Dataset is available on the Consumer Digital Video Library (CDVL). Log in with a free account and then use the Advanced Search option with key word “FOWR”; or go here: <https://www.cdvl.org/members-section/view-file/?id=3034> Download includes videos and individual subject ratings.

Propose to modify ITU-T P.913. Quasi-experimental assessments when cannot have 16 to 24 subjects (e.g., pre-tests, expert assessment, and resource limitations). Propose 4 subjects 4 times each on different days (similar to 15 subject test), or 6 subjects rate test 5 times each on different days (similar to 24 subject test).

Discussion ensued about the potential impact genuine differences among subjects. Werner noted atest where subjects differed in opinion on rebuffering (see [Robitza, Werner, Marie Neige Garcia, and Alexander Raake. "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm." 2015 seventh international workshop on quality of multimedia experience (QoMEX). IEEE, 2015.](https://www.researchgate.net/profile/Werner-Robitza/publication/277475666_At_Home_in_the_Lab_Assessing_Audiovisual_Quality_of_HTTP-based_Adaptive_Streaming_with_an_Immersive_Test_Paradigm/links/55d5f02908aed6a199a34c26/At-Home-in-the-Lab-Assessing-Audiovisual-Quality-of-HTTP-based-Adaptive-Streaming-with-an-Immersive-Test-Paradigm.pdf)), but this is expected mainly to be impactful for QoE tests (more realistic, less lab). Margaret noted genuine differences in order are not commonly observed in a typical test, as opposed to the impact of random errors on rating behaviors which are easily observed. That is, the level of concern exceeds the likelihood based on prior observations. Ioannis observed differences in preference for blocking artifacts and blurring artifacts. Short term and long term variability appear to be different effects, so repeating a test once a week may yield different results than on subsequent days.

Call for other labs to repeat these results (i.e., conduct experiment with both the FOWR method and the conventional 15 to 24 subjects). This will aid our proposal to include this method in ITU-T Rec. P.913, and to discover FOWR method viability, scope, and limitations.

#225. “(Tentatively) New Subject Removal Method” by Lucjan Janowski & Irene Viola, AGH & CWI

Analyze performance of subject screening methods. Can screening method detect subjects whose scores are randomly matched to the wrong stimuli (e.g., data recording or processing error)? How sensitive are the methods to changes in subject bias? How sensitive are the methods to changes in sigma (random error associated with the subject)?

Analyses note problems with ITU-T Rec. BT.500 subject screening method (Annex 1). Cannot recommend this method for ACR tests.

P.913 subject screening method works find. Need to further investigate threshold, because the threshold is arbitrary.

Future work: investigate impact on MOS. Relationship between discarding subject and ability of ratings to detect significant differences among videos.

#222. “Future of SAM” (Discussion)

Proposal (Margaret): confidence intervals for subjective tests, new methods for comparing subjective tests at 2+ labs, and how to extend these results to better understand the precision of objective metrics. Baseline for discussions available here: <https://www.its.bldrdoc.gov/research-topics/video-quality-research/documents-dynamic.aspx>

Proposal (Zhi): crowdsourcing platform, how to make rating results more reliable, how to improve cost effectiveness (e.g., active sampling of stimuli, on-the-fly analyses indicate what stimuli to next be shown to a particular subject). This could have long term impact.

Proposal (Lucjan): working on general score distributions. Consider different distributions (other than normal distribution), speed the estimation process, etc.

Proposal (Ioannis): Supports Margaret’s and Zhi’s proposals, interested in contributing.

Agreement was reached: SAM group wants to continue discussions, meeting every 4 weeks. Propose resume SAM meetings on Monday at previous time. Zhi will establish meeting invitations. Start January 10, 2022.

Day 2, Dec. 14, 2021

NORM

#205, “Contrast Aware Multiscale Banding Index (CAMBI)” by Lukas Krasula at Netflix

VMAF and PSNR do not detect banding. BBAND responds but there is room for improvement. CAMBI focuses on high-quality video and works for dithering. Open source code is available on GitHub: <https://github.com/Netflix/vmaf/blob/master/resource/doc/cambi.md> CAMBI has not yet been tested on user generated content. More information can also be found at: <https://netflixtechblog.com/cambi-a-banding-artifact-detector-96777ae12fe2>

#207, “User-Generated Content (UGC) / In-The-Wild Video Content Detection” by Mikolaj Leszczuk at AGH University

How can we automatically detect UGC? Note different responses (e.g., to NR metrics) from professionally generated content (PGC). Collected UGC and PGC videos, adding annotation, will make dataset available and begin modeling soon.

Note 3rd label: occupationally generated content (OGC).

#209, “Video complexity analysis for streaming applications” by Vignesh Menon & Hadi Amirpour, AAU

Open-source project. Online prediction for live streaming. Pearson correlation between SI and bits per frame around 0.79. Run speed of SI-TI features too slow for live 4K applications. Goal: video complexity analyzer (VQA) fast enough for live applications, feed into encoder. This metric is also useful for shot detection.

Will release code before March 1, 2022. Adding CUDA/OpenCL support, adding multi-threading support.

#210, “Perceptual quality assessment of internet videos” by Jing Li at Alibaba

Challenges: how to collect subjective scores, how much data, how to run enough subjective tests with limited budget, and how to remove outliers. **Youku-V1K** database. Full and uniform coverage of Youku content. Objective metric **STDAM**, currently internal use.

#202, “Why NR metrics fail: comprehensive analysis ” by Margaret Pinson, NTIA/ITS

Wishlist for what a NR metric should deliver (gathered from VQEG discussions).

Conclusions from survey of existing metrics:

* Most NR metrics only trained on 1–4 (mostly 1) dataset
* Independent evaluation results in a large range of correlation values against subj. Scores
* → NR metrics should be evaluated on much larger dataset

Analysis of 12 datasets for NR metric training and testing, with 27 different metrics. Results show poor correlation with originally reported results. Overview of select metrics’ performance against all datasets given.

Presentation of new version of own metric Sawatch (3) within NR metrics framework (<https://github.com/NTIA/NRMetricFramework>).

Usefulness of metrics for root cause analysis depends on use case (e.g. size of subjects, or some kind of distortion might be appreciated more or less depending on use case).

Comments:

* Ioannis: Training on 1 dataset is a risk due to overfitting. Suggestion to replace linear combination of features with SVM/SVR. Focus on fewer individual aspects with the most impact (compression artifacts) for bigger use cases.
* Lucjan: Combining datasets, problems with the way subjective tests are done
* Ioannis: Recommendation for everyone to publish their metrics as open-source, since that means others can help improve it

#212, “YouVQ: A new no-reference metric for UGC” by Balu Adsumilli & Yilin Wang, Youtube

The handcrafted feature approach to NR metric development is difficult & time consuming. Machine learning suitable for large UGC datasets. Transfer learning framework: pre-train on non-UGC quality data, then re-train on UGC quality, then fine tune on UGC data.

Output: content labels, distortion types, compression level, and quality score.

<https://media.withyoutube.com/> has videos and subjective scores

The authors hope to make the YouVQ metric available early in 2022, open source.

#203, “Standard video quality metadata” by Ioannis Katsavounidis, Facebook (company name changed to Meta)

Vision is to make metric measurements available in video streams. Video quality metrics are often computed (e.g., on video encoder ASIC) but discarded. Life-cycle of UGC video includes many processing steps, each with its own processing and metric calculations. History of metric values would be valuable for transcoding pipeline. Avoid re-calculating which adds significant compute overhead.

VQEG is now an official T.35 “terminal provider”. <https://www.itu.int/rec/T-REC-T.35-200002-I> Registered with US street address of ITS (325 Broadway, Boulder, CO, USA), contacts: Kjell, Ioannis, and Margaret. Mfg code first (MSB) byte: Hex: 48. Mfg Code second (LSB) byte: Hex: 10.

Proposal: three payloads. Summary payload, metric instance descriptor, and metric instance value.

This is an ongoing effort of NORM, with monthly audio calls.

* Audio call minutes [here](https://docs.google.com/document/d/154y5_CElepcZV6TX-3g_HfZCAjWGN8I-Vb-4gjub6Rw/edit?usp=sharing).
* Working document is [here](https://docs.google.com/document/d/1v02cd7tFz-YozfctAy2OkKz1dGX8eHuk-1lueaxobYE/edit?usp=sharing).

We will continue this discussion on Friday, December 17.

#204, “Updating SI/TI for modern video systems” by Lukas Krasula, Cosmin Stejerean, Werner Robitza, Netflix, Facebook, AVEQ/TU Ilmenau

Goal is to clearly specify SI/TI algorithms, in response to changes in video technology. Harmonize and unify for different implementations (e.g., handling of edges, full or limited range conversions), how applied to HDR content, how to handle more than 8 bits per channel. New python software developed. Stable main branch and current branch “siti2020” for current updates. See <https://github.com/Telecommunication-Telemedia-Assessment/siti-tools/> and <https://github.com/Telecommunication-Telemedia-Assessment/siti-tools/tree/siti2020>

Video recording includes a demonstration of the software running. Open issues: performance is a bit slow, and some issues with range of SI values.

* [Minutes (Google Doc)](https://docs.google.com/document/d/1pjAJet6YMznf1pPZ_5Xp0L3UiJvCh78x2LRIHvKApro/edit#heading=h.pod7m6xer5vu)
* [Python code](https://github.com/Telecommunication-Telemedia-Assessment/siti-tools/tree/siti2020) repository

How to compare different encoders against output of new metric? Shahid: H.264 is basically legacy now and could be compared against. Cosmin: Idea would be to be able to do it at the fraction of the cost of actually encoding the sequence.

January 17, continue SI/TI discussion, at usual time (16h UTC, 17h CET)

January 24, continue standard metadata discussion, at usual time, discuss time of day for future meetings

#206, “Datasets for NR metric research,” by Margaret Pinson, Mikołaj Leszczuk, Lukas Krasula, Nabajeet Barman, Maria Martini, Jing Li

The presentation contains a list of datasets for NR metric research including publication and the dataset’s download link.

Another dataset

**AVT-VQDB-UHD-1**: A dataset consisting of 4 different tests.

Paper: <https://ieeexplore.ieee.org/document/8959059>

Dataset: <https://github.com/Telecommunication-Telemedia-Assessment/AVT-VQDB-UHD-1>

#233, “Encoding complexity of short video sequences” by Shahid Satti, OPTICOM GmbH

A low cost CRF based encoding complexity measure was introduced. The measure was shown to perform better than SI/TI in terms of pearson and spearman correlation. The CRF complexity measure is a feature in ITU-T p.1204.5 Hybrid no-reference model.

Lucjan: This can be used to analyze the src complexity in subjective tests

Alex: This was already done in P.NATS2

Shahid: Yes, a similar method was used to create complexity classes of srcs in PNATS2

Day 3, Dec. 15, 2021

Quality-related ITU Activities

Presentation on ITU-R activities (Chulhee Lee)

* Activities: special focus on HDR
* Less attendance and fewer contributions mostly due to COVID-induced delays in subjective testing
* Discussion on ITU-R BT.500 – how to capture other environments like operational environments or at home

Presentation on ITU-T activities (Alexander Raake, Chulhee Lee)

* Questions on models for even longer sessions (> 5 minutes), is something like this planned? → mostly restricted in terms of test content and ways to test this
* Andy Quested: Tests running at BBC Research (Katy Noland) in various contexts with crowdsourcing assessment of longer content

IGVQM Databases (Ioannis Katsavounidis)

VQEG’s Implementer’s Guide to Video Quality Metrics (IGVQM) project

“This project attempts to offer some guidance to the video coding community, including standards setting organization (SSOs), on how to better utilize existing objective video quality metrics to better capture the improvements offered by video coding tools.”

Overview document: [VQEG - IGVQM](https://docs.google.com/document/d/1w3rgGxDHFehwdwtVXENtngAPsMhGnPlAsiU2v0Tu6nU/edit)

Spreadsheet: [IGVQM databases](https://docs.google.com/spreadsheets/d/1wYVjaqKrSB0z7D2eJLfAd1g6y_831IF7TH6GhAKdKxA/edit#gid=0)

Call for the community to contribute databases and metrics.

AVHD

#208: Technology for Secure and Reliable Delivery of Professional Audio/Video Contribution Live Transmissions With Lowest Possible Latency (Mikolaj Leszczuk, AGH University)

Comment by Ioannis pointing out existing study from [[1701.01500] VideoSet: A Large-Scale Compressed Video Quality Dataset Based on JND Measurement](https://arxiv.org/abs/1701.01500)

IEEE Videoset: [VideoSet video quality dataset | IEEE DataPort](https://ieee-dataport.org/documents/videoset)

Dynamic Optimizer Netflix tech blog: [Dynamic optimizer — a perceptual video encoding optimization framework | by Netflix Technology Blog | Netflix TechBlog](https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f)

#223: 8K Image Quality (Andy Quested)

Not a presentation, rather a discussion on how to assess the video quality for very high resolution (8K, 16K, 32K…) monitors for interactive applications.

Ioannis: While 0.75\*H remains to be the correct distance to view the image to absorb details of the image without pixelation, it can cause ergonomic issues. The ability to be able to zoom in into the different parts of the image is the key.

ITU-T SG6 is working on a vision document….(Andrew to be add more text in here)

Margaret argued that high details do not necessarily justify 8k monitors if the possibility of zooming in is available.

5GKPI

#226, “Proposal for KPI message format for gaming QoE over 5G networks” by Werner Robitza, AVEQ GmbH

Draft specification: [GQOE KPI Message Format — Draft](https://docs.google.com/document/d/11xeFlfLklIruAkOm7_9Qw06hU6H382SoQdYfMgUZ94A/edit#heading=h.oeonok5jrsp4)

* David: Consider display framerate too, not just decoded frames
* Saman: Consider device characteristics, properties like resolution, refresh rate of the display.
* Nabajeet: If not mmwave, the results might be expected, maybe slightly better than 4G
* Pablo: developed an opinion model (presented at QoMEX)
* Margaret: 5G technology is still maturing, results might change
* Lucjan: Was this used to make decisions on network slicing? No, not yet
* Pablo: Network providers would be interested to get this kind of data, but they would have some limitations in terms of what they could with the data (net neutrality). Also a matter of standardization – the number of tools you have a available in production for reacting are limited

#218, “A parametric quality model for Tele-operated Driving” by Pablo Pérez, Nokia Bell Labs

Automated driving, model proposed and analyzed based on performance measurements on 5G networks. The model has been designed based on assumptions on QoE-QoS and state-of-the-art and used to evaluate the feasibility of doing Tele-operated Driving over different networks.

Reference: Pérez, P., Ruiz, J., Benito, I., & López, R. (2021). A parametric quality model to evaluate the performance of tele-operated driving services over 5G networks. *Multimedia Tools and Applications*, 1-17.

#221 “Introduction to 5G” by Margaret Pinson, NTIA/ITS

Overall description of 5G technology.

* Differences in spectrum allocation per country impact the propagation and responsiveness and throughput of 5G devices
* Pablo provided more details on causes for the limitations on possible bandwidth improvement
* Ioannis provided more details on ORAN Alliance (how to leverage ORAN metrics/APIs to optimize video delivery QoE).

#217, “ITU-T GSTR-5GQoE "QoE requirements for real-time multimedia services over 5G networks”  by Pablo Pérez, Nokia Bell Labs

Presentation of the status of the Technical Report which is being presented to ITU-T SG12.

* Ioannis: it’s important to add energy consumption KPIs, especially for battery-operated use cases (and in general for CO2 emissions).
* Pablo: it might be not possible in a use-case basis, but we can try to include some general guidelines an KPIs. To be analyzed.

Day 4, Dec. 16, 2021

CGI

#231, “Updates on LCEVC for Gaming Video Streaming Applications using subjective datasets” by Nabajeet Barman KU/Brightcove

The work was presented in the last meeting, this time, the subjective results will be presented.

* The work is a comparison of compression efficiency of LCEVC with existing practical implementation of video codec standards H.264/AVC and H.265/HEVC.
* LCEVC is a two layer enhancement model, which is supposed to improve video compression efficiency compared to base codecs.
* Codec decision: Medium preset for H.264, and veryfast for HEVC.
* LCEVC-x264 offers 42% bitrate saving compared to Native-x264  based on the BD-BR analysis using VMAF.
* LCEVC-x265 offers 35% bitrate saving compared to Native-x265  based on the BD-BR analysis using VMAF. However using PSNR, the x265 outperforms LCEVC-x265.
* Subjective test results indicate that at bitrates lower than 3 Mbps, LCEVC x264 outperforms x264. At higher bitrates, the performance of the two codecs are quite similar.
* In terms of MOS ratings, no significant difference was found between LCEVC-x265 and x265 codecs.

Ioannis: 1. Quite low bitrates (such as 800 kbps) for 1080p resolution are not practical (considering multiple-bitrate encoding resolutions, a lower resolution such as 720p or 540p could make more sense and hence the practical results might be quite different.

2. Per-frame bitrate can give a better indication of frame quality variation

To be checked by Nabajeet and Saman.

David: 1) Question about command settings

— (NB) not specific settings used for either of the compared codecs.

    2) Check for P and B frames for the codecs!!

        NB to follow up

Administration

Lu Zhang of IETR - INSA Rennes, France, has offered to host a face-to-face meeting in Spring, 2022.

Proposal: stream video to remote participants, remote participants submit questions via text, to encourage conversation in-person

Poll of interest in attending a face-to-face meeting in Spring. 12 in favor. 6 against. 4 undecided. We will conduct a Doodle Poll on the VQEG reflector.

Weeks considered:

* May 2-6 = potentially ok
* May 9-13 = potentially ok
* May 16-20 = potentially ok
* May 23-27 = problematic, Christian holiday, very expensive and crowded, + INSA Rennes closes on 27th and 28th

3GPP meets May 11 to 20

Agreement was reached: close Joint Effort Group (JEG) and Independent Lab Group (ILG).

Discussion on online meeting time

* 2 hour later start time would improve participation form US west coast
* Asian participation would be worse, but participation from North America is generally higher than Asia
* Propose avoid midnight to 5:00am in all timezones, do not have contiguous 5 hours (separate blocks of time)

This topic will be added to the Doodle Poll: time of day for online sessions, dates for spring meeting, and attendance of face-to-face (if still possible in spring)

#237, Call for papers (Quality and UX): Crowdsourced and Remote User Studies for Quality of Experience and Usability Research <https://www.springer.com/journal/41233/updates/19733142>

QAH

#227, “Objective quality assessment for medical images and videos: Review and challenges” by Lucie Lévêque LS2N - Nantes Université & Rafael Rodrigues Instituto de Telecomunicações - UBI

Summary of what has been done in 2021. Note:  medical datasets are available on the web page (presentation: [206-medical](https://www.its.bldrdoc.gov/media/66824/vqeg_nrorm_2021_206_medical-database-for-nr.pptx))

1) Topical review paper “Comparative study of the methodologies used for subjective medical image quality assessment” (Lévêque *et al*.), published in July in *Physics in Medicine & Biology*.

2) Special session submitted to *ICIP 2022* in December: Quality assessment for medical imaging applications (notification of acceptance early January).

3) Topical review paper “Objective quality assessment of medical images and videos: Review and challenges” (Rodrigues *et al*.), submitted in December to *Medical Image Analysis*.

* Survey of objective quality assessment (QA) methods for medical images and videos;
* Categorisation of QA methods into visual quality-based (FR, RR, NR metrics) and task-based approaches (model observers);
* Analysis of key findings and recent advances in medical imaging objective QA;
* Discussion of limitations of the reported studies and future work considerations (artifact simulation, deep learning methods, characterisation task…).

Margaret: hard to get content… how to improve this?
Main problem is linked to ethics… maybe we can try to launch a call? Through VQEG/QAH.

Healthcare professionals are extremely busy now with the COVID-19 pandemic…

+ Demanding in terms of time etc for researchers (need to go to hospitals).

Human-in-the-loop systems? Software to collect evaluation while doctors do diagnosis?

#228, “Model observers for the objective quality assessment of medical images” by Lu Zhang IETR - INSA Rennes

Presentation initially made for another workshop where Lu advertised the QAH group.
Complement of Rafael’s talk on task-based approaches.

3 tasks: detection, localisation, characterisation.

Model observers: IO (ideal observer), HO (hotelling observer), CHO (channelized HO).

Even the FDA uses model observers! :-D

AVHD

#220, “Unified VMAF - exploring reducing VMAF complexity by unifying features in the wavelet domain” by Abhinau Kumar UT Austin & Cosmin Stejerean Facebook

Day 5, Dec. 17, 2021

IMG

#211, “Projects at the AVT group of TU Ilmenau” by Alexander Raake, TU Ilmenau

#219, “360° Video Database with Higher-Order Ambisonics Spatial Audio” by Ashutosh Singla, TU Ilmenau

#236, “Update on IEEE standardisation activities on HFVE” by Maria Martini, Kingston University London

#234, “Follow-up on the new test plan” by Pablo Pérez & Jesús Gutiérrez, Nokia Bell Labs / UPM

Ludo Malfait (Dolby) developed an interactive speech quality test method that encouraged interaction between subjects by making the test into a game. You might look for his research or reach out to Ludo directly.

#239, “Legibility in Virtual Reality” by Kjell Brunnström, RISE

VR legibility and perception of speech-to-text by Defa and hard of hearing.

NORM

#235, “The KonVid-150k Video Quality Assessment Dataset & Objective Metrics” by Franz Götz-Hahn, Universität Konstanz/Universität Kassel

Overall, conclude that subjective tests with few ratings per subject are very valuable for training NR metrics. Infer that diversity of content more important than precision of MOS.

<<reference to be added, to paper describing this dataset>>

Continued discussion on standard metadata

Changes marked in the working document

Meeting Closed