VIDEO QUALITY EXPERTS GROUP

Progress report 2013

Copyright Information

http://www.vqeg.org

For more information contact:

Arthur Webster	webster@its.bldrdoc.gov	Co-Chair VQEG
Kjell Brunnström	kjell.brunnstrom@acreo.se	Co-Chair VQEG

Introduction

VQEG's work during 2013 was very productive. The Hybrid project neared to a conclusion with all of the subjective testing and model evaluations almost finalized. A VQEG eLetter was started and the first issue will be published in January 2014.

Since the three most recent meetings have been in December 2012, July 2013, and January 2014, there was only one face-to-face meeting in 2013. This was the Ghent, Belgium meeting hosted by Ghent University iMinds. This 5-day meeting had 27 (+5 remote) participants and handled 50 documents. (see http://www.its.bldrdoc.gov/vqeg/meetings/ghent,-belgium-july-8-12,-2013.aspx). In addition to the face-to-face meeting, several VQEG Groups held numerous conference call meetings and worked extensively via email.

VQEG also started a LinkedIn group in 2013 which now has over 100 members. The VQEG's main reflector has around 500 members. VQEG has 11 email reflectors in all.

VQEG has currently 9 active projects and 4 support groups. Their activities and progress during 2013 will be described below in this report.

Active Projects

Audiovisual HD Quality (AVHD)

Co-chairs: Margaret Pinson (NTIA/ITS), Chris Schmidmer (Opticom), Quan Huynh-Thu (Technicolor)

The AVHD Quality project seeks to benchmark quality metrics suitable for various cases including:

- Video-only quality in HD resolution
- Audio-visual quality of HD videos with accompanying sound
- Adaptive video streaming quality assessment methods

The AVHD group is currently in the middle of defining appropriate test plans for the benchmark test. A first overview can be found in the project synopsis document of the former multimedia 2 project: This document is specific for audio-visual quality. Except for the audio component, most of the content will be valid for the video-only metrics as well.

The Audiovisual HD group also investigates improved audiovisual subjective quality testing methods. This effort led to Consented Draft Recommendation ITU-T Rec. P.913, which is currently under AAP. The AVHD project is now examining an immersive method for audiovisual subjective testing. Presentations on this topic are encouraged at all VQEG meetings.

High Dynamic Range (HDR)

Co-chairs: Phil Corriveau (INTEL), Patrick LeCallet (IRCCyN)

Proposal on to produce tone-mapped PVSs, and run previously validated FR models on them. There is a problem: what should be used as the original video? The desire is to use this as a starting point (e.g., model in its entirety, or use individual parameters—still to be determined).

Hybrid Perceptual/Bitstream project

Co-chairs: Jens Berger (SwissQual), Chulhee Lee (Yonsei University)

The goal of the hybrid project is to develop hybrid models (FR, RR, NR), which use bitstream data and the decoded video signal as input. The models were designed to perform accurate and fast measurements for quality monitoring of various video services. Four proponents submitted models and the validation process is under way and expected to conclude in early 2014.

During weekly scheduled audio calls remaining details about the evaluation data sets were discussed. Ten data sets were processed and exchanged, and the corresponding subjective experiments were performed. Subjective data and model predictions are available for model validation.

It is expected that a draft final report will be available end of January 2014.

JEG-Hybrid

Co-chairs: Marcus Barkowsky (IRCCyN), Lucjan Janowski (AGH University), Nicolas Staelens (Ghent University-iMinds-IBCN)

10000 H.264 coded video sequences database available on the ftp server. Amy Reibman has joined the group and will probably work together with Lucjan Janowski to learn about which of the 10000 video sequences can be trusted to be evaluated by objective measurement and which ones need to undergo subjective assessment.

The current existing database with encoded H.264/AVC video sequences is being extended with HEVC encodings. This also includes running different video quality measurement tools on the generated PVSs.

The wiki (wiki.vqeg-jeg.org) has been migrated and is now being administered by Ghent University - iMinds.

Kongfeng Zhu proposed a new methodology for wavelet-based No-Reference Image and Video Quality Estimation. She is planning to present her work during the JEG-Hybrid session at the Boulder VQEG meeting (Jan 21-24).

Enrico Masala proposed a new method for impacting video bitstreams by removing NAL Units directly in the decoder which allows for the decoder not to crash. He is going to present his work during the JEG-Hybrid session Wednesday morning during the Boulder meeting.

Marie-Neige Garcia proposed to contribute a Python implementation of P1202.2.

Monitoring of Audio Visual Quality by Key Indicators (MOAVI)

Co-chairs: Silvio Borer (SwissQual), Mikolaj Leszczuk (AGH University), Emmanuel Wyckens (Orange Labs)

Current work of MOAVI includes the topics and results below:

- Implementation of 7 metrics for following artifacts:
  - Blockiness – the probability of correct classification: 98.48%
  - Blur – the probability of correct classification: 80.52%
  - Exposure time distortion
  - Noise
  - Framing
  - Freeze
  - Blackout
- Initial values of thresholds for particular metrics were settled
- Development of metrics for audio artifacts (mute and clipping) in Matlab environment
- Development of metrics for block loss and interlace artifacts in Matlab environment
- Preliminary tests of subjective opinion with the purpose of improving the approach to thresholds
- Design and construction of the website where the metrics are publicly available (vq.kt.agh.edu.pl)
- Writing paper regarding MOAVI project for SIGCOMM conference in Hong-Kong and VPQM conference in Arizona
- SIGCOMM and VPQM conferences reviewers have provided some feedback comments that should be analysed and taken into account for future steps of MOAVI project. The most important weakness detected is the lack of any presentation of actual results in the articles, although there is a set of metrics of artifacts ready.
- Therefore, a set of video and audio files has been created to test the metrics developed in previous months (Mute, Clipping and the Voice Activity Detector). These results of the metrics on those videos are ready to be compared with some ground truth determined by the researchers or eventually the results coming from subjective results.
- In the case of the Voice Activity Detector particularly, its accuracy detecting the voice activity in the audio clips extracted from the database has been measured comparing the results obtained from the detector with the ground truth determined by both the observation of the waveforms and the listening of the sound.
- The metric to detect the Lip Activity from the videos has been enhanced during this month and the results of the temporal activity in the region of the mouth for the videos of the database have been stored for its future analysis. The main goal of the latter is being the establishment of a threshold to consider the video frame as “lip active” or not.
- A set of test videos has been created with the following characteristics:
  - Frontal view of talking faces.
  - Duration around 20 s.

- Real delay introduced to make the tests compared with the delay detected by the metric:
    - Average deviation = 130 ms.
    - The metric discriminates positive and negative delays.
- For the supercomputing cluster calculations we had to move the Temporal Activity and Spatial Activity metrics to C++, which we think, may also belong to the small progress in the MOAVI project.
- Also solely to create all databases with the results of the MOAVI project metrics require the use of the project applications, which can be considered as a solid test (for a total of more than 7500 videos).

Below, the results of key indicators verification test are presented. For each metric the test consists of two parts: one is setting of the threshold of distortion visibility; the second is key indicators checking process. Before the test the results of subjective experiments are randomly split into two independent sets for each part of the test. These two sets are training set and the verification set respectively.

Setting metrics threshold value

For each metric the procedure of determining visibility threshold includes the following steps:

1. For all video sequences from the appropriate subjective experiment the values of the metric is calculated.
2. We assume each successive value of metric as candidate thresholdsth_{TEMP}. For values less than th_{TEMP} we set the key indicator to 0 and to 1 for same or above.
3. For each th_{TEMP} we calculate the accuracy rate of resulting assignments. It is the fraction of key indicators, which match with indications given by humans from the training set.

$$accuracy(th_{TEMP}) = \frac{number\ of\ matching\ results}{number\ of\ results}$$

Eq. 1

4. We set the threshold of metric to the candidate th_{TEMP} with the best (maximum) accuracy. In the case of several th_{TEMP} values with the same accuracy we select the least value.

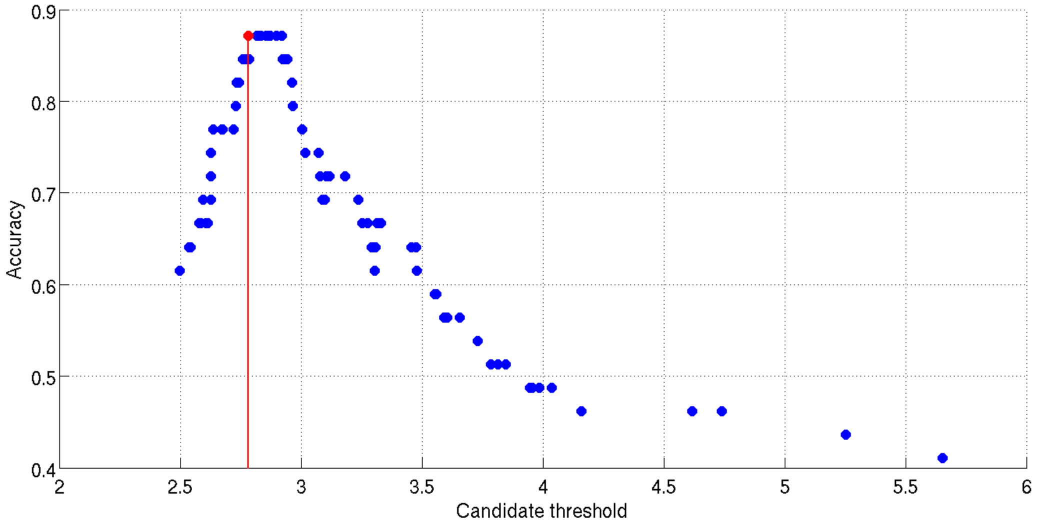


Fig. 1 Blur metric threshold determination. Points represent the relation between candidate thresholds and accuracy. The line is drawn at the best candidate, which is chosen to be the metric threshold

Fig. 1 illustrates the procedure of determining threshold for the blur key indicator. The threshold values are shown in Tab. 1.

Key indicators verfication

In the second part of the test the correctness of the key indicator is checked. Accuracy of indicator is calculated according to Eq. 1, compared with indications from the verification set. Tab. 1 presents the verification results.

Tab. 1 Key Indicators verification – probability of distortion detection

Metric	Probability of distortion detection	Value of threshold
Blur	0.86	2.78
Exposure Time Distortions	0.81	78 and 178
Noise	0.85	3.70
Block loss	0.84	5.3
Blockiness	0.94	0.85
Freezing	0.80	0
Slicing	0.85	7

Quality Recognition Tasks (QART)

Co-chairs: Joel Dumke (NTIA/ITS), Mikolaj Leszczuk (AGH University)

As the project's objective is to develop statistical models that will be able to estimate the quality of a video with regards to its usefulness in discerning visual information, for this purpose, some data must be gathered, which includes results of automatic quality assessment and recognition rates from experiments involving humans. This part of the project concentrates on performing calculations with objective measures and writing the results to a database.

A massive database has been created, storing several parameters coming from two different experiments: recognition of license plates, vehicle maker and vehicle colour in a parking lot, and recognition of different objects in different scenarios. Both experiments have been converted into a common format, as far as possible. In total, there are:

- 126 SRCs -> original video sequences
- 40 HRCs -> sequences modified (cropping, compression...)
- 1860 PVS -> derived clips
- 193 subjects answered in all experiments
- 69236 answers (data: rows in the database)

This database is available in order to get parameters from there (such as bitrate, resolution...), and also to update it with other parameters (which need to be calculated).

Apart from that, all this data has been converted into a single numerical matrix in Matlab which will be the base for starting the modelling process.

Quality evaluation must focus on the areas where essential information is found. Some work has been done on following selected objects. It can be achieved in quite a simple way when standard foreground determination is possible. However, the method should also enable tracking in case of unstable background, and the objects may be nested in larger ones. For this reason, another approach tried was the use of optical flow. The implementation of this method found in Matlab did not ensure sufficient accuracy. A more sophisticated algorithm is needed to reliably accomplish the task. Probably the best way to overcome the problem is to use a function from the OpenCV library.

There was also some attention paid to additional tools, like Kalman filter, which will be used to improve the accuracy and correct random errors. Methods of keeping track of multiple interacting objects were also analysed in case they are needed in the future.

To connect Matlab code to database, external Java driver had to be installed and some configuration done.

Available data about target recognition results is heterogeneous. Its structure is not certain at the moment, so the code for accessing it could not be completed. However, some of the design questions have already been resolved.

All the data available since the realization of the experiments (single values), i.e. answers from the viewers for every PVS (SRC-HRC), bitrate, resolution, scenario, etc. has been joined in a Matlab structure to the results of having run the NR and FR metrics (single value for every frame -> arrays for every PVS). Three different sub-experiments (not in laboratory, at AGH University and by not practitioners respectively) from the objects recognition subjective experiment have been added to the two original sub-experiment (both in laboratory) included in the data. Taking into account the parking lot experiment, there is a total of 6 different experiments. Thereby, in just one Matlab variable all the data required for the modeling process is accessible. This structure has a total of 69236 sub-structures with

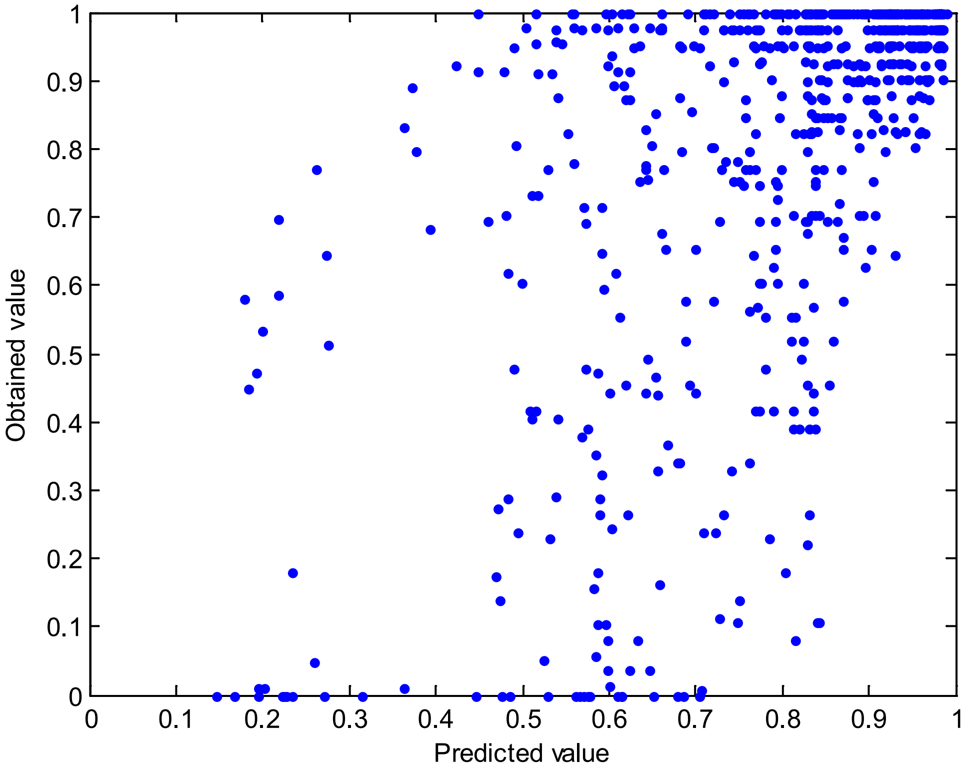
23 fields each one: SN, SRC, HRC, recognized object, original object, bit rate, color, make, file size, Levensthein distance, viewer, resolution, scenario, correct, experiment, incomplete, blockiness, blur, exposure, spatial activity, temporal activity, SSIM and VIF.

Analyzing the likelihood of correct recognition from the 6 different experiments, the first decision taken has been to start searching for a model just for the 3 experiments which have similar probability of correct recognition (around 80%).

First calculations (means, medians, and quantiles for NR and FR results) and conversions (creating arrays storing the answers for every PVS instead of doing it for every viewer) have been done to make easier to find the first model. Afterwards, just the mean will be used.

The first results showed us that there is no a clear correlation between any of the metrics used and the probability of correct recognition. This is the first conclusion exposed: better unique metrics should be used to obtain direct results.

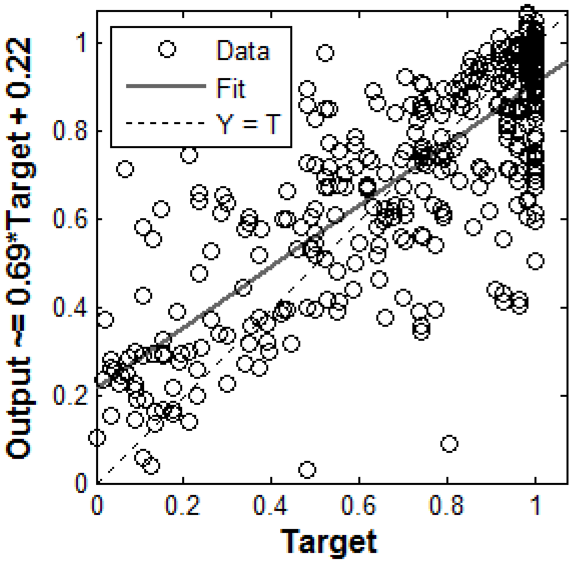
Trying to create a generalized linear model was the next step. For this task, it was necessary to decide which set of videos was going to form the training set and which one the test set. Some linear models, starting from the most basic one, were tried. However, even the results from the most complicated of the models were not good at all. This model was taking into account every parameter, the square number of every parameter, and the total multiplication of all of them (both single and square). The results obtained were the following:



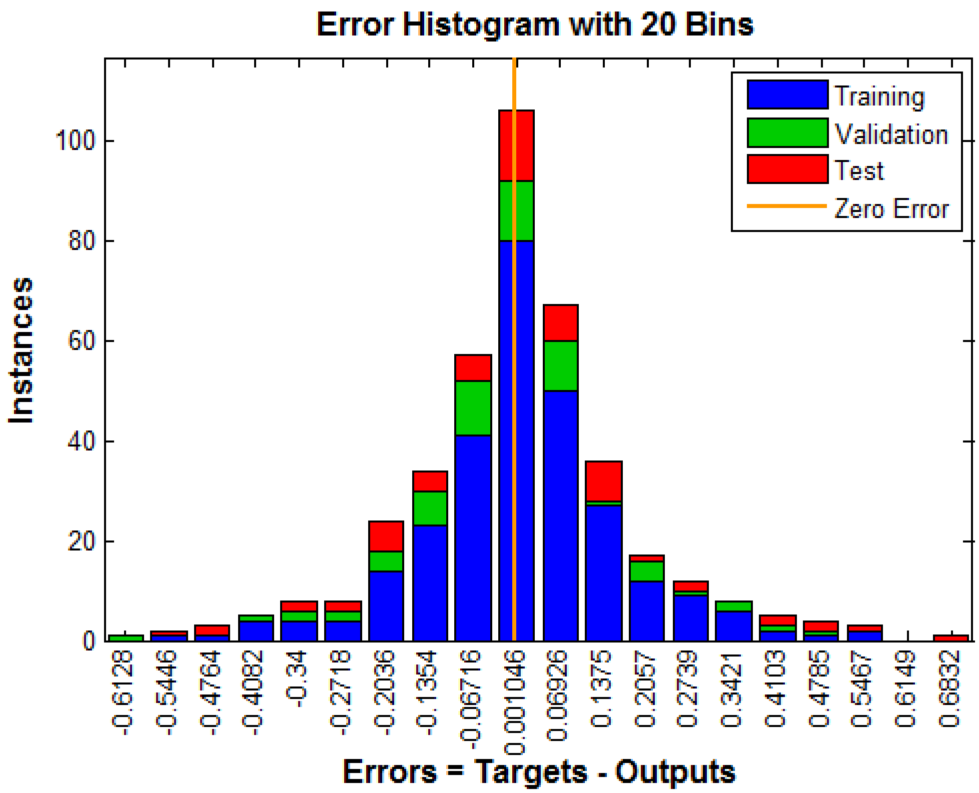
As this model was totally unsuccessful, another approach was tried: neural networks. After running a program which calculated which neural network (set of parameters and number of hidden neurons used) was the one providing the best coefficient of determination R^2, the best model found is the following:

- Parameters used: Blur, Exposure, Spatial activity (SA) and Temporal activity (TA).
- Network structure, number of neurons in every layer: 7 -8-1 (input layer – hidden layer – output layer respectively)

This combination provided R^2 = 78.7%. The results obtained with this models are the following:



The following figure shows the error histogram for this model:



The main inconvenient of neural networks is its instability.

Real-Time Interactive Communications Evaluation (RICE) project

Co-chairs: Kjell Brunnström (Acreo), Ulrich Engelke (CSIRO), Dave Hands (Skype/Microsoft)

A meeting was held between the RICE Co-chairs and ITU-T Q10/12 Rapporteur Gunilla Berndtsson (Conferencing and telemeeting assessment) and her colleague Mats Folkesson, June 3 2013, at Acreo. It was recognized that RICE and Q10/12 could benefit from collaboration and most interesting in the short perspective was to continue develop the subjective testing methodologies.

Kjell Brunnström presented methods for depth-based tasks during the Ghent meeting for 3D video conferencing.

A new Co-Chair Dr Ulrich Engelke CSIRO (www.csiro.au), the Commonwealth Scientific and Industrial Research Organisation, was elected to the project in Sept 2013. The exact focus of the project is still under discussion and how it can benefit from CSIRO's involvement. A couple of audio calls have been held for discussing this.

Ultra HD

Co-chairs: Vittorio Baroncini (FUB), Naeem Ramzan (UWS)

A new Co-Chair Dr Naeem Ramzan (UWS) was elected to the project in July 2013. Three activities are defined within the scope of Ultra HD project: (1) Creation of Ultra HD database, (2) Defining subjective quality testing methodologies for Ultra HD, (3) Objective video quality metrics for Ultra HD.

1. Creation of Ultra HD database: Currently Ultra HD has focused on the creation of a data base of 4K Raw videos. We have a plan to enrich the database to 10 UltraHD video sequence which can be encoded with different encoding parameters of HEVC. The encoded sequences will be uploaded to VQEG ftp server. This database will be described and will be freely available for download from VQEG ftp server.
2. Defining subjective quality testing methodologies for Ultra HD
   Work will be carried out in second half of 2014
3. Objective video quality metrics for Ultra HD
   Work will be carried out in second half of 2014

3DTV

Co-chairs: Marcus Barkowsky (IRCCyN), Patrick LeCallet (IRCCyN), Quan Huynh-Thu (Canon, CiSRA)

Development of subjective assessment methods was discussed and a presentation is planned.

Ground Truth Dataset – Preparation of distribution of Pairs and Common Set analysis by Jing Li using a newly developed subset selection method. A presentation in the 3DTV session is planned for the Boulder meeting.

Coding and Spatial Degradation Dataset: One more experiment performed by Chaminda Hewage from Kingston University.

Support Groups

Independent Lab Group (ILG)

Co-chairs: Phil Corriveau (INTEL), Margaret Pinson (NTIA/ITS)

ILG is focused on assisting the Hybrid effort.

Joint Effort Group (JEG)

Co-chairs: Alex Bourret (IP-label), Kjell Brunnström (Acreo), Patrick Le Callet (IRCCyN)

Promotes the idea of joint collaboration within VQEG. Discussions are underway on how to increase visibility. Proposal is to change VQEG group names to reflect whether or not the effort is currently collaborative, through a "JEG-" prefix

Tools and Subjective Labs Setup

Co-chairs: Glenn Van Wallendael (Ghent University-iMinds-Multimedia Lab), Nicolas Staelens (Ghent University-iMinds-IBCN)

A tool for lossless transmission of video bitstreams using H.264 has been publicly made available (sourceforge: definitely_lossless). The feature of this tool is to ensure that the encoding and the decoding leads to lossless reconstruction of the YUV422 video sequence which is assured by using a hash value (SHA512). This tool has been used successfully in the Hybrid project to exchange PVS while significantly reducing the required data transfer volume. See the VQEG website for available tools.