VQEG Meeting Minutes

Shenzhen, 14-17 Oct, 2019

Day 1: Monday 14 Oct 2019

Welcome, general information, and introduction of participants

Groups Introduction

IMG (immersive media group) - Pablo Perez (vice-chair)

- Biggest milestone - subjective methodology for 360 video to be recommended to ITU-T

5G KPI - Pablo Perez (chair)

AVHD (Audiovisual HD) - Kjell Brunnström (no chairs present)

- Objective models for adaptive streaming

PsyPhyQA (Psychophysical Quality Assessment) – Kjell Brunnström (no chairs present)

SAM (Statistical Analysis Methods) – Lucjan Janowski (chair)

- How to make subjective experiments and the data analysis more precise.
- Goal is to adjust the recommendations

CGI (Computer Generated Imagery) – Saman Zadtootaghaj (chair)

- More focus devoted to analyzing and evaluating computer-generated content
- Two ITU-T Recommendations already produced

NORM (No-Reference Metrics) – Mike Colligan (new vice-chair)

JEG-Hybrid – Enrico Masala (vice-chair)

- main outcome so far - large-scale datasets for objective metrics development

QACoViA (Quality Assessment for Computer Vision Applications) – Mikolaj Leszczuk (chair)

Building objective model for predicting performance of computer vision algorithms

Support and dissemination groups

Overview provided by Kjell Brunnström

Discussion on how to approach ITU with concrete proposals

It has been decided that the best approach would be to send the liaison from VQEG as a whole, as well as approach ITU through members of both ITU and VQEG

Discussion on possible certification of datasets

Patrick Le Callet:
We can compute content characterization measures
We could state if some data is missing (e.g. confidence intervals of MOS)
Providing the results of objective metrics (to avoid having many different performances in different papers)

Strongly supported by other members

Zhi Li:
Proposing to tie su-JSON with Qualinet databases – putting everything into common format to make things easier for users to work with

Suggestion – this effort could be driven by SAM, potentially together with ILG

Resources can be coming from common pool of SAM, ILG, and potentially Qualinet databases task force (if agreed upon)

Christos Bampis:
we could standardize the exact steps that are used to calculate metrics
we could provide a table with what information are necessary for calculating different indicators

Patrick Le Callet:
If we are to provide guidelines for benchmarking the metrics – we should get rid of mapping of objective scores to subjective scores and include the confidence intervals

Kjell Brunnström:
This should be done through revising ITU recommendations and approach them with a concrete proposal

Pablo Perez:
We need to provide tools for people to use.
These tools are already on git – but need to be promoted!

Zhi Li:
We could also be providing clusters/computational resources so people could upload the datasets or metrics and the indicators will be calculated.

Patrick Le Callet:
IEEE Dataport can be an inspiration, if we want to go this way.

Saman Zadtootaghaj:
Computer Vision datasets are another example of datasets evaluation/benchmarking. Deep learning-based metrics should be benchmarked on independent databases.

Is this the effort that we should do?
The decision has been made – YES

The leader of the effort is to be identified later on this week! (after more discussions, especially during SAM sessions)

JEG-Hybrid

Presentation #8 gave status update and presentation of the published research activities on MOS range estimation from objective metrics:

Computing Quality-of-Experience Ranges for Video Quality Estimation, QoMEX 2019.

A Neural Network Based Approach for Observer Behaviour Investigation in Media Quality Assessment, submitted to Signal processing: Image communication journal, 2019

Presentation #16 "Characterising the performance of objective metrics with large- scale database" This was a status update on the collaboration between Sky affiliates and academic parties to understand the performance behaviour of different objective metrics.

IMG

VQEG-IMG Pre-Test - discussion. Participating labs include Nokia, UPM, TU IImenau, Roma3, RISE, U.Surrey, Wuhan U., U.Ghent and CWI

Decisions

Removing Thieves and Segovia

Increase training session by including middle quality and non-uniform quality

Initial questionnaires

Post questionnaires

Fix the orientation issue

Selection of playlist

Fix some PVS which are broken

Day 2: Tuesday 15 Oct 2019

NORM Session: NO reference metrics

Mike Colligan (new chair) gives his views about the NR metrics

---

Some ideas about how we could go forward

- What is the root cause is more important than just saying MOS = e.g. 3.5; useful for troubleshooting
- Lukas: do not focus only on impairments, there is also good quality when no impairment is clearly visible (e.g. when sharpening, color saturation, deblocking etc...: can increase but also decrease quality, there is tradeoff that could be interesting for NR metrics)
- Mike: anything that yields to a higher score is interesting
- Lukas: blur can be both enhancement in noisy conditions, or a distortion
- Saman Zadtootaghaj: study some of the enhancement techniques (filters, models) to understand if they improved the quality or not, mut maybe they deviate from the reference
- Florence: previous scope going to change? Maybe slightly, to advance faster than before.

Metrics should be connected to the use of the image: e.g. first responder: does image serve its purpose? This is different for entertainment video, etc.

Micolaj: particularly overlapping with QaCoVia

Christos: where the improvement will come from? Deep learning? or other? Depends on the type of use cases? First responder

Phil: still focused on first responder use case? yes

Florence: originally: 3 use case: first responders, entertainment, user generated content (Youtube). Which use cases can we focus on? First responders is the main use case, but all cases are still interesting.

Lukas: identify the next steps, to decide what to discuss in next calls (we just identified some sub-metrics in the past)

Phil: have presentation first, then discuss

Check if Margaret can join from remote on Thursday morning at 9am, and do an additional **NORM session on Thursday morning 9am (duration 1 hour).** Ioannis is also interested.

Presentation #18 by Jari Korhonen (Shenzhen University)

Two Level Approach for No-Reference Natural Video Quality Assessment

Several NR-VQM models available, however few learning models with implementation available, mostly focused on compression and transmission, and they too complex for practical applications.

Proposal: 2 levels: low complexity features (derived from motion but also spatial characteristics) for group of frames + representative frame selection to compute high complexity features

Performance in line with more complex algorithms, however still somehow slow, but can be improved in C++/OpenCV. Other proposals are faster but perform worse.

Future work could use CNN.

Published in IEEE Transactions of Image Processing 2019, implementation available in github

Saman: dataset is small (1,000 contents) but transfer learning techniques could be used for the CNN. Jari: issue with same MOS assigned to all frames, but quality might be fluctuating

Lucjan: limited precision of subjects in doing NR assessment. Jari: Typically outliers are exotic content (e.g., animated content, time lapses).

Christos: with no frame selection: what is the performance? Jari: probably no better performance in general, in this database it does not improve; cross-database test? Jari: Yes, results not that good, but probably because of content resolution and maybe different types of artifacts.

Christos: very easy to overfit in this kind of experiments, so how do we create algorithms, that do not overfit? Jari: here we use SVR, not so prone to overfit, we must be careful with NN.

Ioannis: which features are scale-dependent? Jari: sharpness is scale-dependent if you have large resolution. Other example is noise.

Lucjan: resolution play a role in terms of estimation of distributions if you have 100 or 1,000 pixels to consider

Ioannis: recommendations? Jari: probably build different models for different resolutions

Roberto Nery da Fonseca: also frame rate? Jari: Matters for temporal features.

Roberto: test on interlaced? Jari: interlaced content (even if not common in consumer generated content) must be included in the features, otherwise prediction is difficult.

Mike: scene detection to have 1 frame per scene? Jari: in user-generated content there is typically no scene change, just moving the device, gradual change, so periodic sampling seems a good compromise.

Christos: who designs datasets could include some content available from other datasets, so there is a way to align scores from different datasets.

SAM

Presentation #17: Jing Li, Alibaba Group

UPGC content selection and processing criteria for constructing a quality database

UPGC image/video quality assessment

PGC Professional Generated Content (e.g. professional bloggers)

Issues with converting horizontal to vertical content, with other types of presentation, make quality assessment very difficult.

Need to construct a database with these peculiarities

How to display content in a subjective experiment? Horizontal is ok (just rescale), but vertical?

Also, how to select content? Caption included? Tag-based analysis? Indicator based sampling (brightness, contrast, spatial activity, temporal activity, blur, blockiness, noise, exposure, colorfulness)? Quality based sampling using NR metrics?

Should we draft a recommendation for UPGC?

Asking for comments/suggestions:

Ioannis: Facebook and Instagram: very similar issues. Last conference in Taiwan: Youtube presented how to select user-generated content by Balu. Jing: Actually Jing knows the paper. Youtube does not have some ways of presenting, e.g., repeating the video vertically.

Patrick: purpose of Youtube dataset is check how UGC react to coding, while Jing focuses on quality of presentation (in subjective experiment) and also about what should be included in the dataset

Jing: another problem about using DeepLearing for super-resolution or HDR, … what happens for content never seen in the training dataset

Ioannis: Youtube uses sampling of indicators and standard clustering and selection approach.

Ioannis: for display: different depending on use cases: Instagram -> phone only, Facebook: it depends.

Lucjan: showing vertical and horizontal (image) on a PC screen, results are strange. Probably people distracted by the content, so quality evaluation was difficult. Interested in exploring the issue.

Roberto: what is the issue with caption? Are not they created at end user device? Jing: issue when it is part of the video itself

Ioannis: e.g. Instagram: you can add effects including text.

Phil: quality of caption is good enough … (I did not get the comment)

Lukas: just image quality: do not include caption. But include it for QoE

Patrick: video quality itself might be affected by low quality caption, not just QoE.

Jing: 50-year old with caption: they improve the caption, and the user is so happy even if the image is the same.

Lukas: with UPGC, for the PGC there is the issue to mess with the artistic intention of the professional. Jing: they know which users are "professional". Some "users" are AI technology, e.g., to summarize video. User = might also be algorithm.

Roberto: consider also fps (some 60, some 30).

Patrick: presenting horizontally and vertically on mobile device, there is scaling in the process.

Ioannis: they force rotation in full-screen

Zhi: experimental with vertical video before selection, but the source is horizonal. Things can be quite complicated.

Lucjan: this is interesting to test (horizontally and vertically), if tested on a mobile phone. On PC or TV, you do not have the option to rotate. So, it is really an issue with mobile.

Ioannis: scaling is always there on the phone, e.g., capture 1080p and screen is not 1080. NB: most of mobile devices is 16:9.

QACoViA Session

Chairman?

Was set-up in Madrid, Mikolaj was not attending. Pablo and Mikolaj were appointed.

Pablo has step down.

Decision: Patrick appointed as vice chair of QACoViA.

Presentation #5 by Mikolaj, AGH

Evaluation of video summarization

Infobesity (information obesity): like taking a drink from a fire hydrant. E.g. too much video on the Internet. So, summarization is needed.

Purpose: summarization and content translation (from Arabic (subtitled in English, with machine translation) to English and French)

Mainly news reports and interviews, quite short, 10-15 minutes -> to about 1 minute

4 architectures (approaches) differing for the order of components, 4 different results. Which is the best one? No metrics available, only ask people for testing.

Test: No comparison of videos, people get bored to see the same content. Already tried. So, ACR approach was used (with safety check 1 question to see if people watched the video). 1 summarized video per user per original video.

Request for participation:

Find 10 minutes to help us… http://amis.kt.agh.edu.pl

Deadline in 2 weeks (end of October 2019)

Interested expecially if you do not speak Arabic (target group of people)

Florence: target number? incentives? Mikolaj: most of the people will not evaluate many videos, but it is still useful even if you evaluate few videos

Lukas: guidance about what is a good summary? Any ways of training people? Mikolaj: maybe discard the first scores? True that the idea of a good summary is subjective. An experiment before showed there is no winning scenario but there were some bad scenarios (significantly worse).

Florence: methodology is sound, but typically user cannot see difference between SD HD…

Ioannis: there are professional editors that can do summarization, to have a sort of reference. Mikolaj: not easy to find for Arabic etc… Some trailers of movies can be used, but the content is different

Roberto: engagement: one way is to feedback the information you get (e.g. 30% of people have chosen the answer you gave etc.)

Patrick: link with QACoViA which is about quality assessment for computer vision applications? Mikolaj: Task based … and computer vision, more linked to old "QART".

Presentation #6 by Lucjan, AGH

Objective Video Quality Assessment Method for Recognition Tasks

Presents a framework for doing that:

Scenario: Entertainment, Surveillance, then going to algorithm. Typically test the accuracy of the model.

But we want a quality indicator before running the ML algorithm, to know e.g., that some image quality gives less chance to detect things.

Face recognition, Object recognition, Automatic license plate recognition.

Different DBs. Some datasets for autonomous driving. Nobody in this scenario cares about compression.

Quality indicators form AGH, and from LIVE.

Face recognition: dlib

Object recognition: YOLO trained on COCO database

License plate: openALPR

Roberto: camera correction should include rotation etc.? Not now, but could be considered.

Patrick: chromatic aberration… should be considered. Demosaicing can have strong effect.

Patrick: test also BIECON from Yonsei, probably they can give the implementation (full CNN approach), also should be in ML frameworks such as Caffe.

Pablo: distortions: 2 types: out of your control, or those that affect the system. Mixing both in the problem can cause issues. Lucjan: some detection algorithms use lower-resolution

Patrick: one thing is optimizing the system, another is the quality of the image that comes from the sensor during the operation.

Patrick: add some enhancement algorithm in HRCs

Lukas: be careful when you add effects like rain on a good image, rather than a real image with rain.

SAM

chaired by Lucjan Janowski (AGH University of Science and Technology) and Zhi Li (Netflix):

Welcome and introduction to SAM by Lucjan Janowski and Zhi Li

Presentation #13 of Zhi Li: "Overview of SAM Activities". Discussion on the findings, led by both Zhi Li and Lucjan Janowski

Break:

Introduction of Haiqiang Wang (Tencent) by Kjell Brunnström (RISE)

Group photo

SAM, continued:

Presentation #27 of Kjell Brunnström: "Multiple Comparisons and Planning Number of Test Subjects", tool presentation

Discussion on the ITU normalisation plans, between Lucjan Janowski, Zhi Li, Kjell Brunnström and Patrick Le Callet (Université de Nantes).

Decision: Liaison statement needed with ITU including ITU-T P.910, P913, P.1401 revision and ITU-R BT.500.

Presentation #10 of Lucjan Janowski and Jakub Nawała (AGH University of Science and Technology, remote): "suJSON - a uniform JSON-based subjective data format"; discussion on implementing standards

Presentation of Jing Li (Alibaba Group): "Preference aggregation using Pair Comparison: an overview"; discussion on pair comparison

Presentation of Lukas Krasula (Arm Ltd.): "Using Pair Comparison Data for Objective Metrics Training and Testing"

Presentation of Pablo Perez (Nokia Bell Labs): "Subjective Assessment of Adaptive Media Playout (AMP) for Video Streaming". Discussion afterwards with Zhi Li and Patrick Le Callet.

Presentation of Lucjan Janowski "Generalised Score Distribution"; discussion with Lukas Krasula, Pablo Perez, Saman Zadtootaghaj (TU Berlin) and Kjell Brunnström

Decision on further steps on the Liaison is deferred to the Thursday session on Liaison drafting.

Day 3: Wednesday 16 Oct 2019

AVHD

Presentation #12: Netflix – Live Research Project – Christos

Perceptual Optimization using Deep Compression Model

Lots of great conversations and questions from the meeting attendees

There is a disconnect between our community and the Compression community – where they are solely using PSNR and that we need to be more vocal around the advantages and disadvantages.

Action – there should be a liaison to the compression community from the quality community – going to MPEG/JPEG AOM being the most open to the communication and discussion around metrics.

List and methodology to those communities. Encoding tools and metrics - ITU (MPEG and AOM)

Ioannis – are you able to start to draft the text for this liaison statement – AVHD group – Kjell will be involved.

Decision – Ioannis with Kjell will produce a draft Liaison for the group to review before sending.

Lucian – would we propose that we have a website with a frozen set of metrics that people can use versus providing equations or versions of VMAF.

Ioannis – we can have a repository – but must have golden versions that are open source and functional by the community. VMAF – open source packages which include several tools but still need verification.

Pablo – there are missing participants from today's meeting around tool development and deployment. We don't want to endorse metrics that have not been validated and verified by VQEG. AVHD team can be consulted before moving forward. We want to ensure we follow the classical VQEG validation process etc..

Presentation #11: C3DVQA – From Tencent Media Lab by Haiqiang Wang

IQA and VQA work.

Great discussion – no decisions or direction taken

15 min coffee break – till 10:55am

IMG

IMG Work Plan:

Phase 1: Short Sequences

Decided July 2018 – to consider the use case of subjective quality evaluation of 360-deg video joint work was established

Ready for the first round – short sequences – Contribution to P.360-VR (not covering long seq in the test plan or complex concepts)

Reviewed the test plan plus decisions taken from the Monday meeting around source removal.

Pablo edited his document directly.

Ioannis offered devices if that was a bottle neck for Oculus device.

Data analysis will be covered at different time.

The labs will need to be willing to give the data collected to the VQEG body and publicly available and shot for a journal article and this will be agreed prior to the testing being run and who will be listed in the publications based on work completed etc.. Issues need to be sorted prior to the tests.

Everything will be on the reflector and will be discussed on the audio call.

Schedule was presented – shoot for Q2 2020 for the article submission (Kjell said that the schedule looks tight based on resources, holiday's etc.) (might need to adjust the timeline for the labs and result provisioning)

Only one meeting in SG12 which is in April 2020 – which means we will miss 2020 and hit 2021 for the contribution. Deadline for SG 12 – is April need to finish by the end of March – with data analysis and text for contribution.

Question 30 and 40 might have interim meetings SG12 meeting in Nov and the dates will be decided in Dec of 2019.

Has a list of decision points final foil of presentation. (23) – try and prepare a pres with the conclusions and that can be shared with ITU-T session tomorrow.

Decisions

We agree on the objective - will setup expectations in the training phase and then look at the results carefully – then look at if they need to make changes to the content or change the training session.

Assigned the tests for the 9 labs that are currently involved in the foils.

Tentative decision on data being publicly available – however Pablo needs to email all labs to ensure tracking of the responses yes or no and then you have a thread of the agreement for the direction of the work.

Summary of agreed parts of the test plan

- Agreed on baseline test methodology

Use of ACR and DCR
Simulator Sickness protocol and questions
Session structure and duration
Use of Miro360 app interface

- Agreed on baseline SRCs

There will be 8 SRCs
5 of them will be used for longer tests
Based on the ones from pre-test
Remove Segovia and Thieves
Maybe remove others with stitching artifacts

- Agreed on list of HRCs and their subsets
- No additional PVSs for specific test cases Agreed on test conditions

Select which ones will be tested (see next slides)
Each lab will test one test condition
Minimum number of subjects: 30 per lab
ACR methodology will look for compression artifacts
Explained in the training part so that we can search for more values equal to 5
Some modifications are needed from pre-test (see next slides)

- Agreed on exploitation of results

Journal paper and public data set
See following slides for details

- Agreed on rough schedule (tentative)
- Agreed on main test conditions

ACR: 10s vs 20s
ACR: 20s vs 30s
DCR: 10s vs 20s
DCR: 20s vs 30s
HMD vs HMD (ACR 20s)
HMD vs HMD (ACR 20s)
HMD vs HMD (ACR 20s)
With vs without audio (ACR 20s)
Scoring interface vs voice (ACR 20s)

- Pending decision: which pairs of HMDs to compare
- Agreed on extra test conditions if new labs can join

ACR: 10s vs 30s
DCR: 10s vs 30s

Conclusions of the pre-test discussion

- Remove some SRCs: Segovia, Thieves.
- Fix some PVSs which are broken
- Miro360 changes/fixes: initial orientation and selection of playlists
- Improve training session to

Include more qualities (especially non-uniform coding)
Train people better on the defects that they have to assess

- Improve pre- and post-session questionnaires

Exploitation of results

1. We will write a contribution for ITU-T P.360-VR
    1. Target date: end of March 2020 (before next SG12 meeting in April 202)
2. One journal paper

    1. With the conclusions of the study
    2. Authors: people from each participant lab which have contributed significantly to the work
    3. Editors/writers (first authors): Jesús Gutiérrez & Pablo Pérez
3. Public data set
    1. All test result files (raw data) and all PVSs
    2. Linked to the journal paper (for citation)
4. Rule: each participant lab, *before starting the test*, must
    1. Agree on the publication of the data set
    2. Propose the names of authors for the paper, and their expected contributions

Tentative schedule

- Finish preparations/modifications during October
- Try to kick-off tests by the beginning of November

Target date to have tests ready: end of December
Hard deadline: end of January

- Contribution to ITU-T ready by end of March

Next SG12 meeting in April 15th -24th

Admin

Date for next f2f meeting

- Some days in April are inconvenient
    - ITU-T: Apr 15-24
    - NAB: Apr 18-22
    - Holidays: around 12nd - 19th
- Decision: Target a week in march

Options: 2-6, 9-12, 16-20, 23-27 March 2020

[Note from email after the meeting
"Hi Kjell,

March in Seattle is still rainy and cloudy, and sometimes it even snows. I would recommend late May or early June when Seattle is in the super beautiful summer, if possible. Besides that, my calendar is open for the 1st, 2nd and 4th week in March 2020.

Please let us know whether late May or early June is preferred.

Thanks,
Yongjun"]

- Depending on the host
- Decision: Kjell will send an email to reflector for vote

Should next meeting be 4 day or 4 1/2 day?
Discussion:

- Lucjan: 4 1/2 days. Use Friday for common work: write liaisons.
- Lukas: 4 1/2 days. Friday can be kept as buffer.
- Kjell asks how many people would stay on Friday with any pre-defined agenda: about half of the audience would.

Decision: 4 and 1/2 days

Fall 2020 meeting:

- Potential candidate: Politecnico di Torino. Enrico will check the availability.

- Backup candidate: Sky

Qualinet

Kjell Brunnström. **Presentation # 20** about Qualinet activity.
All info: http://www.qualinet.eu
Description of Qualinet:

- European organization targeting QoE.
- One f2f meeting per year one day colocated with QoMEX

Achievements

- QoMEX conference
- Definition of QoE (ITU-T P.10)
- Collection of databases: http://dbq.multimediatech.cz
- Journal

Presentation of task forces

IMG

Presentation #19 by Kjell Brunnström. "Quality of Experience Assessment of 360-degree video"

- Test eye-tracker on 360 video visualization on HMD under quality degradations, freezing videos and different contexts.
- Subjective experiment
    - 3 sources, several impairments
    - Subjective assessment
    - Objective measurement based on PCA of eye-tracking statistics
    - 32 participants, 36 ratings per subject
- Analysis of subjective and objective metrics wrt to the different contents and videos
- Discussion/Q&A: Pablo, Ioannis, Patrick

CGI

Presentation #1: Saman Zadtootaghaj. "NDNet Gaming - Development of a No-Reference Deep CNN for Gaming Video Quality Prediction"

- Latency: capture RGB data from frame buffer -> GPU encoding -> Fixed MB size (32x32)
- Task-specific network protocol for reliable UDP
- Specific spatial and motion patterns
- CNN-based No-Reference metric
    - Pre-trained with VMAF, fine tune (transfer learning) with subjective ratings
- Tested in general and gaming-specific datasets
- Discussion: Christos

Presentation #2: Steve Göring. "nofu – A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content"

- Metric: lightweight handcrafted features -> temporal pooling -> selection + random forest
- Evaluation using GamingVideoSET —> good results
- Q&A: Christos, Patrick

Presentation #3: Steven Schmidt, "Updates on the ITU-T activities with respect to gaming quality assessment"

- ITU-T G.1032 (10/2017) - Influence factors on gaming QoE
- ITU-T P.809 (05/2018) - Subjective evaluation for gaming quality
- G.OMG (Q 13/12) - Opinion model for gaming applications
    - Network planning tool
    - Non-professional gamers, non-VR gaming.
- Q&A: Lucjan, Pablo

Presentation #4: Kumar Awanish, "No reference metric for gaming video content"

Day 4: Thursday 17 Oct 2019

NORM

Make Colligan: Norm

The key is not only is MOS but why we obtained this number, what to do to increase it.

Ioannis suggested we should have two classes of NR metrics. One related to case where we don't have access to the source, like user generated content. Other is compression when the source exist.

Margaret prepares a git repository with a framework supporting NR metrics. If a bitstream parser will be available in MATLAB. There is discussion should it be in MATLAB or python. It seems that python getting more attention and MATLAB is expensive which is problematic. Zhi pointed that VMAF has ability to add NR metrics to the framework already created for FR metrics.

Question: If you propose a metric you should provide videos or images with examples of distortions. Answers from Margaret: No. Florence asked if we can add synthetic impairments to CDVL content? The answer is yes, you can do any modification you need.

There is an option that Pablo will help with software development and some tests. AGH will provide their metrics and there is an option with support from students.

JND usage for NR metric development is discussed. After couple of opinions presented the final conclusion is that SAM should work on it. The first stem is identifying the publications related to the topic. Some work was already done.

Can bit stream reader should work on a file or in the live scenario? Answer: File is ok.

Decisions about next steps for NORM.

1. Calls are organized at the same time as SAM i.e. Monday at 8 am at pacific time (PST or PDT). This may change to accommodate also Asian participation. Action: Send the invitation to create a calendar invitation.

2. Agenda should be created before the NORM teleconference

SAM

Proposal for ITU

Proposals:

Subject model

Zhi: Alternative to subject rejection base it on SAM analysis. Synthetic results show that you do not have to remove subjects, but you can use MLE. We should start with simple model. We can use different error models, but we should start from simple model, with bias and subject consistency. For such models we can be sure that we can recover the data. We are not linking MLE to the p-value analysis. True quality should be an alternative to MOS. MLE takes two steps, subject rejection and calculating MOS as a single step.

Lucjan: proposed to call MLE analysis as superior to MOS.

Kjell: proposed to say exactly what we believe what should be in the recommendation.

Zhi: MLE can provide strange results by error in the conversion. In such case going back to mean can make sens.

Ioannis: Pointed that in Physics it is normal to estimate different parameters of errors we measure. We could also add comparing new method and MLE.

Pablo: There is big difference between Physics and Social science. We do not have theory from which we extract parameters. MOS goes from audio quality. The main idea is to deal better with strange subjects. We should have strong evidence that MLE has significant advantage.

Lukas: We should focus on the Zhi example with random users.

Pablo: Ok, but we have to show it very clearly what is the advantage.

Lucjan: We are not ready to send it to the recommendation. Simulations can answer the question.

Zhi: We should test it on more databases.

Christos: Advantage of MLE is that you do not have to worry about the analysis of the rejection and about the quality of subjects.

Pablo: We should exploit and better understand the advantages because MOS will be in any subjective experiment paper anyway. People will add something only if it is clearly better.

Work to do:

1. Synthetic data to test the recovery

2. Use synthetic data to test the confidence interval, especially with random voters (with different proportion of random votes)

3. Show evidence on the real data

4. Propose a tool which is easy to use so the users do not need to install MLE

Framework for Objective Metric Evaluation

Lukas: It will be liaison to inform that VQEG will propose it before the April ITU meeting.

Lucjan: We should add more simulations with synthetic data.

Kazuhisa: P.1401 is still open.

Action: Zhi will draft a Liaison for sending to ITU in Nov 26 meeting of the intention to submit contribution for the April ITU-meeting, based on the MLE-work.

Action: Lukas will draft a Liaison for sending to ITU in Nov 26 meeting of the intention to submit contribution for the April ITU-meeting, based on the *Framework for Objective Metric Evaluation*

Database evaluation

Tentative decision: Phil - We can have it under ILG if Margaret is ok with it and AGH will contribute with student resources.

IMG

Open session at 2:00

ITU-T Q13/SG12

Presentation from Rachel Huang

G.QoE-VR

Provided an overview of the document sections.

Request for any feedback on the structure of the recommendation.

The document will be shared with Pablo who will share to the VQEG IMG reflector.

The deadline for the conribution is November 13th.

Therefore comments must be submitted by the end of October.

5G KPI

Presentation #15 from Pablo Perez.

Deconstructing AR applications for 5G

Discussion of types of 5G networks.

Discussion of AR applications.

This project focuses on 5G Ultra dense networks.

QoE = f(KQIs) where KQI = key quality indicators.

Discussion of an example of Microsoft HoloCall.

This will serve as the AR app for deconstruction.

Shared interest with Q13/SG12.

Closed session at 3:20.

AVHD

AVHD P.NATS Phase 2 Project

Opened session at 4:10

Presentation from Shahid Satti

Description of the project and timeline.

Plan to submit the model winners to SG12 in November.

Description of competition and model performance evaluation.

Description of results: 5 models met performance criteria, 4 did not.

Results will be detailed in the SG12 submission.

Closed session at 4:50.

eLetter

Presentation #28 by Kjell Brunnström. Description of eLetter issues and process.

Looking for a group for the next issue.

Suggestion that it could be SAM or CGI. CGI tentatively accepted

Decision: Also looking for a volunteer to cut the eLetter PDF into chapters for indexing.

Mikolaj will volunteer.

Minutes were approved.

Kjell thanked the host once again and closed the meeting.