VIDEO QUALITY EXPERTS GROUP

Progress report January – June 2013

For more information contact:
Arthur Webster webster@its.bldrdoc.gov Co-Chair VQEG
Kjell Brunnström kjell.brunnstrom@acreo.se Co-Chair VQEG

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 either:

- Video-only quality in HD resolution, or
- Audio-visual quality of HD videos with accompanied sound

The AVHD group is the direct successor of the former individual projects HDTV 2 and Multimedia 2. The scopes of the two projects are very similar and a lot of synergies between the former projects is expected. The new group AVHD with a broadened scope has therefore been established.

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 Audio-visual HD group also investigates improved audiovisual subjective quality testing methods. This effort may lead to a revision of ITU-T Rec. P.911. 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 individual parameters)

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 bit stream data and the decoded video signal as input. The models were designed to perform accurate and fast for quality monitoring of various video services. Four proponents have submitted models and the validation process is under way.

During regularly scheduled audio calls remaining details of the evaluation procedure were discussed. The source scene pool selection is finished, as is the test design for the databases.

Almost all databases have been processed and exchanged. Subjective testing started in June, 2013. The subjective tests for 7 out of 11 databases are expected to be finished before the meeting. The remaining

subjective tests are expected to follow soon after. Therefore, it is expected that validated models will be determined this summer.

JEG-Hybrid

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

JEG Hybrid has focused on the creation of a large scale data base. This data base contains currently 10 SRC sequences which have been encoded with more than 1000 different encoding parameter combinations, using both JM and x264. This database has been described and is freely available for download at: ftp://ftp.ivc.polytech.univ-nantes.fr/VQEG/JEG/HYBRID. Details can be found in the test plan (ftp://vqeg.its.bldrdoc.gov/Documents/Projects/jeg/jeg_hybrid_evaluation_testplan_june2013.zip). Several different video quality measurement tools are currently run on these 12900 videos and results are stored in a simple tabular format to allow easy access. The goal of the next project phase is to analyse the results of the objective metrics in order to learn for which video sequences the quality can be sufficiently well predicted using objective algorithms and which sequences are particularly difficult and thus require subjective evaluation. Fusion algorithms will be evaluated to learn about the possibility of combining several different algorithms. In parallel, the XML files for bitstream analysers (HMIX2) have been created and further progress on a Hybrid metric is expected.

Monitoring of Audio Visual Quality by Key Indicators (MOAVI)

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

- 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).

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.

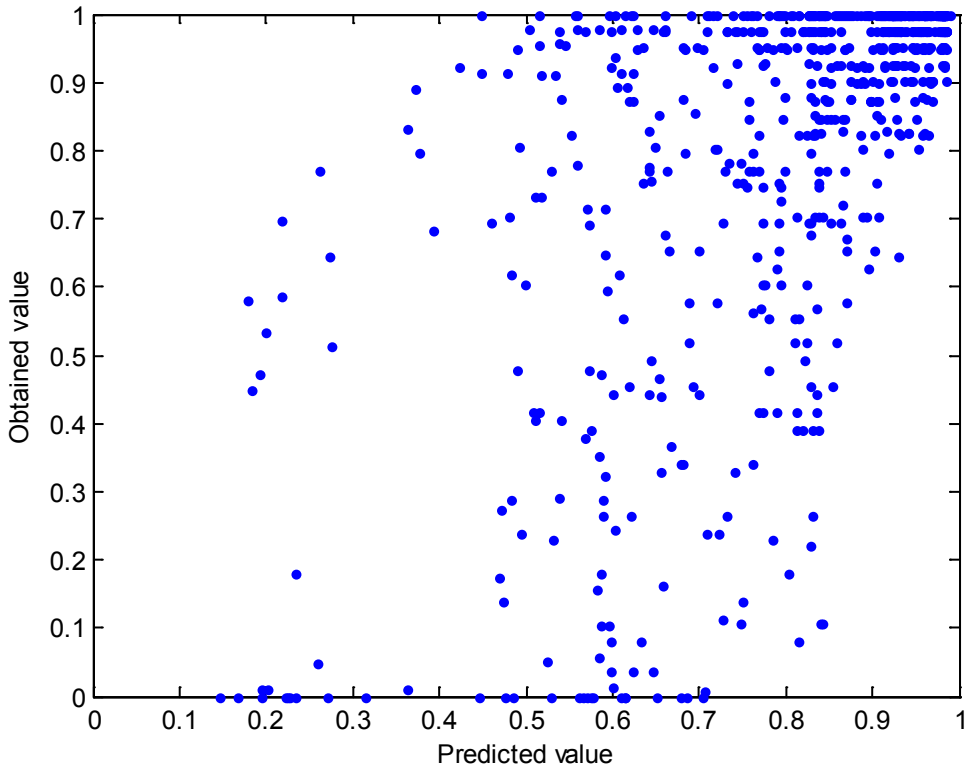
All the data available since the realization of the experiments (single values), i.e. answers from the viewers for every PVS (SRC-HRC), bitrate, resolution, scenario, etc. has been joined in a Matlab structure to the results of having run the NR and FR metrics (single value for every frame -> arrays 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:

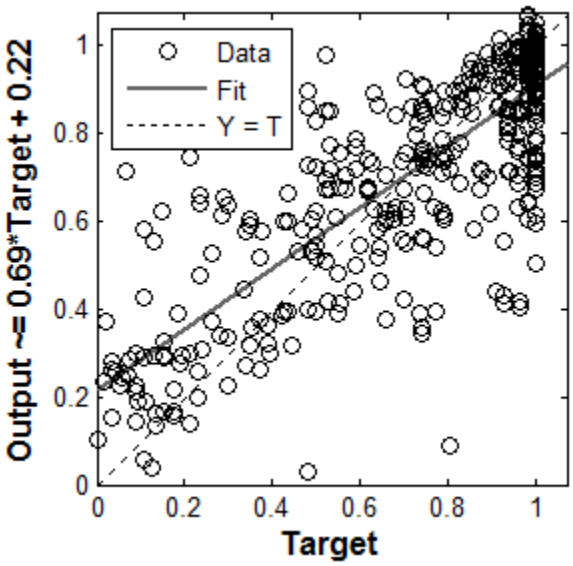


As this model was totally unsuccessful, another approach was tried: neural networks. After running a program which calculated which neural network (set of parameters and number of hidden neurons

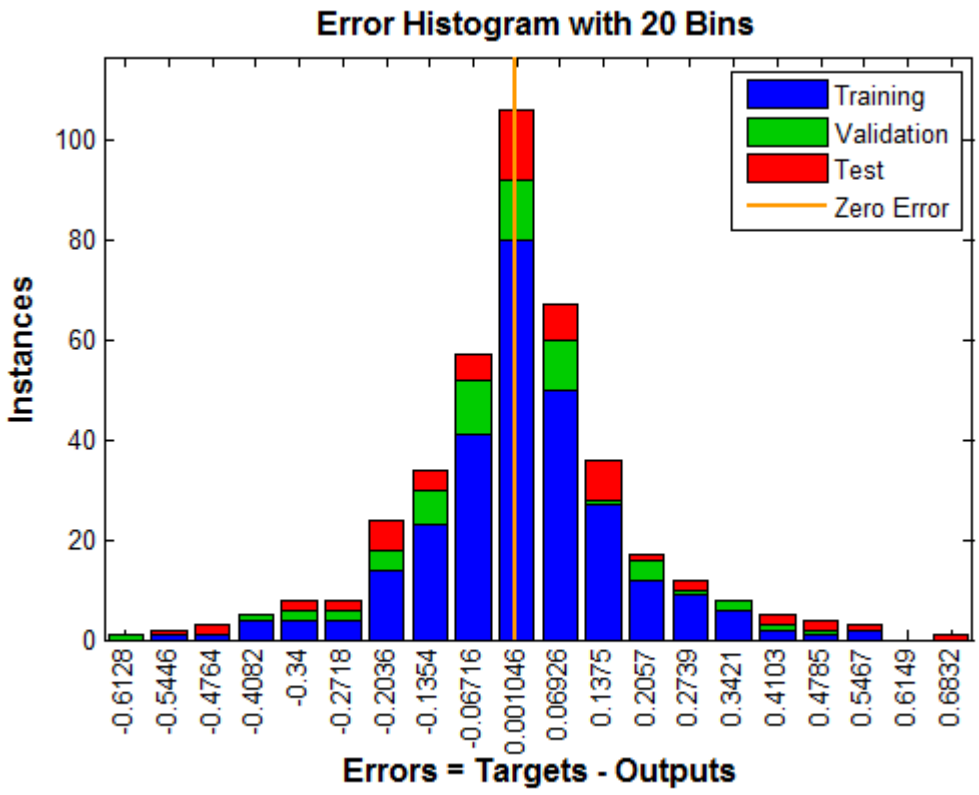
used) was the one providing the best coefficient of determination R^2, the best model found is the following:

- 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), 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.

Ultra HD

Co-chairs: Vittorio Baroncini (FUB), Osamu Sugimoto (KDDI Labs)

Project started end 2012

3DTV

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

Has three activities: (1) viewing environments, (2) ground truth for subjective testing methodologies and scales, and (3) objective model validation. 3DTV subjective tests to address (1) and (2) are being run in different laboratories, under the coordination of IRCCyN.

The current status of these activities is as follows:

1. Viewing environments:
    - The purpose is to investigate the influence of viewing environment, test set-up and display equipment on subjective quality:
    - A unique set of 3D test sequences (COSPAD dataset) is currently used by different research labs
    - Each test lab studies variables related to viewing environment/display equipment, and report experimental results for comparison with other labs' results
    - First datasets are available and have been presented, further work regarding the influence of viewing distance and various 3D reproduction technologies are ongoing.
2. Subjective quality testing methodologies and scales for stereoscopic 3D video:
    - The first step is to conduct a large-scale experiment using the pair-comparison methodology, as it is expected that human observers can provide easily and confidently a judgment of overall preference.
    - The second step will use the results of the pair-comparison testing as a ground truth database to investigate which more time-efficient subjective testing methodology and/or scales can be used to predict the results of the pair-comparison test, as well as which perceptual dimensions can be judged.
3. Objective video quality metrics for stereoscopic 3D:
    - A first step will evaluate only full-reference metrics for picture quality of S3D video: 3D viewing experience is related to several perceptual dimensions such as visual quality, depth rendering (depth quality, depth quantity), and visual comfort. This first phase of evaluation will focus on the visual quality dimension and evaluate full-reference media-layer metrics (use of decoded pixel information) to predict/monitor this visual quality at the head-end.
    - This work will progress in parallel with the work on subjective testing methodology.
    - Discussions currently cover the development of a test plan and terms of reference

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.