VQEG eLetter
Volume 1, Issue 2 • December 2014

Contents[1]

Stockholm VQEG meeting, July 2015

Storyboard

An Era of Change

Philip Corriveau

Changes in leadership and younger membership drives VQEG forward.

A new chapter for the Video Quality Experts Group, starts with changes in leadership and increased younger membership. I am Philip Corriveau, one of three remaining founding members of VQEG. It has been over 15 years of research results and excitement since the founding of VQEG. The last meeting in Sweden marked the start of a transition for this group. Arthur Webster announced, and it was ratified that Margaret Pinson would join the senior leadership as a Co-Chair of VQEG. During this transitional period, Arthur will migrate to a new position (yet to be defined) where he will continue to contribute to VQEG success. I would like to congratulate Margaret on this new role in VQEG.

Margaret has been a Co-chair with me on ILG since she joined VQEG and has driven results generation in the team for years. She joins Kjell Brunnström, who has been in a Co-chair position for several years, keeping us all focused and productive. I personally have known Margaret for many years and have worked with her on many pieces of research. I am personally very excited about what she brings to the senior leadership table.

I would like to thank Arthur for being a solid fearless leader for all these years, during many of which I was co-chair with him. Without his position and ability to navigate the standards bodies we would not have the strong group of participants from Industry, Government and Academia.

Philip J. Corriveau is a Principal Engineer in Experience Development and Assessment in SMG at Intel. He now directs a team of human factors engineers conducting user experience research across Intel technologies, platforms and product lines. He was a founding member of and still participates in VQEG.

Another great change that has been developing over the last few meetings is the growth of younger membership within VQEG. These new members to the group are unique in the sense that their perspective on the problem spaces we need to tackle are tactically different than the current mind set of those of us who have been here for a while. Another milestone for this meeting was the participation of technical women in the group. I personally find it gratifying see more and more technical woman driving forward these specialized fields in engineering and psychology.

All of this to say: the future of VQEG is bright and I encourage you all to come and join us as we move the needle on Quality of Experience forward.

Issue Overview: Verification and Validation

Kjell Brunnström and Marcus Barkowsky, Editors

Verification and validation of subjective and objective video quality assessment are two very important aspects that strongly interact with the topics of the first issue of the VQEG eLetter which focused on "best practices" for training sessions during a subjective video quality test. Verification and validation is an often neglected part when presenting new or improved methods in scientific methods. VQEG has worked on this topic since its start and would in this issue give an overview of recommended good practices, but also new and interesting ways to further improve the process. We are proud to present a number of excellent contributions on the topic.

"QoE Models' Performance Evaluation" by Dr. Irina Cotanis starts the issue out by presenting state-of-the-art hands-on methods that are already available and standardized in the ITU-T Recommendation P.1401. This is an important milestone for the area to formalize a set of statistical based tools that

should be the basis for every serious evaluation of objective metrics,

"Strategies for testing image and video quality estimators" by Amy R. Reibman introduces a new way of studying, identifying, and isolating the shortcomings of video quality estimators, by introducing a three-stage testing strategy for evaluating the accuracy and effectiveness of them.

"Dreamed about training, verifying and validating your QoE model on a million videos?" by Glenn Van Wallendael, Nicolas Staelens, Enrico Masala, Lucjan Janowski, Kongfeng Berger, Marcus Barkowsky describes a fantastic effort and a valuable resource for further testing objective quality estimators, building up a huge database of processed video sequences consisting of almost 60000 sequences.

"Validation of reliable 3DTV subjective assessment methodology - Establishing a Ground Truth Database" by Jing Li, Marcus Barkowsky and Patrick Le Callet describes another ambitious project in establishing a ground truth database for stereoscopic 3D video. This is of fundamental importance if we are going to understand the multidimensional aspect of quality of experience of 3D video and its reliable subjective and objective assessment.

"Reliably combining quality indicators" by Adriaan Barri, Ann Dooms, and Peter Schelkens discusses an often neglected topic: Selecting and fusing quality indicators for objective video quality estimators in a reproducible and reliable way using Machine Learning. They present the new concept of Locally Adaptive Fusion that put strict regulations on the machine learning behavior.

"T1A1 Validation Test Database" by Margaret Pinson and Arthur Webster documents a subjectively annotated dataset that is freely available on the Consumer Digital Video Library. Thoroughly prepared and conducted in 1993-1994, it offers

today the opportunity to test video quality estimators for their universal applicability.

"Multimedia Quality of Experience for Target Recognition Applications" by Mikołaj Leszczuk and Lucjan Janowski highlights that Quality of Experience is not universally applicable. They describe in detail the ongoing efforts in Target Recognition Video concerning challenges, available databases, standardization, and subjective assessment in this particular context.

"A New Subjective Audiovisual & Video Quality Testing Recommendation" by Margaret H. Pinson and Lucjan Janowski introduces the new ITU-T Rec. P.913 which focuses on the separate and combined subjective assessment of audio and video media in controlled or uncontrolled environments. Special emphasis is given on best practices.

"New ITU-T Rec. P.1302 for Audio and Audio-visual Call Quality Testing" by Sebastian Möller and Benjamin Weiss briefly explains the advantages and use cases of the new ITU-T Rec. P.1302, notably the focus on the content instead of the transmission quality while including time-varying transmission channels.

"Blind Image Quality Assessment: Unanswered Questions and Future Directions in the Light of Consumers Needs" by Michele A. Saad, Patrick Le Callet and Philip Corriveau describes an exciting innovative approach to holistically analyze the quality at the human receiver by considering isolated influence of complex interactions of each prior step such as intent, capture, conversion, transmission, and display in real-world consumer terms, a mission for a new workgroup within VQEG.

Issue Editors

Kjell Brunnström, Ph.D., is a Senior Scientist at Acreo Swedish ICT AB and Adjunct Professor at Mid Sweden University. He is an expert in image processing, computer vision, image and video quality assessment having worked in the area for more than 25 years, including work in Sweden, Japan and UK. He has written a number of articles in international peer-reviewed scientific journals and conference papers, as well as having reviewed a number of scientific articles for international peer-reviewed journals. He has supervised Ph.D. and M.Sc students. Currently, he is leading standardisation activities for video quality measurements as Co-chair of the Video Quality Experts Group (VQEG). His current research interests are in Quality of Experience for visual media in particular video quality assessment both for 2D and 3D, as well as display quality related to the TCO requirements.

Marcus Barkowsky received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the Image and Video Communications Group at IRCCyN at the University of Nantes in 2008, and was promoted to associate professor in 2010. He currently co-chairs the VQEG 3DTV and Joint Effort Group Hybrid activities.

QoE Model Performance Evaluation

By Dr. Irina Cotanis

Initiated during the VQEG Multimedia QoE Models project, then extensively refined, tested, and validated during ITU-T SG12 POLQA, P.NAMS, and P.NBAMS projects, the ITU-T P.1401 recommendation uses state of the art statistics to define methods, metrics, and procedures for the statistical evaluation, qualification, and comparison of objective quality prediction models, regardless of the assessed media type—e.g., voice, video-audio/multimedia. The recommendation describes an evaluation framework, provides guidance on model selection, and discusses special use cases.

Voice and video-audio (multimedia) QoE modeling experts contributed throughout the years to the development and continuous improvement of a stable and self-sustained statistical evaluation procedure for QoE model comparison. The final work resides with the ITU-T P.1401 recommendation, released in July 2012.

Evaluation Framework

Evaluation framework assumes that subjective tests in place are taking into consideration all new types of degradations that have emerged from a rapid technology evolution, one that brings with it a large variety of multimedia services which impact users more and more in a non-traditional way (e.g., re-buffering effect for multimedia streaming). In addition, it is assumed that aspects related to objective models, such as model type (e.g. parametric, perceptual), evaluation scope (e.g. comparison between

Based on well-established aspects related to both subjective tests and objective model development, an evaluation framework covers data preparation techniques, analysis types, numeral scale predictions, statistical evaluation metrics, and evaluation metrics' associated statistical confidence and significance.

models or against pre-defined performance thresholds), and application type, are well-defined prior to the evaluation process.

Data preparation

Known to drastically impact the evaluation results, the content of the databases is recommended to cover conditions related to the main scope of the QOE models (e.g., network design/deployment, performance evaluation and/or monitoring) as well as simulated conditions specific to the network's design/deployment life phase and the real live recordings required by the evaluation/monitoring phase. In addition, each experiment should contain conditions with quality levels that uniformly cover the 1-5 MOS scale. A thorough cleansing that removes unexpected subjective outlier scores ensures the quality of the databases.

Analysis types

There are four main analysis types that are dependent on the application and model types. Analysis per individual experiment and across multiple experiments are required regardless of the application or the model type. Analysis per media sample is necessary for live recorded databases, while per condition analysis is needed in the case of simulated databases. However, for live recorded databases, a recorded sample can equate to a field condition.

Prediction on a numerical quality scale

Prediction on a numerical scale is a determining factor of the accuracy of the QoE models' evaluation and involves the following relevant topics:

- The comparison of MOS values from different experiments
- The scale calibration of a QoE model

- The compensation for variance between subjective experiments in the evaluation process

The systematically observed differences between MOS scores from different experiments, even when the experiments followed the same guidance, can be grouped into three problem categories: bias (offset), different gradient, and different quantitative rank order. Bias represented in the result of the "overall" quality experiment is generally caused by different listening gear or environmental noises. A different gradient, defined as the relative quality distance between two identical stimuli or conditions during two experiments, is usually caused by a test design that does not cover the entire quality range. A different quantitative rank order is caused by MOS scores' statistical uncertainty expressed in the confidence interval, which needs to be considered when quality ranking is required. Ranking relies only on statistically significant differences, and resolutions finer than 0.3 MOS are not expected since a MOS confidence interval is usually in the range of 0.15 MOS. A generally adopted strategy to minimize scaling effects, such as biases and differing gradients, is to introduce defined anchor and reference conditions in two experiments; this can then be used to align the scores of the two experiments. In addition, other alternatives, such as MOS score normalization across experiments and design constraints to make the distribution of distortion types and quality ranges comparable between different experiments, are under discussion.

The scale calibration of QoE models is needed due to the fact that objective models predict quality based on technical information, and often partial results of individual analysis are combined in a late aggregation step into a single value that is generally dimensionless and not tied to the numerical 1-5 MOS quality scale. The scaling involves multidimensional optimization against the statistical evaluation metrics across a large pool of media samples (e.g., voice, video, audio) carefully selected to uniformly cover all test conditions for

which the algorithm has been designed. The scaling procedure is based on a large number of well-balanced subjective reference experiments, and it is calculated such that the prediction widely follows the scale interpretation of the reference experiments, e.g., by choosing a scaling function that results in a minimum root mean square error (rmse) between the subjective reference experiments and the scaled objective predictions. Therefore, the selection of reference experiments is essential to how the model uses or interprets the quality scale.

The compensation for variance between subjective experiments in the QoE model evaluation process is required due to the inevitable differences between the objective QoE model, which predicts an average MOS value across many experiments as described above, and the subjective MOS value obtained in an individual experiment. As a strategy to minimize this dependency on subjective experiments, an individual compensation is used. The basic assumption is that well-balanced and well-designed subjective experiments are reproducing the qualitative rank-order with high accuracy, while the actual scale range and the gradient, as explained above, may be subject to individual interpretation. Both can be compensated for by individual mappings, where bias and gradient become aligned towards a generalized scale as used by the objective model. Usually, a monotonous linear, or a more sophisticated monotonous part of a third order polynomial, or a logistic mapping function can be applied. The purpose of the mapping function is to minimize the rmse or another metric as well as compensate for offsets, different biases, and other shifts between scores without changing the rank-order. The function is usually applied to the predicted scores before any statistical evaluation metric is calculated.

Per Experiment Statistical Evaluation

The recommended statistical metrics for objective quality assessment need to cover three main aspects: accuracy, consistency, and linearity against subjective data.

It is recommended that the prediction error be used for accuracy; the outlier ratio (OR), or the residual error distribution, for consistency; and the Pearson correlation coefficient for linearity. In addition, confidence intervals, as well as the statistical significance tests, are required for the comparison of these metrics calculated for different QoE models. The ITU-T P.1401 recommendation provides details on how these metrics should be calculated and compared.

Statistical Evaluation in the Context of Subjective Uncertainty: Epsilon-insensitive rmse

For stricter performance evaluation, ITU-T P.1401 introduces the ***epsilon-insensitive rmse*** (***rmse*)*** statistical metric, which considers differences related to an epsilon-wide band around the target value, with ***epsilo***' defined as the 95% confidence interval of the subjective MOS value, which reflects the uncertainty of the MOS scores. The ***modified*** rmse (rmse*) uses as ***modified*** prediction error (Figure1)

$$Perror(i) = \max(0, |MOSLQS(i) - MOSLQO(i)| - ci_{95}(i)),$$

where ci_{95} is the 95% confidence interval of the individual MOS scores. The rmse* is calculated per database, and it describes how the prediction error exceeds the ci_{95}. As a modified rmse, the statistical significance of the difference between two rmse* values is calculated as in the traditional rmse case.

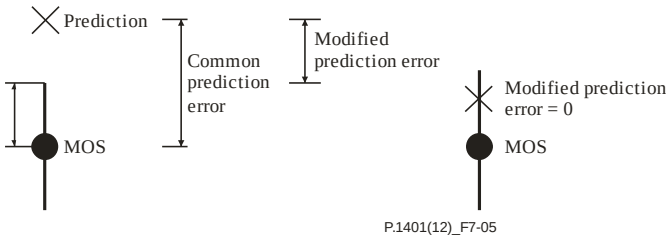


Figure 1. Rmse* calculation.

Statistical Evaluation of the Overall Performance

The overall performance of a model is defined by its performance across each experiment (i.e., test database) as well as across all experiments. Therefore, results per experiment should be aggregated in an overall figure of merit. In order to do so, three steps need to be performed:

- Weighting of databases based on their importance within the QoE model evaluation scope
- Calulation of the aggregated statistical significant distance measure (SSDM) per experiment
- Calculation of the overall performance and statistical significance testing between QoE models.

The SSDM represents the figure of merit of a model per experiment and can be calculated as follows:

$$d_{k,v} = \sum_{i=1}^{Nmetric} W(i) * \max(0, StatMetricF(0.05, N_k, N_k) Result)$$

where $StatMetricF(0.05,N_k,N_k)Result$ denotes the result of the statistical significance test for each evaluated metric i=1...Nmetric (e.g., correlation coefficient, OR, rmse). The index k denotes the experiment, while index v denotes the objective model. F(0.05, n1, n2) is the tabulated value of the F-distribution for n1 and n2 degrees of freedom and 95% significance level. N_k describes the number of considered samples (files or conditions) in experiment k. The function $W(i)$ represents the weight that is allocated to each statistical metric based on their importance to the evaluation process. The highest importance should be allocated to the primary metric which the QoE models have been optimized against.

The overall performance for an algorithm v is defined as

$$p_v = \sum_{k=1}^{M} w_k \times d_{k,v}$$

where M is the total number of databases across the sets, k is the index of the database, $d_{k,v}$ is the distance measure of the model v for the database k, and w_k represents the weight of the database k.

The statistical significance test is applied to the aggregated distance measures p_v calculated for all models. The value p_v is the aggregated distance for v model, p_{min} is lowest p_v in the evaluation and the value K describes the degree of freedom of the F distribution:

$$t_v = \max(0, \frac{p_v}{(p_{\min} + c)} - \mathrm{F}(0.05, K\ , K\ )).$$

If $t_v = 0$, the model v is considered as statistically equivalent to the model with $p = p_{min}$. If $t_v > 0$, the model v is considered as significantly statistically worse than the lowest $p = p_{min}$. The constant c is recommended to be set to 0.0004 based on proved calculations performed for the speech QoE models.

Guidance on Models' Selection

Selecting a best-performing QoE model depends on a variety of factors, such as scope of the evaluation, media and model type, approach used for the QoE model development, etc.

To select the best performing model, it is recommended to consider ***per experiment*** and ***overall performance***, as well as the analysis of ***the worst performance cases***. The models with statistically equal lowest SSDM values per experiment perform the best for that particular experiment. The overall best performing models should exhibit the lowest statistically-equal overall figure of merit calculated as the aggregated SSDM across all experiments. The analysis of the worst performance experiments ensures that the best performer does not show as the worst case in any of the evaluation instances (e.g., per one experiment).

In addition, the evaluation process should use both test databases (e.g. databases used to train the models) as well as validation databases (e.g. databases that are completely unknown to the model). After the selection process is accomplished and a winner is selected, then a characterization phase should take place, with the scope of identifying strengths and weaknesses of the best performing model.

Special Cases

In the case of models designed to estimate the subscriber's perception of various dimensions of quality degradation (e.g., blurriness and blockiness in video, or loudness and coloration in voice), the evaluation is required for each degradation type, as well as on the overall performance.

Special evaluation cases refer either to models with multi-dimensional outputs or to scenarios when only one model is evaluated. In both cases the same framework and same statistical metrics are used.

The second special case refers to the evaluation of one single model. In this scenario, the comparison is performed against pre-defined minimum performance thresholds defined based on previous experiences, whenever available. These scenarios include the case of either a new or improved standard, or a parametric (including planning) or hybrid model when a perceptual model is already in place. In this case, the role of the "best performing model" is played by the minimum performance thresholds defined *a priori* to the evaluation process.

Dr. Irina Cotanis is principal technologist with Ascom Network Testing CTO Office. She holds a Doctorate in Electrical Engineering, and a score card of more than 25 years of experience in wireless and radio communications systems, statistical signal processing and analysis, and statistics, as well as more than 10 years as an active member in standardization organizations, and several publications in IEEE conference proceedings, standards, and text books. She has also acted as reviewer to IEEE papers as well as session chair for various IEEE conferences.

Strategies for testing image and video quality estimators

Amy R. Reibman

Introduction

This note describes a three-stage testing strategy for evaluating the accuracy and effectiveness of image and video quality estimators (QEs).

Assessing the performance of a quality estimator (QE) is essential at three times: during the design process, when selecting the most appropriate QE for a specific application, and when trying to understand the limitations of a selected QE. While the current method of using specification-based subjective testing [1] has been useful, it also is insufficient to fully test and understand the overall performance of a QE. This note describes a three-stage testing strategy for evaluating the accuracy and effectiveness of image and video QEs that supplements the existing methodology.

While our examples and results primarily focus on the full-reference (FR) case, all methods described in this note are applicable for any type of QE: full-reference (FR), reduced-reference (RR), or no-reference (NR), including parametric bitstream QEs. Further, although examples are discussed in terms of images, the strategies also apply to videos.

Principles of software testing

The three-stage testing strategy is motivated by the principles of software testing. In software testing, the following principles are well known [2]:

- The goal of software testing is to find errors, not to demonstrate that the system satisfies its specifications.

- To find errors, it is important to include both positive and negative tests.
- Since it is impossible to use exhaustive testing to find all errors, it is useful to consider a cost-benefit approach.
- The process should be fully automatic.

When applying principles of software testing to evaluate QEs, it is first necessary to understand what constitutes a "bug". Specifically, a bug is a misclassification error [3], [4] defined for a *pair* of images. These errors include false ranking or false ordering (FO) (the objective QE rates an image pair opposite to the humans), false differentiation (FD) (the objective QE rates an image pair as different but the humans do not), and false tie (FT) (the humans rate an image pair as having different quality but the objective QE does not).

Methods for designing image test pairs include white-box and black-box testing, where the underlying mechanism of the QE being tested is either leveraged or not, respectively. Tests can also be classified as *domain-specific* or *adversarial*. Domain-specific methods target specific models within a QE, while adversarial methods use one (or multiple) more accurate QE to systematically identify weaknesses in another QE (or QEs). In adversarial methods, a more accurate QE functions as a "proxy" for the actual, unknown subjective quality. The most flexible, effective methods use a combination of techniques. Further details can be found in [5].

Three-stage testing strategy

We believe a comprehensive testing strategy for quality estimators has three stages, which are presented below in order of increasing cost. The strategies in all stages should be applied, but cheaper strategies should be used first to learn as much as possible about potential weaknesses in a QE during the design process. More costly strategies can be used later, to evaluate a QE for a specific application. Finally, once it is

decided to deploy a specific QE in an actual system, all these strategies can be applied to quantify the limitations of the QE. A QE with known vulnerabilities may still be the best solution given cost or system constraints.

The first stage consists of black-box computational tests, as described in [6]. These require no subjective testing and are very low cost, but can provide valuable information for a QE designer, particularly if they are applied to many reference images. In [6] they were demonstrated with over 400 reference images.

The second stage is to create small-scale targeted subjective tests as described in [5]. These small-scale targeted tests are pairwise tests that probe for specific weaknesses in one or more specific QEs being tested. Image pairs are systematically generated with the specific intent of causing misclassification errors. The pairs can also be chosen based on the results of the first stage, or using the joint collaborative strategy in [7].

The third stage is the well-known specification-based subjective database testing that was proposed by VQEG. This strategy is best represented in the literature, and a number of databases are now available [8]. This strategy is useful to compare the quality of different QEs. However, it does not provide valuable information on how to improve a specific QE.

Figure 1. These image pairs create False Orderings for several QEs. The noisy image on the left was preferred by 17 people out of 30, relative to the blurry image in the middle, while 22 people out of 30 preferred the blurry image in the middle to the noisier image on the right [7]. On the other hand, PSNR, IW-SSIM and SSIM prefer the image in the middle to the less noisy image on the left, while PSNR-A, PSNR-HVS-M and VSNR prefer the noisier image on the right to the middle image. VIF correctly ordered both pairs.

The third stage includes the unbiased process presented in [8] whereby an independent lab provides a validation process using a secret collection of subjectively-annotated videos. QE designers can obtain the performance of their QE for a fee; the secrecy of the test set is maintained. Performance results are reported using a common template, enabling easy comparison across multiple QEs. This additional step provides a rigorous performance evaluation. However, due to its cost it may be more relevant for a company deploying or marketing a QE than for a research contribution.

Closing thoughts

In software testing, one can never know for certain that a program contains no bugs, yet the software is put into use anyway. Similarly, even after vulnerabilities in a QE have been identified, one may choose to deploy it anyway if it has also been shown to assist in other scenarios. The three-stage testing strategy proposed here will enable an informed choice.

Bibliography

[1] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," 2003.

[2] G. J. Myers, T. Badgett, C. Sandler, and T. M. Thomas, *The art of software testing*.: John Wiley and Sons, 1979.

[3] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video-quality metrics: new methods from ATIS/T1A1," *Signal Processing: Image Communication*, vol. 19, pp. 101-107, February 2004.

[4] ITU-T, "J.149: Method for specifying accuracy and cross-calibration of video quality metrics (VQM)," 2004.

[5] F. M. Ciaramello and A. R. Reibman, "Supplemental subjective testing to evaluate the performance of image and video quality estimators," in *Human Vision and Electronic Imaging*, 2011.

[6] F. M. Ciaramello and A. R. Reibman, "Systematic stress testing of image quality estimators," in *IEEE International Conference on Image Processing*, 2011.

[7] A. R. Reibman, "A strategy to jointly test image quality estimators subjectively," in *IEEE International Conference on Image Processing*, 2012.

[8] R. S. Streijl, S. Winkler, and D. S. Hands, "Perceptual quality measurement: Towards a more efficient process for validating objective models," *IEEE Signal Processing Magazine*, pp. 136-140, July 2010.

Amy R. Reibman has been a Professor in the School of Electrical and Computer Engineering at Purdue University since January 2015. She received the B.S., M.S. and Ph.D. degrees in electrical engineering from Duke University. She was an assistant professor in the Department of Electrical Engineering at Princeton University, and spent 23 years at AT&T Labs -- Research, where she was a Distinguished Member of Technical Staff and a Lead Inventive Scientist.

Dr. Reibman was elected IEEE Fellow in 2005, for her contributions to video transport over networks.

In 1998, she won the IEEE Communications Society Leonard G. Abraham Prize Paper Award. She was the Technical co-chair of the IEEE International Conference on Image Processing in 2002; the Technical Co-chair for the First IEEE Workshop on Multimedia Signal Processing in 1997; the Technical Chair for the Sixth International Workshop on Packet Video in 1994. She was a Distinguished Lecturer for the IEEE Signal Processing Society from 2008-2009, was a member of the IEEE Awards Committee from 2010-2012, served on the IEEE Fellow Selection committee from 2012-2014, and served on the IEEE Fellow Selection Strategic Planning committee in 2013 and 2014.

Dr. Reibman's research interests include video compression systems for transport over packet and wireless networks, video quality estimation, video analytics, and 3-D and multiview video.

Dreamed about training, verifying and validating your QoE model on a million videos?

Glenn Van Wallendael, Nicolas Staelens, Enrico Masala, Lucjan Janowski, Kongfeng Berger, Marcus Barkowsky

Training, verification, and validation of objective prediction models require well-chosen test stimuli. The measured prediction performance depends largely on the congruence of stimulus selection in the three steps training, verification, and validation. Different stimulus selection criteria are discussed: extracting a representative set of stimuli from the scope of application, spreading the range of application scope with equidistant stimuli, or using stressful stimuli for the prediction algorithm. Nowadays, most databases are too small to sufficiently cover even one of these evaluation types; a large-scale database may solve the problem but requires new statistical methods and understanding of quality evaluation.

Although we are not yet at a million videos, gradual additions over time will eventually get us there. In the beginning of the large-scale database effort, in 2012, the main focus was on encoding conditions.

Therefore, it all started with 10 HD-sequences, downscaled by a factor of 4 and 8. They were encoded with 430 different encoding parameters like bitrate, frame rate, encoding structure, encoder implementation, number of slices, and so on, resulting in 12,960 H.264/AVC encoded video streams. These sequences were annotated by Full-Reference (FR) results. The same video sequences were encoded with the H.265/HEVC standard as well, with 5952 different encoding settings leading to another set of 59,520 encoded sequences.

What's the quality of each of these sequences? While a full subjective experiment is prohibitive, objective algorithms may be computed and compared, stimulating research on new types of agreement analysis. Currently, the database features five video quality metrics computed for each encoded video

sequence: Peak Signal to Noise Ratio (PSNR)[2], Structural Similarity Index (SSIM)[3], Visual Information Fidelity (VIF)[4], Video Quality Metric (VQM)[5], and Perceptual Video Quality Measure (PVQM)[6]. Further details are available on the JEG wiki.[7]

Efforts are under way to extend the database in the direction of adding more content, notably Ultra-HD resolution sequences, as well as to provide the same measures for sequences impaired by packetlosses. To this end, an H.265/HEVC robust decoder[8] has been used to produce distorted video sequences on the basis of 25 different loss patterns. Although it is difficult to provide such measures for all loss patterns applied to all the encoded sequences due to the huge processing time required, it is expected that in the next six months at least a significant subset of the original encoded video sequences will have all the quality measures corresponding to the 25 loss patterns.

[2] NTIA / ITS. (2001). A3: Objective Video Quality Measurement Using a Peak-Signal-to-Noise-Ratio (PSNR) Full Reference Technique. ATIS T1.TR.PP.74-2001

[3] NTIA / ITS. (2001). A3: Objective Video Quality Measurement Using a Peak-Signal-to-Noise-Ratio (PSNR) Full Reference Technique. ATIS T1.TR.PP.74-2001

[4] Sheikh, H. R., &Bovik, A. C. (2006).Image information and visual quality. IEEE Transactions on Image Processing, 15(2), 430–444.

[5] ITU-T Study Group 9.(2004). ITU-T J.144 Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. ITU-T J.144

[6] Hekstra, A. P., Beerends, J. G., Ledermann, D., de Caluwe, F. E., Kohler, S., Koenen, R. H., et al. (2002). PVQM – A perceptual video quality measure. Elsevier, Signal Processing: Image Communications 17, , 781–798.

[7] http://vqegjeg.intec.ugent.be/wiki/index.php/JEG_no-reference_hybrid_HEVC

[8] http://media.polito.it/jeg

Development and performance evaluations of objective assessment algorithms

Most industrial and research effort has been spent so far on creating holistic objective assessment algorithms optimized for a particular application scenario. Rarely, the intermediate steps of such complex algorithms have been evaluated separately.

Figure 1 shows a functional overview of the typical development cycle. The cycle, in general, includes a training procedure followed by verification, and after development has finished, validation is performed. In the training procedure, various indicators are developed, pooled over space and time, and then merged to predict the perceived quality. Typical prediction performance measures include linearity (Pearson Linear Correlation Coefficient, PLCC), Rank Ordering (Spearman Rank Order Coefficient, SROCC), and accuracy (Root Mean Square Error, RMSE). The stability of the estimated fitting parameter during training and the appropriateness of its count as compared to the samples available for training may be evaluated by cross-validation of the training process.

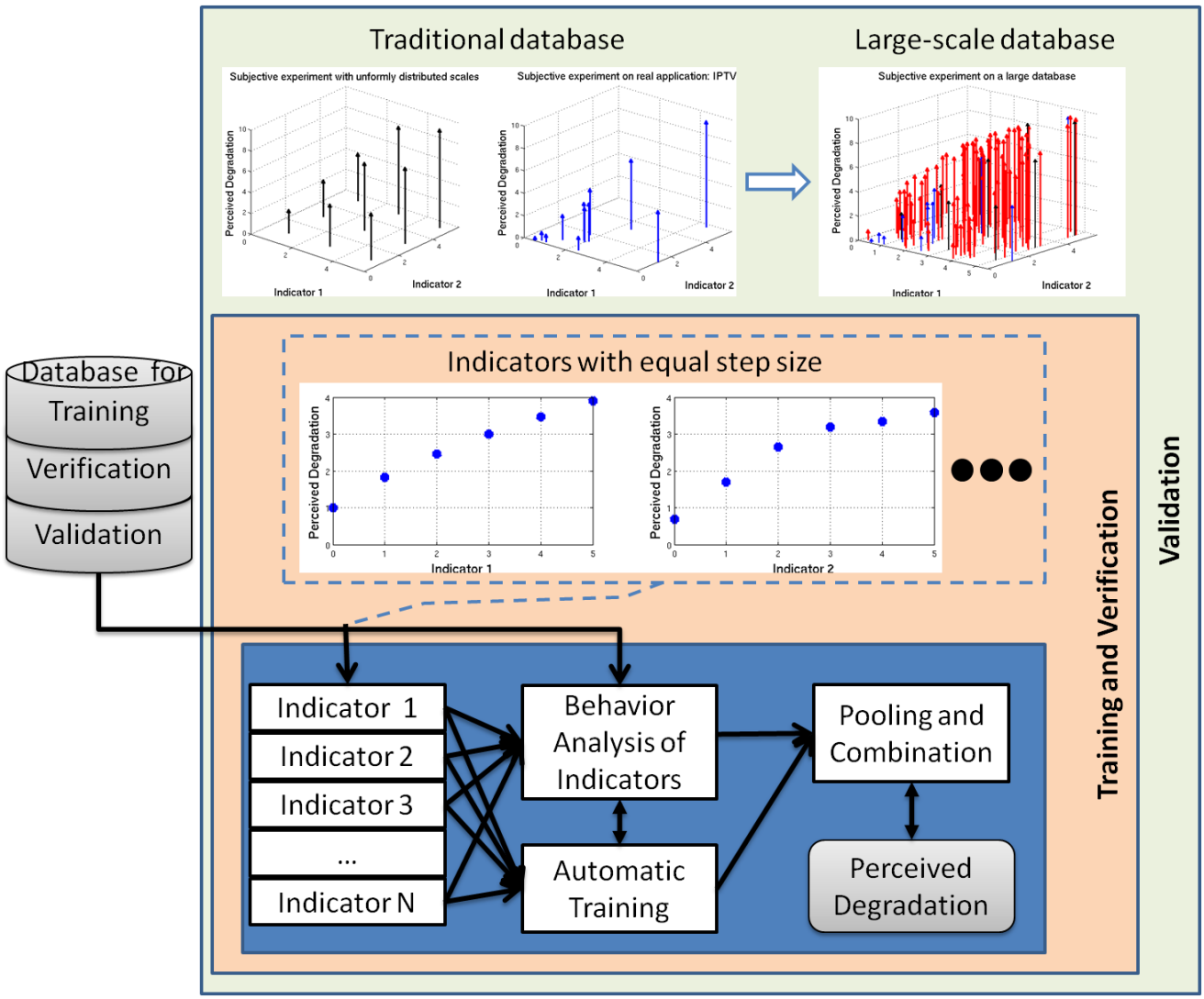


Figure 1: An overview of a typical development cycle of objective quality assessment

Validation requires a different set of samples. In the validation procedure, algorithms of objective quality assessment are often validated using the same performance measures as previously introduced for verification. In addition, more sophisticated measure may be used, for example epsilon-insensitive RMSE (RMSE*), Outlier Ratio with respect to Standard Error as detailed in ITU-T P.1401, and Accuracy Analysis or Resolving Power as specified by ITU-T J.149.

A typical objective video quality assessment algorithm combines several quality indicators where each of them should ideally provide good quality prediction results when

used within its scope of application, rough estimates when used at the boundaries or in an extended scope and each of themshould stay neutral when confronted with degradations out of its specific measurement scope. A typical example would be a perceptual frame rate indicator that correctly predicts constant frame rate settings, that has limited accuracy when the framerate becomes variable, and that stays neutral when longer pauses and skips occur as those isolated events require a different perceptual measurement.[9]

Figure 1 shows the systematic development situation of a quality prediction algorithm in a block diagram. Several perceptual features are identified and experimented in isolated subjective experiments such that the degradations occur equally often in different strengths. The expected behavior of each indicator with respect to subjective results is illustrated by the two plots in the orange verification procedure block. This process may be simplified as a one dimensional training procedure for each indicator algorithm but in practice the indicators are interdependent. For example, the ratio of frame rate reduction is dependent on resolution in the application scenario of IPTV.

How is a large database going to help in the development stage?

Most objective metrics were designed for certain applications, such as compression only,[10] or compression and transmission degradations, additionally including display postprocessing and so on. The existing databases were also built for certain applications. Metrics developed for compression may perform well on the database of compressed videos, and it is very

[9] Barkowsky, Staelens, Janowski, Koudota, Leszczuk, Urvoy, et al. (2012). Subjective experiment dataset for joint development of hybrid video quality measurement algorithms. QoEMCS 2012, Berlin, Allemagne.

[10] K. Zhu, C. Li, V. K. Asari, and D. Saupe, "No-reference video quality assessment based on artifacts measurement and statistical analysis." IEEE Transactions on Circuits and Systems for Video Technology, 2014.

likely that these metrics were tested only on compressed videos. It is of great interest to know how these distortion-specific metrics perform on videos in their extended scope or out of their scope—for example, how a metric designed for H.264 compressed natural videos performs on HEVC compressed videos, videos with packet loss, and computer-generated videos. Observing the performance of distortion-specific metrics on videos in their extended scope and out of their scope calls for a large-scale database with videos impacted by various degradations.

Another problem that may be solved by a large database is machine-learning based algorithms' over-fitting. Machine-learning based algorithms, in general, have good quality prediction accuracy. They are, however, highly prone to over-fitting on the training set, and therefore end up with a low generalization ability.[11] In many cases, the number of videos in the training set is small in comparison to the large number of parameters in the trained algorithm. Additionally, the content of videos in the training set is diverse enough. Consequently, the predicted quality of the model may show large errors with respect to the MOS when a video has different content from the training videos. Both problems, over-fitting and lack of considered content, can be avoided by a large databases. Typically, machine-learning methods' stability is evaluated by cross-validation. For example, the 10-fold cross-validation is an often used strategy to assess how a machine-learning based algorithm performs on unseen data. We noticed that the statistical results of cross-validation are sensitive to cross-validation strategy and the number of video sets in one fold. With a large video database, the number of video sets in one fold is also large, so that the cross-validation results are robust, and, therefore, the estimated general performance of a machine-learning based algorithm on unseen data is robust.

[11]P. Gastaldo and J. A. Redi, "Machine learning solutions for objective visual quality assessment," in the sixth International Workshop on Video Processing and Quality Metrics, Jan. 2012.

How is a large database going to improve the validation stage?

Performance evaluation with respect to the application scenario is the primary purpose of the validation step. Previous VQEG efforts on SDTV, Multimedia, HDTV, and Hybrid models document the enormous effort required for this black box type of independent validation of computational models. [12]

The selection of both the source content (SRC) and the degradation, also called a Hypothetical Reference Circuit (HRC) forms a crucial part of such evaluation. Open questions include whether the coverage of samples shall be uniform with respect to the scope of application (i.e., as many perfect as average as strongly degraded videos) or uniform with respect to the expected application scenario (i.e., more average quality videos than perfect or strongly degraded videos). Figure 1 shows this graphically in the green validation area. The first two diagrams illustrate the situation in the case that the validation database is designed for equally covering the scope of the indicators, which may or may not coincide with equally covering the application scope.

The second diagram illustrates the distribution when focusing on typical examples: usually the perceived quality is above average most of the time and strong degradations occur rather seldom. The third diagram illustrates that a large-scale database allows for both types of evaluations and actually may invert the interpretation: it may provide the answer as to which application scopes an algorithm can be applied to, besides the one that it was designed for.

This question also applies to content. The choice of extreme contents, such as artistic video sequences, may bias the evaluation while allowing for the analysis of the stability of

[12] See http://www.its.bldrdoc.gov/vqeg/reports.aspx

Glenn Van Wallendael *obtained the M.Sc. degree in Applied Engineering from the University College of Antwerp, Belgium, in 2006 and the M.Sc. degree in Engineering from Ghent University, Belgium in 2008. Afterwards, he worked towards a Ph.D. at Multimedia Lab, Ghent University, with the financial support of the Agency for Innovation by Science and Technology (IWT). Currently, he continues working in the same group as a post-doctoral researcher. His main topics of interest are video compression including scalable video compression and transcoding.*

Nicolas Staelens *obtained his Master's degree in Computer Science at Ghent University (Belgium, 2004).In 2006, he joined the Internet Based Communication Networks and Services (IBCN) group at Ghent University where he received a Ph.D. degree in Computer Science Engineering in February 2013. The topic of his dissertation was "Objective and Subjective Quality Assessment of Video Distributed over IP-based Networks".As of 2007, he is also actively participating within the Video Quality Experts Group (VQEG) and is currently co-chair of the Tools and Subjective Labs support group and the JEG-Hybrid project.*

the algorithms. A large-scale database would therefore allow for more detailed analysis including overall suitability of quality prediction algorithms and their behavior at the limits of the application scope.

More detailed analysis may also be obtained with respect to the accuracy of an indicator measuring a technical parameter (e.g., bitrate), a perceptual feature (e.g., blockiness), or a complete algorithm within a certain quality range, i.e. near-lossless or strongly degraded. The combination of several algorithms may be proposed during validation.[13]

The availability of a variety of SRC and HRC used for validation is often a bottleneck in traditional approaches.

A large-scale approach may have such a large selection of both SRC and HRC that conducting a formal subjective assessment on a subset may be considered sufficient for validation. Otherwise, the reproducible processing for the creation of the database may simplify the creation of similar or completely new processed sequences. Evaluating algorithms on each result obtained in the large-scale database allows for drawing a complete picture of its stability, applicability to a certain (sub-)scope, and comparing with other available algorithms. An example would be to provide a resolving power analysis for each application that may be automatically predicted in a next step.

Sample results

To give a rough idea of the possibilities opened by such the currently available large-scale database, a sample validation result is reported here. To give a rough idea of the possibilities

[13]Barri, A.; Dooms, A.; Jansen, B.; Schelkens, P., "A Locally Adaptive System for the Fusion of Objective Quality Measures," Image Processing, IEEE Transactions on , vol.23, no.6, pp.2446,2458, June 2014

Enrico Masala *received the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 2004. In 2003, he was a visiting researcher at the Signal Compression Laboratory, University of California, Santa Barbara, where he worked on joint source channel coding algorithms for video transmission. Since 2011 he is Assistant Professor in the Control and Computer Engineering Department at the Politecnico di Torino. His main research interests include simulation and performance optimization of multimedia communications (especially video) over wireline and wireless packet networks.*

Lucjan Janowski *is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology, in Krakow, Poland. He is a Co-Chair of the VQEG JEG-Hybrid project (http://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx).*

opened by the currently available large-scale database[14], a sample validation result is reported here. When taking any two video sequences from the large scale data set and evaluating their quality with either PSNR, SSIM, or VIF, a rank order can be established. It would be interesting to understand to what extent the three measures agree on the ranking. For three measures, there will be either agreement or exactly one metric which does not agree.

For each measure we calculate the distance between the two sequences in a pair when the measure disagrees. There is a total of six possible cases, i.e., for each one of the three measures, one of the other two does not agree.

The scatterplot in Figure 2 represents all pairs of encoded video sequences for src06 when VIF disagrees with PSNR and SSIM. The grey level represents the number of sequences that do not agree, for a certain difference of the PSNR and SSIM on the x and y axes. Darker shades indicate more disagreement between measures. It can be seen that beyond a certain difference in each measure the quality difference is so pronounced that all metrics agree. This limit is approximately +-2 dB for PSNR and +-0.05 for SSIM on their natural scales.

[14] Leszczuk, M., Janowski, L., & Barkowsky, M. (2013). "Freely Available Large-scale Video Quality Assessment Database in Full-HD Resolution with H.264 Coding." IEEE Globecom 2013

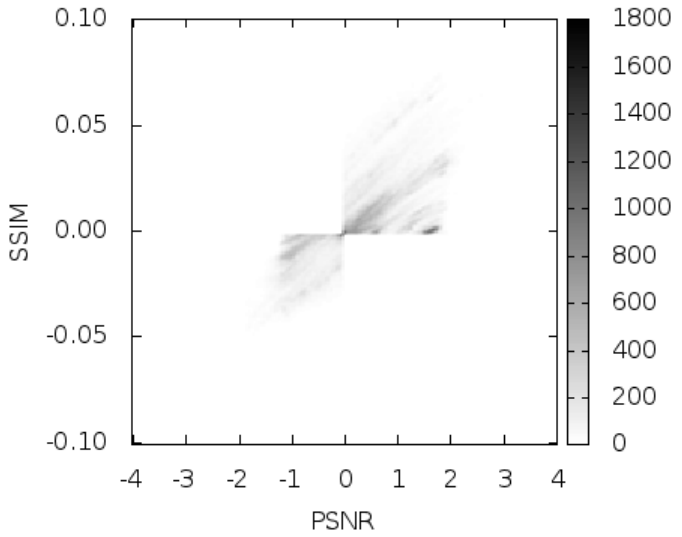


Figure 2: Density plot of the difference of SSIM and PSNR in the pairwise comparison when VIF disagrees

Selecting the 95 percentile value, a reasonable threshold for the prediction consistency of the measure with respect to the two others may be determined. As can be seen from Figure 3, this value is strongly sequence dependent (compare, for instance, seq01 and seq03 for PSNR), and within the same sequence, there can be a large difference depending on the cause of disagreement (see, e.g., seq08).

This shows the advantage of having a large set of coding conditions for measuring the influence of content on a quality measure in validation. Please note that this analysis is purely based on disagreement, subjective experiments are required to determine whether the disagreement of one measure with respect to the two others indicate a failure of that measure and whether an agreement of the three measures is consistant with human observation.

Kongfeng Berger *was born in Shandong, China. She received the M.Sc. degree in communication and information systems from Shandong University, China, in 2006, and the Ph.D. degree in multimedia signal processing from the University of Konstanz, Germany, in 2014. She is currently aa Postdoctoral Fellow at the University of Nantes, France. Her main research interests include visual quality assessment, image processing, motion analysis, natural scene statistics, feature selection, and machine learning.*

Marcus Barkowsky *received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the Image and Video Communications Group at IRCCyN at the University of Nantes in 2008, and was promoted to associate professor in 2010. His activities range from modeling effects of the human visual system, in particular the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays using psychometric and medical measurements. He currently co-chairs the VQEG "3DTV"and "Joint Effort Group Hybrid" activities.*

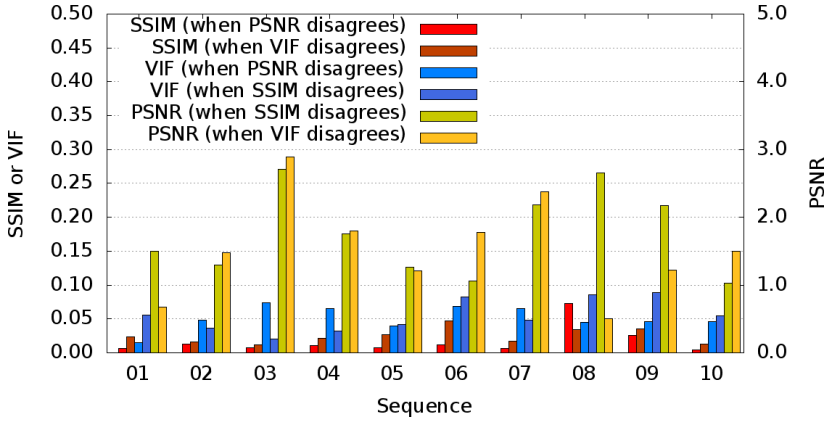


Figure 3: 95 percentile of the two agreeing video quality measures when one disagrees

What's next?

Establishing large-scale databases is a continuous effort; packet losses and higher resolutions as well as more content and encoders need to be added for improving the training, verification and validation process. Further statistical analysis tools should be researched in parallel. Innovative analysis questions may emerge, as shown with the example above.

Validation of reliable 3DTV subjective assessment methodology - Establishing a Ground Truth Database

Jing Li, Marcus Barkowsky, Patrick Le Callet

Subjective assessment methodology for 3DTV

Quality of Experience (QoE) in 3DTV is a multi-dimensional concept which includes image quality, depth quality, and visual comfort. How to measure this multi-dimensional concept is a challenging issue nowadays. In this letter, we introduce a Ground Truth database which is targeted for the standardization of subjective methodologies for QoE in 3DTV.

May the Absolute Category Rating (ACR) be used with 3D stereoscopic content? Experts agree that as long as the degradations are on one single perceptual scale, notably image degradations such as coding artifacts, the previously employed assessment methods such as ACR or DSCQS may be suitable. In 2012, the International Telecommunication Union (ITU) published ITU-R BT.2021[1] for the Subjective assessment methods of stereoscopic 3DTV systems. These recommended methods are derived from ITU-R BT.500 and measure the three primary dimensions of QoE independently: picture quality, depth quality, and visual comfort. However, depending on the transmission conditions, 3DTV may impact all three scales simultaneously. For example, a packet loss may lead to mismatched content in one of the two views leading to an immediate sensation of visual discomfort. In such a case, the previous methods may not be applicable anymore. This is also reflected by ITU-R BT.2021, where the recommended test

[1]International Telecommunication Union - Radiocommunication Sector, "Recommendation ITU-R BT.2021: Subjective methods for the assessment of stereoscopic 3DTV systems", 2012

methods are not suggested for the assessment of naturalness, sense of presence, or the overall QoE. Concerning this issue, in 2013, the IEEE P3333.1[2] Work Group was established to develop novel methods for standardization of subjective quality assessment methodology in 3DTV.

Currently, VQEG experts agreed that the most suitable method for subjective experiments that span several scales is the Paired Comparison (PC) method. Observers just need to choose one sequence in each pair which avoids scale and language interpretation issues; this criterion is also easy to understand. The drawback of PC is the number of visualizations and therefore the length of the subjective experiment, particularly when each pair is visualized in the Full Paired Comparison (FPC) method. To resolve this issue, a new design, Optimized Rectangular Design (ORD),[3] has been proposed to reduce the number of comparisons in PC and is now widely used in the community. In 2014, the ORD was accepted by IEEE P3333.1 Work Group as a standard quality assessment methodology for 3D contents. The basic idea of the ORD method is to arrange the stimuli indices optimally into a rectangular matrix and only compare the pairs within the same row or column. In this way, the number of comparison is significantly reduced compared to FPC.

As it was shown that precision similar to that of FPC can be reached by the ORD method[4], VQEG has therefore decided to run a coordinated subjective experiment on QoE of 3DTV by using the PC ORD method. The obtained results are considered "Ground Truth" for the standardization of subjective assessment methodology for QoE of 3DTV. Thus, the reliability and suitability of ACR, DSCQS or other newly

[2] IEEE P3333.1 WG - Quality Assessment of Three Dimensional Contents based on Psychophysical Studies Working Group, IEEE Computer Society.
[3] J. Li, M. Barkowsky, P. Le Callet, "Subjective assessment methodology for Preference of Experience in 3DTV", IEEE IVMSP, 2013.
[4] J. Li, M. Barkowsky, P. Le Callet, "Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs", Proceedings of the SPIE Electronic Imaging, Stereoscopic Displays and Applications, 2013.

designed subjective methods can be evaluated and validated based on this database.

VQEG GroTruQoE3D database

The database is called VQEG GroTruQoE3D (**Gro**und **Tru**th **Q**uality **o**f **E**xperience in **3D**) database. This database contains a well-chosen set of 3D contents (SRC) exhibiting small and large depth budgets, slow and fast planar movement, various kinds of in-depth movement, fine spatial details, strong contrasts, and dark scenes. They were degraded with 18 degradations (HRC) that were selected by experts in order to target a uniform usage of the three scales and their interaction. The distribution of the degradation levels of the selected HRCs on each dimension (image quality, depth quality, and visual comfort) is shown in Figure 1. The video contents are shown in Figure 2. This database has been made available[5] and is currently under evaluation by VQEG's 3DTV group.

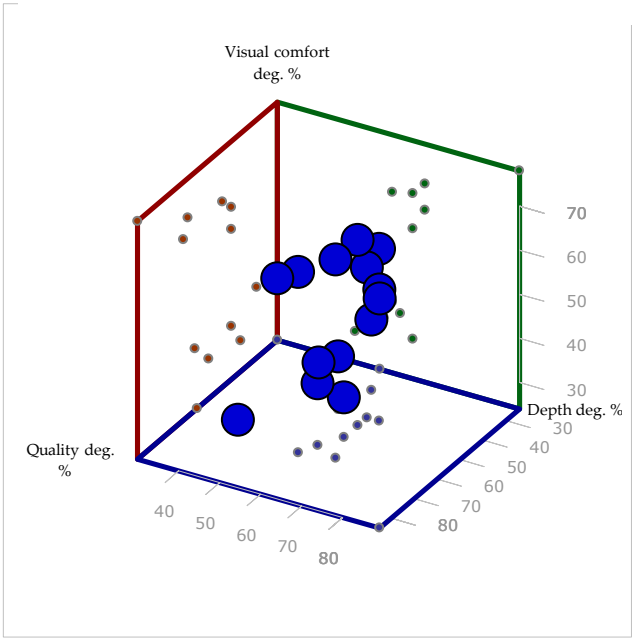


Figure 1. The distribution of degradation levels of the HRCs on each dimension.

[5] ftp://ftp.ivc.polytech.univ-nantes.fr/VQEG_3DTV_GROTRUQOE3D

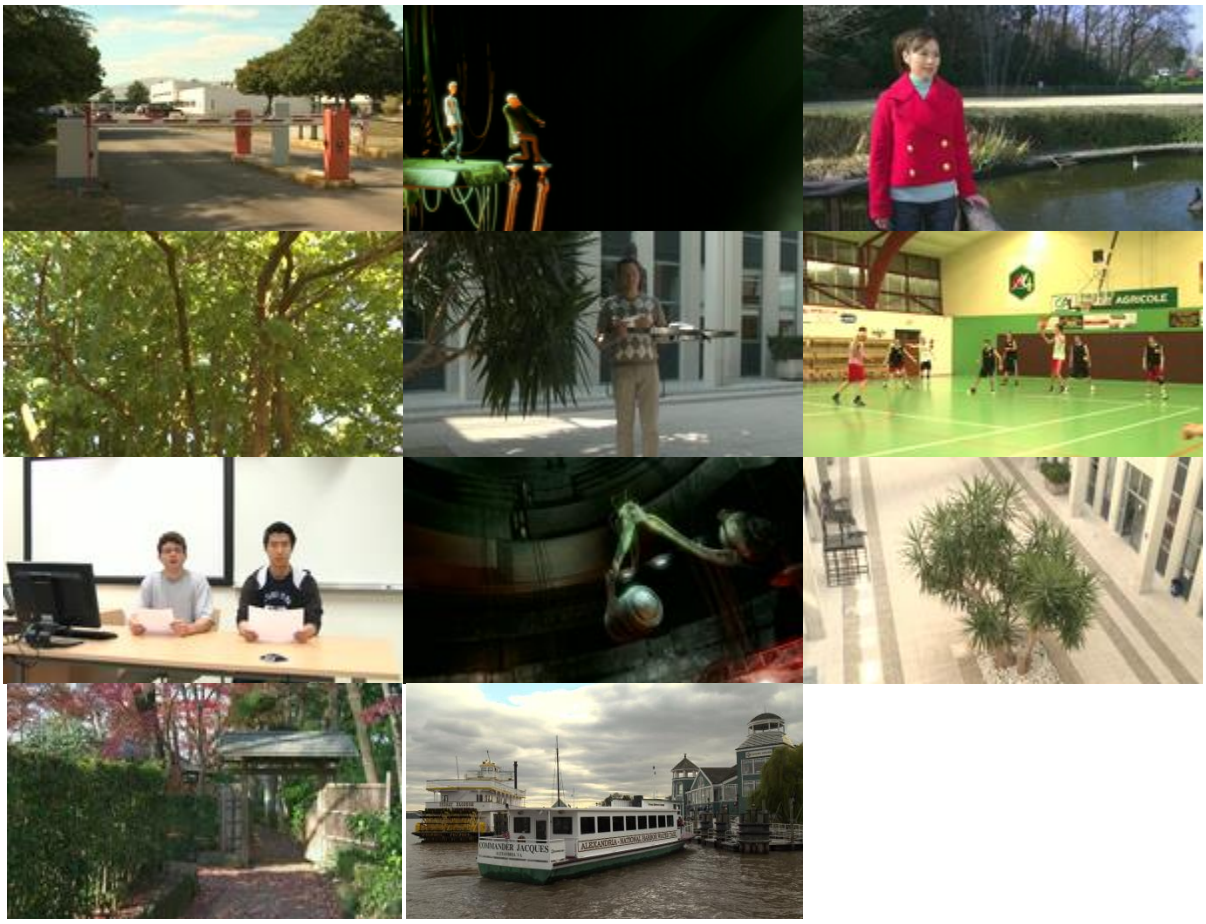


Figure 2. Thumbnails of the VQEG GroTruQoE3D database.

The QoE of the dataset is assessed by the ORD method in such a way that the 18 HRC indices are arranged into a 3 × 6 matrix and only the pairs within the same row or column are compared, which leads to $3 \times \binom{6}{2} + 6 \times \binom{3}{2} = 63$ comparisons per observer. However, considering that there are 11 SRCs altogether, for the ORD method the total number of comparisons would be 11 × 63 = 396 observations per observer, which is still a large number. To make the test feasible, it has been decided to split the workload amongst eight laboratories: IRCCyN (France), INSA (France), Yonsei University (Korea), UPM (Spain), NTIA (USA), T-labs (Germany), FuB (Italy) and BskyB (UK). The construction of a common set of pairs for all

labs is required which will allow for the validation of results among labs.

The common set includes two fixed SRCs and 18 HRC pairs. The two SRCs are selected in such a way that they are sensitive to test environment. The 18 HRC pairs are constructed by a 3 × 3 HRC matrix which is a subset of the whole 3 × 6 HRC matrix. The selected 9 HRCs represent 3 levels in 3 dimensions of the 3D QoE, i.e., image quality, depth quality, and visual comfort.

The obtained data will be collected and then analyzed for two main purposes. The first goal is to evaluate the validity of the acquired data in the different subjective assessment labs, thus allowing the creation of a large common dataset. When two alternative forced choice (2AFC) Paired Comparison is used as the assessment methodology, scale adaptation problems do not arise. The second goal is to establish a scale value for each Processed Video Sequence (PVS). This eases the comparison of the results to assessment methods that use direct scales such as Absolute Category Rating or Double Stimulus Continuous Quality Scale. For the first goal, the main statistical analysis tool used is Barnard's-exact-test, which examines whether the PC preference data obtained from two labs is significantly different. Thus, "outliers" may be detected by determining a threshold on the total number of significantly different pairs. For the second goal, to convert the paired comparison data to scale values, the Bradley-Terry model will be applied. This could provide not only the scales for all PVSs, but also some statistics, including the confidence intervals for each PVS, how well the model fits, etc.

Validation of reliable subjective assessment methodology in 3DTV

The results of the GroTruQoE3D evaluation may be used to verify not only the performance of existing subjective quality

Jing Li *received her M.S. degree in Electronic Engineering from Xidian University, Shaanxi, China, in 2010, and her Ph.D. degree from University of Nantes, France, in 2013. Her research interests include subjective assessment methodologies and objective modeling on QoE, visual discomfort and quality of experience in 3DTV and Ultra HDTV. She is now leading the construction of the VQEG GroTruQoE3D database.*

Marcus Barkowsky *received the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 2009. He joined the Image and Video Communications Group at IRCCyN at the University of Nantes in 2008, and was promoted to associate professor in 2010. His activities range from modeling effects of the human visual system, in particular the influence of coding, transmission, and display artifacts in 2D and 3D to measuring and quantifying visual discomfort and visual fatigue on 3D displays using psychometric and medical measurements. He currently co-chairs the VQEG "3DTV" and "Joint Effort Group Hybrid" activities.*

assessment methods, but also the impact of different perceptual measurement scales, the influence of observer training on the results, etc. New methodologies may be developed based on the results.

The existing quality assessment methods can be validated, for example, using the following criteria:

1) Correlation analysis: By calculating the Pearson Linear Correlation Coefficients (PLCC) and Spearman Rank Order Correlation Coefficients (SROCC), the correlation between the results obtained by Pair Comparison (Ground Truth) and the tested methodology can be obtained, which shows the consistency of the tested methodology with the ground truth.

2) Accuracy analysis: By calculating the Root Mean Square Error (RMSE) between the ground truth and the fitted data, the accuracy of the tested methodology can be evaluated.

3) Distinguishability analysis: The distinguishability of Pair Comparison can be tested by the Barnard's-exact-test, where the significance of the observer's preference on each pair can be shown. For the tested methodology, the distinguishability can be evaluated by confidence intervals or Student's t-test. Another possible way is to convert the results of the tested methodology to PC data and then Barnard's-exact-test may be used. Statistical analysis between the two subjective test methodologies is enabled and the relative performance of the tested methodology can be evaluated.

With this GroTruQoE3D database, a list of verified and validated assessment methods for 3DTV may be established for standardization in ITU Recommendations.

Patrick Le Callet *is a full Professor at Ecole polytechnique de l'Universite de Nantes. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. He is currently co-chairs the "3DTV" activities and the "Joint-Effort Group", driving mostly High Dynamic Range topic in this latest. He is currently serving as associate editor for IEEE transactions on Circuit System and Video Technology, SPIE Journal of Electronic Imaging and SPRINGER EURASIP Journal on Image and video Processing.*

Reliably combining quality indicators

Adriaan Barri, Ann Dooms, Peter Schelkens

Is machine learning (ML) suitable for objective quality assessment?

Objective quality measures based on machine learning (ML) require fewer computations and are less affected by inaccuracies in the HVS models. But they may also yield less transparent quality predictions when the ML responses are difficult to interpret. The absence of interpretability may disguise serious vulnerabilities in the design of the objective quality measure.

In recent years, machine learning (ML) has gained increased attention as a technique to improve the accuracy of objective quality measures. By incorporating ML, objective quality measures can mimic mechanisms of the human visual system (HVS) that otherwise had to be modeled explicitly. As a consequence, ML-based quality measures require fewer computations and are less affected by our limited knowledge of the HVS. On the downside, they yield less transparent quality predictions, because the ML responses are often difficult to interpret. The absence of interpretability may disguise serious vulnerabilities, such as consistency violations, unstable predictions in the high quality range, and severe false orderings. Our recently developed Locally Adaptive Fusion (LAF) method addresses these issues by imposing strict regulations on the ML behavior. This article analyzes the prediction performance of LAF by traditional validation techniques and by complementary stress tests on an unannotated image database. These tests explain the benefits of LAF and illustrate the importance of a thorough validation.

Locally Adaptive Fusion (LAF) when transparency is important

In contrast to traditional ML methods, Locally Adaptive Fusion (LAF) is specifically designed for objective quality assessment. The LAF method predicts quality in two steps. Firstly, the signal is subjected to multiple fusion units. Each fusion unit is a fixed weighted sum of predetermined quality indicators, which are meant for specific content or distortion types. Secondly, the calculated fusion unit values are combined through adaptive weighting, using a second set of weights that change depending on the received signal. This nonlinear response allows LAF to better mimic complex HVS mechanisms.

To ensure interpretability, the weights of LAF are directly related to the quality indicators. The strict regulations imposed by LAF come with three additional advantages: reproducibility, consistency, and computational scalability.

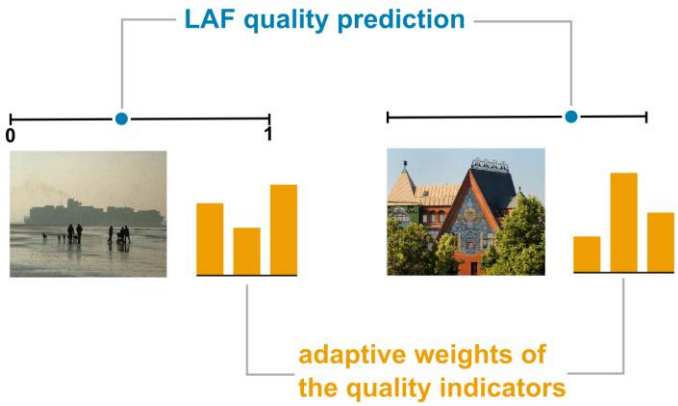


Figure 1. The weights used by LAF are directly related to the quality indicators. These weights change depending on the content and distortion type. In the above illustration, LAF predicts the quality of two still images by adaptively weighting three input quality indicators.

The behavior of LAF is strictly regulated and much easier to interpret in comparison with other nonlinear ML methods (e.g. neural networks). By design, the weights of LAF are directly related to the quality indicators. As a result, the influence of the quality indicators on the quality prediction of the received signal can be visualized (Figure 1).

The imposed regulations of LAF come with three additional advantages: reproducibility, consistency, and computational scalability. Firstly, the training phase of LAF does not require a random initialization. Unlike neural networks, re-training LAF on the same data will always produce the same weights. Secondly, the LAF response is always consistent with the input quality indicators to avoid overfitting. Thirdly, LAF can be easily configured to find the optimal trade-off between computational complexity and prediction accuracy.

Validation methods adjusted to ML-based quality measures

This section compares the reliability of LAF with a one-layer feed forward neural network (FFNN). To avoid the curse of dimensionality, we limited the ML input to three simple quality indicators for still images, one of the no-reference type and the two others of the reduced-reference type. The selected quality indicators respectively measure blocking artifacts, spatial information loss, and contrast similarity. The performance of ML-based quality measures is typically tested on multiple annotated databases. However, these tests revealed no significant differences between LAF and FFNN (Table 1). For a more thorough comparison, we needed complementary stress tests on an unannotated database.

Table 1. The validation tests on LIVE, CSIQ, and TID revealed no significant performance differences. More details are in (Barri A. et al., 2014).

Tests on annotated databases (Pearson correlation)	LAF	FFNN
Repeated cross-validation on the LIVE database	0.96	**0.965**
Database independence *Training set:* LIVE – *Test set:* CSIQ	**0.967**	0.959
Robustness for unknown distortions *Training set:* LIVE – *Test set:* TID	**0.822**	0.790

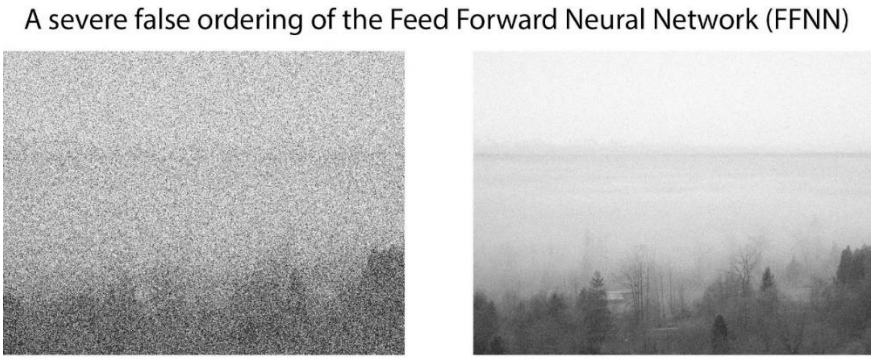


Figure 2. Even when ML-based quality measures obtain high correlation values on annotated quality assessment databases, they may produce severe false orderings on larger, unannotated test databases. In the above illustration, the FFNN prefers the quality of the left image. Such severe false orderings confirm the importance of complementary stress tests during validation.

We evaluated the ML-based quality measures on an unannotated stress test database containing 650 reference and 26,000 distorted images. We acquired three new insights:

- **Traditional ML is often inconsistent.** Given two signals, suppose all input quality indicators systematically give a higher rate to the first signal. Then we proved the LAF method will always agree with the preference of the indicators. Traditional ML tends to ignore the indicators to better fit the training data. For FFNN, we discovered more than 100,000 of these consistency violations.
- **Traditional ML is unstable in the high quality range.** For the quality predictions of barely distorted images, LAF will optimize the weights of the indicators to the high quality range. The FFNN will still employ the no-reference blocking indicator, but this yields unstable quality predictions due to the low visibility of the artifacts.
- **Traditional ML may produce severe false orderings.** The quality predictions should decrease when the distortion rate is gradually increased. On the stress test database, LAF produces fewer false orderings than FFNN (6 vs. 119). Moreover, the false orderings of FFNN were often very severe (Figure 2).

Adriaan Barri is a member of the iMinds research center of Flanders, and the department of electronics and informatics (ETRO) at the Vrije Universiteit Brussel, Belgium. He holds a PhD bursary from the agency for Innovation by Science and Technology (IWT). His research focuses on machine learning and quality assessment.

Ann Dooms is a member of iMinds and holds a professorship at ETRO, Vrije Universiteit Brussel. Dooms leads a research team in Multimedia Forensics, which studies the lifecycle of a multimedia item to answer forensic questions ranging from authenticity and traitor tracing over perceptual quality and compressed sensing to digital painting analysis.

Peter Schelkens is research director at iMinds and holds a professorship at ETRO, Vrije Universiteit Brussel,. In 2010, he joined the board of councillors of the Interuniversity Microelectronics Institute (IMEC), Belgium. In 2013, he obtained an ERC Consolidator Grant focusing on digital holography. He is an elected member of the IEEE Technical Committees IVMSP and MMSP, and is participating in the JPEG and MPEG standardization activities.

What have we learned?

Not all vulnerabilities of ML-based quality measures can be detected by traditional validation methods. Most vulnerabilities can be reduced or even avoided when more interpretable ML methods are used, such as LAF. We firmly believe that LAF is more reliable than other ML solutions for real-life applications. More information can be found in the referenced paper and at www.locally-adaptive-fusion.com.

References

Barri A., Dooms A., Jansen B., Schelkens P. (2014),
"A Locally Adaptive System for the Fusion of Objective Quality Measures,"
IEEE Transactions on Image Processing, Vol. 23, No. 6, pp. 2446-2458.

T1A1 Validation Test Database

Margaret H. Pinson and Arthur Webster

Introduction

In the early 1990s, broadcasters were transitioning from analog to digital systems and ISDN video teleconferencing was an exciting new technology. During 1993-1994, the T1A1 committee conducted an objective video quality metric validation test focused on video teleconferencing applications. T1A1 was a subcommittee of the American National Standards Association (ANSI) accredited Alliance for Telecommunications Industry Solutions (ATIS). T1A1 is now known as PTSC QoSR— Packet Technologies and Systems Committee, Quality of Service and Reliability.

This document summarizes the T1A1 video quality subjective test. We focus on information that a current researcher needs to effectively use this dataset. The T1A1 video sequences and differential mean opinion scores (DMOS) are available on the Consumer Digital Video Library (CDVL, www.cdvl.org, [1]). The test plan and analyses appear in [2]-[4]. A future paper will document this test in more detail.

Scenes and Impairments

The T1A1 validation test analyzed standard definition video as per the NTSC broadcasting standard. The test focused on video teleconferencing applications. The source video sequences (SRC) are all in the public domain. Three are ITU-R Rec. BT.802 standard test sequences. The other 22 SRCs were donated by NTIA/ITS, Delta Information Systems (DIS), PictureTel Corp, and Compression Labs Inc. (CLI).

Twenty-five SRCs were chosen to represent five content categories (see Figure. 1). Most videos were filmed using broadcast quality cameras; however some intentionally included production problems (e.g., light level fluctuation, analog noise, deterioration typical of old film). These 25 SRCs later became the ANSI 801.1 standard test sequences. The digitized videos on CDVL contain occasional analog impairments stemming from the age of the tapes when it became practical to convert the analog Betacam-SP tapes into a digital format (uncompressed AVI, 4:2:2).


A. One person, mainly head and shoulders.


B. One person with graphics and/or more detail..


C. More than one person.


D. Graphics with pointing.


E. High object and/or camera motion (e.g., broadcast TV).

Figure 1. Sample frames of the 25 SRC.

The T1A1 subjective test plan [2] specifies an exact list of the 25 hypothetical reference circuits (HRC) (see Table I). T1A1 used the term HRC to intentionally eliminate vendor information from publications. The HRCs included hardware coder/decoder pairs, a VHS tape dub, and the null impairment (i.e., the original video dubbed from one Betacam-SP recorder to another). All 25 SRC were recorded to a Betacam-SP tape and separated by mid-level grey. The entire SRC tape was played through each HRC, and the output video recorded to another Betacam-SP tape.

Some of the hardware codecs used changed the system delay and frame rate in response to coding difficulty (e.g., longer delay and lower frame rate for difficult-to-code scenes). However, the delay always varied around a single system delay. There were no rebuffering delays or other mean delay changes. The T1A1 video clips include a wider range of dynamic frame rate changes than are found in modern codecs, for example dropping to ≈1 fps during high motion.

The SRC as played into the encoder had three seconds of extra content at the beginning. The extra SRC content ensured that encoding problems and errors would not cause mid-level grey to propagate into the sequence.

Table 1. HRC Descriptions

HRC	*Algorithm (vendor)*	*Resolution*	*Total kbps*	*Audio kbps*	*Video kbps*	*Coding Mode*	*Frame Rate*	*FEC*	*Burst Errors*
1	Null	—	—	—	—	—	30	—	Off
2	VHS	—	—	—	—	—	30	—	Off
3	Proprietary	V. High	45,000	—	—	—	—	—	Off
4	Proprietary	Med.	128	—	—	VQ	—	—	Off
5	Proprietary	High	336	—	—	VQ	—	—	Off
6	Proprietary	Med.	112	—	—	—	—	—	Off
7	Proprietary	Med.	384	—	—	—	—	—	Off
8	Proprietary	Med.	768	—	—	—	—	—	Off
9	Proprietary	High	768	—	—	—	—	—	Off
10	Proprietary	High	1536	—	—	—	—	—	Off
11	H.261 (diff)	QCIF	128	56	70.4	INTER+MC	—	On	Off
12	H.261 (same)	QCIF	128	56	70.4	INTER	10	On	Off
13	H.261 (same)	QCIF	168	48	118.4	INTER+MC	—	On	Off
14	H.261 (diff)	QCIF	384	56	326.4	INTER+MC	—	On	Off
15	H.261 (same)	CIF	112	48	62.4	INTER+MC	—	On	Off
16	H.261 (same)	CIF	128	56	70.4	INTER+MC	—	On	Off
17	H.261 (diff)	CIF	128	48	78.4	INTER+MC	—	On	Off
18	H.261 (same)	CIF	168	48	118.4	INTER+MC	—	On	Off
19	H.261 (same)	CIF	256	56	190.4	INTER+MC	15	On	On
20	H.261 (same)	CIF	384	56	326.4	INTER+MC	—	On	Off
21	H.261 (same)	CIF	384	56	326.4	INTER+MC	—	On	On
22	H.261 (diff)	CIF	768	56	710.4	INTER+MC	—	On	Off
23	H.261 (same)	CIF	768	56	710.4	INTER+MC	—	On	On
24	H.261 (diff)	CIF	1536	56	1478.4	INTER+MC	—	On	Off
25	H.261 (same)	CIF	1536	56	1478.4	INTER+MC	—	On	Off

"Null" is the original SRC recording compared to itself; "VQ" is vector quantization; "FEC" = forward error correction; "INTER" = inter-frame coding; "MC" = "motion compensation"; "Burst Errors" = bursts of bit-errors; "—" = variable not specified; "same" = same coder and decoder manufacturer; and "diff" = different coder and decoder manufacturers.

Subjective Testing

The T1A1 subjective test was conducted according to ITU-R Rec. BT.500-5 using the double stimulus impairment scale (DSIS). Although the currently in-force BT.500-13 excludes DSIS, this method appears in ITU-T Rec. P.910 under the name degradation category rating (DCR). The test was conducted using Betacam-SP tapes, written scoring sheets, and a broadcast quality CRT monitor.

The entire test includes 625 processed video sequences (PVS), which was too much for any single subject to comfortably rate. Instead, the PVSs were divided into three pools of 10 HRCs

each. Overlapping HRCs promoted consistent scoring between subject pools, but those extra scores were discarded. Three subjective labs (NTIA/ITS, GTE, and DIS) each gathered one-third of the data for each pool. The T1A1 subjective data includes ratings from 30 subjects for each PVS (i.e., ten from each lab). An analysis by Cermak and Fay [3] found that the data from these three labs were not statistically different.

The DMOS scores for HRCs 2 through 25 are available on CDVL with the video sequences. The DMOS scores for HRC 1 (null, now labeled "original") were misplaced. At this time, the raw subjective data is only available from one of the three labs (in Contribution T1A1.5/94-143 [5]).

The videos on CDVL contain, for each PVS, all available content on the HRC tape: mid-level grey frames, 3 sec pre roll, 9 sec sequence, 1 sec post-roll, and mid-level grey frames. A spreadsheet (redistributed with the video sequences) lists the following information for the 600 PVSs:

- DMOS
- Standard deviation of differential opinion scores
- Spatial shift in frame lines vertically and pixels horizontally
- Luma gain & level offset values
- Time aligned segment (start frame & stop frame)

These calibration values were calculated using the NTIA/ITS full-reference temporal registration algorithms [6], followed by a manual inspection. This algorithm finds a typical delay for the entire sequence. The time alignments used for the viewing tape edits were chosen by eye and so differ slightly from the spreadsheet values.

Conclusion

Is the T1A1 dataset valuable today, since it examines 20 year old technology? The NTIA/ITS philosophy is to encourage the development of technology independent metrics. If an

objective model is rooted in the image receptors of the eye, and the visual cortex and image processing centers of the brain, then it should be accurate for the T1A1 dataset. Such flexibility indicates resilience: an objective model whose performance will degrade gracefully as coding technology continues to change.

References

[1] M. Pinson, "The Consumer Digital Video Library [Best of the Web]," IEEE Signal Processing Magazine, vol. 30, no. 4, pp. 172-174, Jul. 2013. doi: 10.1109/MSP.2013.2258265

[2] T1A1.5/94-118 R1 "Subjective test plan (tenth and final draft)," Detailed Test Plan Ad Hoc Group and Data Analysis Ad Hoc Group, (A.C. Morton, Editor) 3 Oct. 1993. See ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/

[3] T1A1.5/94-148 "T1A1.5 video quality project: GTE Labs Analysis," GTE Laboratories, Inc. (Gregory W. Cermak and David A. Fay), 20 Sep. 1994. See ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/

[4] T1A1.5/94-152 "Analysis of T1A1.5 subjective and objective test data," NTIA/ITS (Coleen Jones, Ned Crow, Stephen Wolf, Arthur Webster), 3 Oct. 1994. See ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/

[5] T1A1.5/94-143 "DIS/NCS subjective test data," Delta Information Systems / National Communications Systems (DIS/NCS), 14 Jul. 1994. See ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/

[6] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Transactions on Broadcasting, vol.50, no.3, pp. 312-322, Sept. 2004. See http://www.its.bldrdoc.gov/publications/2576.aspx.

Margaret Pinson (top) and Arthur Webster (bottom) participated in the T1A1 validation test as proponents. The NTIA/ITS General Model for video quality (VQM) was trained on this dataset. Both are with the National Telecommunications and Information Administration, Institute for Telecommunication Sciences in Boulder, Colorado.

Multimedia Quality of Experience for Target Recognition Applications

Mikołaj Leszczuk and Lucjan Janowski

Introduction

A decade ago, the telecommunications industry believed that high-performance Quality of Service (QoS) techniques would resolve any recurrent problems of low-quality multimedia services. However, within a few years, it became clear that optimization of QoS parameters such as throughput, packet loss, delay, or jitter is not the best way of improving the quality experienced by users. The problem of low bandwidth can be compensated for by more efficient codecs. The impact of packet loss is strongly dependent on their distribution, and the use of redundancy coding and transmission. For many applications, buffering multimedia data streams can alleviate major delays and jitter.

Since discovering that QoS is not an adequate metric of network quality, most proposals have suggested that quality should be measured at the user level. This process was named Quality of Experience (QoE) [1] [2]. Such measurement calls for special structures (frameworks) for integrated assessment of the quality of video sequences [3]. These structures are increasingly being filled with solutions that attempt to model overall quality, operating at the intersection of QoS and QoE areas [4] or only in the area of QoE. However, it has become obvious that such a general approach simply does not work for many visual applications such as target recognition (utility) applications (video surveillance, telemedicine/remote diagnostics, fire safety, backup cameras, games, etc.) [5] [6].

In fact, QoE – the way quality of multimedia services is perceived – depends on a number of objective and subjective contextual parameters [7]. Only a full understandingof these parameters makes it possible to obtain results consistent with the expectations of service users, and, consequently, to optimize quality, but that is usually only possible when strong limitations are placed on the QoE modelling application. [8]. Unfortunately, the large number of contextual parameters means this research question is still open.

Target Recognition Video

In many visual applications, the quality of the motion picture is not as important as the ability of the user to perform specific tasks for which the visual system was created, given the processed video sequences. Such sequences are called Target Recognition Video (TRV). Regardless of the different ways in which the concept of TRV quality is understood, its verification is necessary to perform dedicated quality testing. The basic premise of these tests is to find TRV quality limits for which the task can be performed with the desired probability or accuracy.

Such tests are usually subjective tests (psychophysical experiments) with a group of subjects. Unfortunately, due to the complexity of the issue and our relatively low degree of understanding of human cognitive mechanisms, computer modelling of TRV quality has not yet achieved satisfactory results beyond very limited areas of application.

Given the use of TRV, qualitative tests do not focus on the subject's satisfaction with the quality of the video sequence, but instead they measure how the subject uses TRV to accomplish certain tasks. Purposes of this may include:

- Video surveillance – recognition of vehicle license plate numbers
- Telemedicine/remote diagnostics – correct diagnosis

- Fire safety – fire detection
- Backup cameras – parking the car
- Games – spotting and correctly reacting to a virtual enemy

The human factor is a significant influence; therefore it is necessary to ask questions on the procedures that must be followed to make a subjective assessment of TRV quality. In particular, questions arise on:

- Method of selecting the TRV source from which the test TRV (with degraded quality) arises
- Subjective testing methods and the general manner of conducting the psychophysical experiment
- Method of selecting a group of subjects in the psychophysical experiment, especially identification of any prior knowledge of the task
- Training subjects before the start of the experiment
- Conditions in which the test will be carried out
- Methods of statistical analysis and presentation of results

Methods for Subjective Evaluation of TRV

The questions formulated in the previous section are addressed by Recommendation ITU-T P.912 [9] "Subjective Video Quality Assessment Methods for Recognition Tasks", published in 2008. In addition, Recommendation P.912 organizes terminology related to subjective TRV testing, introducing appropriate definitions for the methods of testing (psychophysical experiments).

Unfortunately, Recommendation P.912 is only the first step in the standardization of methods of subjective TRV testing. In the opinion of the authors, based on research results (their own and independent) and observations conducted during numerous experiments with TRV, many claims of Recommendation P.912 are formulated at too high a level of

generality. What's more, selected statements are not supported by research results and are significantly disputable. In this situation, the authors propose amendments to Recommendation P.912. We would like to invite all researchers working on TRV-related topics to join us in the process of improving P.912.

References

[1] E. Cerquiera, S. Zeadally, M. Leszczuk, M. Curado, A. Mauthe, "Recent advances in multimedia networking", *Multimedia Tools and Applications*, vol. 54, pp. 635-647, 2011

[2] M. Grega, L. Janowski, M. Leszczuk, P. Romaniak, Z. Papir, "Estimation of the perceived quality of service (QoE) for multimedia communication", *Review of Telecommunications, Telecommunication News*, vol. 81, p 142-153, 2008

[3] M. Mu, P. Romaniak, A. Mauthe, M. Leszczuk, L. Janowski, E. Cerquiera, "Framework for the integrated video quality assessment", *Multimedia Tools and Applications*, vol. 61, pp. 787-817, 2012

[4] M. Leszczuk, L. Janowski, P. Romaniak, Z. Papir, "Assessing quality of experience for high-definition video streaming packet loss under diverse patterns", *Signal Processing-Image Communication*, vol. 28, pp. 903-916, 2013

[5] M. Leszczuk, "Image quality for utility applications: definitions, testing, standardization and current trends", *Review of Telecommunications, Communications News*, vol 83, pp. 242-246, 2010

[6] S. Moeller, A. Raake (ed.), *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, Berlin 2013

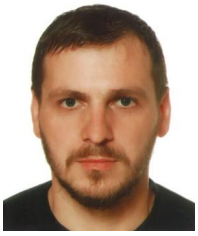
[7] P. Le Callet, S. Moeller, A. Perkis (ed.), *QUALINET White Paper on QoE Definitions*, European Network on QoE in Multimedia Systems and Services (COST Action IC 1003), Dagstuhl 2012

[8] M. Leszczuk, "Optimizing task-based video quality: a journey from psychophysical experiments Subjective

quality is objective optimization", *Multimedia Tools and Applications*, vol. 68, pp. 41-58, 2014

[9] ITU-T *P.912: Subjective video quality assessment methods for recognition tasks*, Geneva 2008

Mikołaj Leszczuk, PhD, is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology (AGH-UST), Krakow, Poland. He has participated actively in several national and European projects. He is a member of the VQEG Board and a co-chair of VQEG QART (Quality Assessment for Recognition Tasks) and MOAVI (Monitoring of Audio Visual Quality by Key Indicators) Group. *The Polish National Centre for Research and Development funds his work, Contract No. C2013/1-5/MITSU/2/2014 under the international EUREKA programme: MITSU – Next Generation Multimedia Efficient, Scalable and Robust Delivery.*



Lucjan Janowski *is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology, in Krakow, Poland. He is a Co-Chair of the VQEG JEG-Hybrid project (http://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx).*

A New Subjective Audiovisual & Video Quality Testing Recommendation

Margaret H. Pinson and Lucjan Janowski

Introduction

ITU-T Rec. P.913 is a new subjective video quality testing standard that was approved in January 2014. This Recommendation focuses on the evaluation of flat screens, laptops, and mobile devices. P.913 emphasizes flexibility of environment, rating scale, display technology, and stimulus modality (video, audio, or audiovisual). To balance this flexibility, P.913 includes mandatory reporting requirements.

This paper introduces ITU-T Rec. P.913. The reader is assumed to have some knowledge of subjective video quality testing. Pinson et al. [1] provides a suitable tutorial on this topic. ITU-T Rec. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment." is freely available on-line at http://www.itu.int/rec/T-REC-P.913/en.

Environment & Reporting

When testing consumer grade devices, most aspects of the viewing environment have a minimal impact on mean opinion score (MOS) [2]. Consequently, P.913 does not rigidly constrain the environment and does not include monitor calibration procedures. Instead, the experimenter chooses an environment that is suited to the experiment. This alternate paradigm encompasses distracting environments, monitors

that cannot be calibrated, mobile devices that only play highly compressed signals, questions that can only be answered using modified rating scales, and mixed evaluation of video and audio.

P.913 includes two environment choices: controlled and public. A controlled environment is non-distracting: a comfortable and quiet room that is devoted to conducting the experiment. Examples include a sound isolation chamber, a laboratory, a simulated living room, a conference room, or an office. The P.913 controlled environment allows experimenters to choose an environment where the subject could imagine using the device under test. Lighting is chosen by the experimenter to suit their situation.

A public environment intentionally includes distractions. A public environment can change over time or include people not involved in (or unaware of) the experiment. Examples include a cafeteria, a bus, a busy office, the subject's home, and an otherwise controlled environment with intentionally distracting background noise (e.g., crowd noise, traffic noise, sirens). A public environment should represent a distracting environment where a person would reasonably use the device under test.

The importance of the public environment can be seen in Harrison et al. [3]. This literary overview summarizes a large variety of studies that evaluate the usability of mobile applications. Of the 163 studies discussed in [3] and conducted from 2008 to 2010, 50% were performed in controlled environments, and 27% were field studies.

Because the experimenter has full control of the environment choice, P.913 mandates that subjective test results carefully document the environment. The report should include:

- a picture of the environment
- type of environment (controlled or public)
- noise level (e.g., quiet, bystanders talking)

- lighting level measured in lux
- viewing distance in picture heights
- type and size of video monitor
- type of audio system
- placement of speakers

Also, a full description may not be possible; for example, if each subject takes a mobile device to their home. Depending upon the type of stimuli, some of these values may be inapplicable.

Types of Stimuli

Quality evaluations of mobile devices and modern video systems can include multiple types of stimuli.The subjective quality test methods used for video are very similar to those used for speech and audio (see for example ITU-T Rec. P.800, ITU-R Rec. BS. BS.1534). One option is to design a series of experiments, as suggested in ITU-T Rec. P.1301, "Subjective quality evaluation of audio and audiovisual multiparty telemeetings."

Another option is to design a single experiment that includes multiple stimuli, and P.913 encompasses this solution. P.913 can be applied to video-only stimuli, audio-only stimuli, audiovisual stimuli, and 3D video stimuli. These can be evaluated in separate sessions or mingled into a single session. Naturally, other ITU Recommendations are better suited to experiments that only evaluate speech or audio quality. Special consideration for 3DTV subjective tests is the focus of several Recommendations that are nearing completion.

Vision Testing

BT.500 and P.910 require that all subjects have normal visual acuity (e.g., on a Snellen chart) and normal color vision (e.g.,

using Ishihara plates). By contrast, the visual screening of subjects is optional within P.913.

We are not aware of a definitive study that analyzes the impact of abnormal visual acuity and/or abnormal color vision on subjective quality ratings. Cermak and Fay [4] analyzed the T1A1 dataset's 625 processed video sequences (PVS) and 114 subjects. They concluded that visual acuity and color vision should not be used to screen subjects, because those subjects' data was not significantly different from the rest of the population's data. This hypothesis is supported by [2] and private communication from Cermak describing later experiments. Bovik [6] questions the validity of vision screening, because the general population includes people with normal vision and people with impaired vision. The usual goal of behavior research is to choose a pool that is representative of the general population.

P.913 leaves the choice of visual screening to the researcher, based upon the purpose of the experiment. Visual screening may be desirable when fine tuning compression algorithm improvements yet undesirable when performing a cost / benefit analysis on a product.

Rating Scales

P.913 includes four rating scales that answer different questions (see Fig. 1):

- Absolute category rating (ACR): the subject views one video sequence, then rates the quality on a 5 level scale (excellent, good, fair, poor, bad).

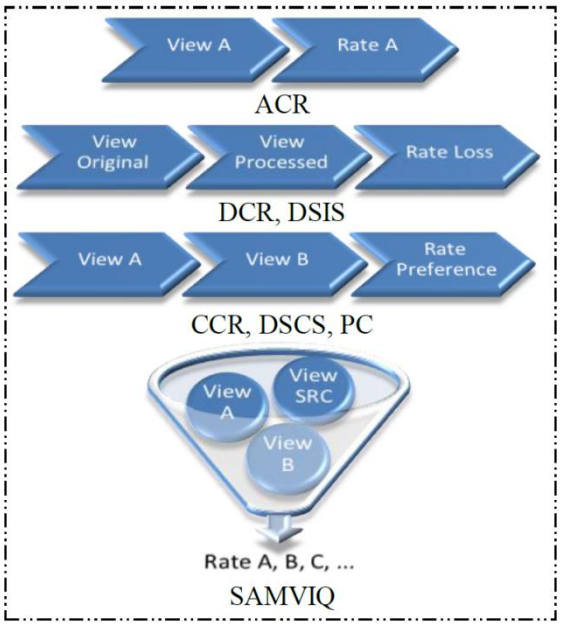


Figure 2. Rating sequence is shown for the four subjective scales in ITU-T Rec. P.913.

- Degradation category rating (DCR) method, also known as the double stimulus impairment scale (DSIS) method: the subject views the original video, views the processed video, and then rates the amount of impairment perceived on a 5 level scale (imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying).
- Comparison category rating (CCR) method, also known as the double stimulus comparison scale (DSCS) or as pair comparison (PC): two versions of the same source video sequence are viewed in a random order, then the subject rates the second sequence relative to the first on a 7 level scale (much worse, worse, slightly worse, same, slightly better, better, much better).
- ITU-R Rec. BT.1788 (SAMVIQ) and ITU-R Rec. BS.1534 (MUSHRA): a computer interface presents multiple versions of the same source stimuli. The subject may play each stimulus multiple times and chooses the order in which stimuli are rated. SAMVIQ and MUSHRA use a continuous scale with ACR labels.

Each method has a unique design goal. ACR focuses the subject on the task of rating one stimulus in isolation. DCR is an explicit comparison between the reference and impairment. PC allows a direct comparison between two impaired stimuli. SAMVIQ and MUSHRA allow multiple stimulus ratings to be adjusted relative to each other.

P.913 acknowledges that some experiments require modifications to these methods. Some modifications are explicitly identified as acceptable, because prior studies have proven their reliability.

Alternate wording of level labels is the first accepted modification. ITU-T Rec. P.800 has long specified two alternate wordings of the 5 level ACR scale for speech quality tests: listening effort and loudness preference. The MPEG video compression testing [7] used DCR with ACR labels excellent, good, fair, poor, and bad. Other examples are translating into another language, using an unlabeled scale (e.g., endpoints are marked with "+" and "-"), and using a scale with numbers but no words.

Zielinski et al. [8] examines multiple sources of subjective test bias, including prior studies into the impact of the words associated with rating levels. The translation of level descriptors into multiple languages raises a concern that the translated level descriptors will have different distributions in terms of linguistic quality meanings, and that this could bias the MOS ratings. Contrary to this expectation, [8] found that the differences between labeled and unlabeled scales were "negligibly small," indicating that this fear is unfounded. Zielinski theorizes that subjects ignore the verbal level descriptors and either interpret the levels linearly or only take the end points into account. Pinson et al. [2] was also unable to find language or culture based biases. The apparent biases indicated by speech quality experiments, such as Cai et al. [9], can be explained by the use of different speech samples by each lab.

A second accepted modification is ACR with hidden reference (ACR-HR). The source stimuli are rated, and a differential mean opinion score is calculated between the original and processed ACR values. The Video Quality Experts Group (www.vqeg.org) successfully used ACR-HR to validate video quality models. These efforts resulted in ITU-T Rec. J.247, J.246, J.340, and J.341, as well as ITU-R Rec. BT.1866 and BT.1867. This ACR variant has proven value when the choice of method must be a compromise between competing priorities. Examples include measuring difference MOS (DMOS) yet minimizing session duration, and evaluating no-

reference and full-reference objective video quality models on the same subjective dataset. See [1] for more information about the advantages and disadvantages of ACR-HR method.

Increasing the number of levels is discouraged but allowed. An example is implementing ACR as a 9 level, 11 level, or continuous scale. Huynh-Thu et al. [10] and Tominaga et al. [11] compared discrete scales with different numbers of levels (e.g., 5 level, 9 level, 11 level) with continuous scales (e.g., 100 point scales). These studies concluded that continuous scales contain more levels than can be differentiated by people. Increasing the number of discrete levels did not improve the accuracy of the MOS or the corresponding confidence interval. An increase in the number of levels was detrimental, in that the rating task is slower and more cognitively difficult [11].

New Best Practices

Testing of mobile video devices usually requires lossy video playback. That is, the mobile device's video playback introduces quality impairments on the stimuli. P.913 allows for the use of lossy video playback when no alternative exists. Such lossy playback impairments will confound the data being measured, which must be considered during the data analysis.

The detrimental impact of a distracting environment is a reduction in accuracy. P.913 compensates by increasing the number of subjects. Based on [2], P.913 recommends that 24 or more subjects should be used when ACR, DCR, or PC are conducted a controlled environment. This increases to 35 subjects when using a public environment or a narrow range of audiovisual quality. Based on a study by Péchard et al. [12], a minimum of 15 subjects should be used for SAMVIQ and MUSHRA. For any method, smaller numbers of subjects are suitable for pilot studies, to find trending.

Improved procedures for subjective video quality testing have been developed over the last decade of validation tests

performed by the Video Quality Experts Group (VQEG) and the International Telecommunication Union (ITU). These are included in P.913.

- Intermittent impairments should be avoided during the first 1 sec and last 1 sec of a video sequence. These may not be perceptible as impairments in the artificial environment of a subjective test.
- Subjects may be screened (rejected) by calculating the Pearson linear correlation between each subject and MOS calculated from all subjects. If a subject has a low correlation, their data is discarded. The ITU-R Rec. BT.500 screening method is also allowed.
- Long and short stalling events are perceived differently (e.g., 5 sec versus 0.5 sec). Special care should be taken with the instructions, to avoid differences in subject rating behaviors. For example, one subject could assume rebuffering, while another assumes an unintended problem with the subjective test video playback system.

Basic ethical principles should be considered in any experiment involving human testing. In the U.S., the legal requirement for informed consent resulted from the Belmont Report [13]. Informed consent refers to a document that tells subjects of their rights and gives basic information about the experiment. P.913 lists the information that would typically be included and provides an example.

Conclusions

Researchers are encouraged to try the methods standardized in ITU-T Rec. P913 and send the authors feedback on what they liked and disliked, either informally or formally. Question 12 of ITU-T Study Group 9 welcomes contributions that identify improved methods for conducting subjective testing of modern video devices and systems. See

http://www.itu.int/en/ITU-T/studygroups/2013-2016/09/Pages/rapporteurs.aspx for contact information.

References

[1] M. H Pinson, L. Janowski, and Z. Papir, "Video quality subjective testing of entertainment scenes," IEEE Signal Processing Magazine, January 2015.

[2] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: an international study," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, Oct. 2012, pp. 640–651.

[3] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," Journal of Interaction science, vol. 1, 2013.

[4] T1A1.5/94-148, "Correlation of objective and subjective measures of video quality," GTE Laboratories Inc. (G.W. Cermak and D. A. Fay), Sept. 20, 1994.

[5] T1A1.5/94-118 R1, "Subjective test plan (tenth and final draft)," AT&T Communications (A. C. Morton), Oct. 3, 1993. Available: ftp://vqeg.its.bldrdoc.gov/Documents/OLD_T1A1/

[6] A. K. Moorthy, L. K. Choi, A.C. Bovik and G. de Veciana, "Video quality assessment on mobile devices: subjective, behavioral and objective studies," IEEE Journal of Selected Topics in Signal Processing, vol.6, no.6, p.652-671, Oct. 2012.

[7] C. Fenimore, V. Baroncini, T. Oelbaum, and T. Tan, "Subjective testing methodology in MPEG video

verification,"' SPIE Conference on Applications of Digital Image Processing XXVII, 2004.

[8] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests—a review," Journal of Audio Engineering Society, vol. 56, no 6, Jun. 2008.

[9] Z. Cai, N. Kitawaki, T. Yamada, and S. Makino, "Comparison of MOS evaluation characteristics for Chinese, Japanese, and English in IP Telephony," 4th International Universal Communication Symposium (IUCS), Oct. 2010.

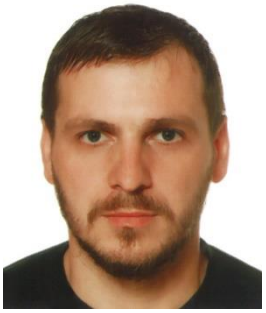
[10] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," IEEE Transactions on Broadcasting, vol. 57. No. 1, p. 1-14, Mar. 2011.

[11] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," Quality of Multimedia Experience (QoMEX), Jun. 2010.

[12] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm." IMQA 2008. Available on-line: http://www.mi.tj.chiba-u.jp/imqa2008/

[13] U.S. Department of Health & Human Services, "Ethical Principles and Guidelines for the Protection of Human Subjects of Research," Apr. 18, 1979. http://www.hhs.gov

Margaret Pinson is an Associate Rapporteur of Questions 2 and 12 in ITU-T Study Group 9. She was the editor for ITU-T Rec. P.913. She investigates improved methods for assessing video quality at NTIA/ITS, in Boulder, Colorado, USA.



Lucjan Janowski is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology, in Krakow, Poland. He is a Co-Chair of the VQEG JEG-Hybrid project (http://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx).

New ITU-T Rec. P.1302 for Audio and Audio-visual Call Quality Testing

Sebastian Möller and Benjamin Weiss

The ITU-T Recommendation P.1302: "Subjective method for simulated conversation tests addressing speech and audio-visual call quality" was consented in the last SG12 (Performance, QoS and QoE) meeting and approved in October, 2014. This recommendation provides a methodology for assessing subjective quality of speech and audio-visual telephone calls with time-varying transmission characteristics. It is an extension of the ETSI TR 102 506 technical report to wide-band and audio-visual telephony.

Instead of simulating real conversations with two participants, one participant experiences conversational structures by viewing and listening to typically five prerecorded stimuli (8..12s). Between these stimuli, the participant is asked to verbally answer multiple choice questions related to the content of the stimulus just perceived. At the end of the "call," a typical ACR scale is applied to rate the quality of the whole "call" instead of answering a question.

The aim of this method is to elicit conversational structures (turn-changes), including active speech production of the rating participant, thus providing a more valid situation with attention on the content, not only on the transmission. Whereas this method currently does not allow for evaluating effects of echo or delay, it does allow for well-defined profiles of time-varying transmission characteristics. Sample material for producing the stimuli and questions is provided.

Sebastian Möller studied electrical engineering at the universities of Bochum (Germany), Orléans (France) and Bologna (Italy). He received a Doctor-of-Engineering degree in 1999 and the Venia Legendi with a book on the quality of telephone-based spoken dialogue systems in 2004. In 2005, he joined Deutsche Telekom Laboratories, TU Berlin, and in 2007, he was appointed Professor for Quality and Usability at TU Berlin. His primary interests are in speech signal processing, speech technology, and quality and usability evaluation.

Benjamin Weiss studied communication science and phonetics, educational studies and Scandinavian studies at the universities of Bonn, Trondheim and Berlin. After his graduation in 2002, he was with the "Graduiertenkolleg" at the Linguistics department at Humboldt University Berlin, doing his dissertation on speech tempo and pronunciation. He received his Ph.D. in Linguistics in 2008. Currently, he is working on Vocal Likeability, Speech Transmission Quality and Human-Computer Interaction at the Quality & Usability Lab.

Blind Image Quality Assessment: Unanswered Questions and Future Directions in the Light of Consumers Needs

Michele A. Saad, Patrick Le Callet and Philip Corriveau

Motivation

Are proposed no-reference models accurate enough to be standardized for all use cases? What remains to be solved?

This past decade has seen significant progress in the field of image and video quality assessment. While full- and reduced-reference models (for images and videos), which emerged earlier than blind/no-reference ones, have managed to achieve significant quality prediction accuracy as measured by correlations with subjective quality ratings, there is still much progress to be made within the no-reference realm. In addition, the use cases covered by most standardization efforts are largely related to the content delivery chain, excluding acquisition and enhancement issues, and focusing more on compression or transmission impairments. The industry has been demanding the move towards blind assessment with the hope of being unshackled from requiring a reference. With the overwhelming ecosystem that now supports acquisition and consumption of media on a myriad of devices and context (e.g. viewing conditions) this move becomes even more urgent.

Indeed, some promising approaches to blind quality assessment have been proposed. These methods include, but are not limited to, LBIQ [Tang H, (2011)], CBIQ [Ye P. et al. (2011)], BLIINDS-II [Saad M. et al. (2012)], and NIQE [Mittal A. et al. (2013)]. These methods perform well on the databases on which they have been developed and on similar types of

images and distortions. The generalizability of their performance on many types of consumer images breaks however, for understandable reasons. These methods were developed and designed to achieve competitive performance on existing databases such as [Ponomarenko N. et al. (2013), Sheikh et al. (2005)], which are designed for quality assessment research; they should not be expected to perform well on images that are significantly different from images in those databases. These databases however, do not contain many of the distortions that are expected in many of the rapidly increasing consumer devices (most notably mobile devices as well as compact and higher end cameras). These databases that are the most widely used for algorithm design do not describe many of the very popular consumer usage models—the millions of images captured by mobile devices, for instance. Images and videos from consumer devices contain multiple distortions that are very complex in nature. These distortions may be a result of the optical system, the post processing that happens in the devices after the signal is captured, and the storage and display of the data. Simulating these distortions collectively is an extremely challenging task, and it would be necessary to create a comprehensive corpus of image distortions if one is to design algorithms for these types of images. Creating this database faces hurdles in and of itself due to privacy and sharing rights, and the requirement for a constantly new pool of test content to validate new models.

Existing subjective testing datasets for quality assessment are not suitable to validate NR models that test consumer devices.

Unanswered Questions

Going beyond fidelity: revising the methodologies

In full- and reduced-reference problems the question addressed is essentially that of fidelity: how close is a test image/video to a reference one. In blind assessment, on the other hand, prior to predicting quality, one needs to define what "better quality" is. This is all the more critical when

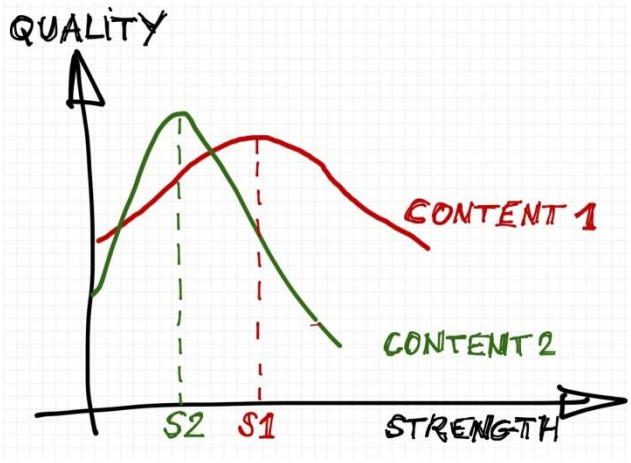


Figure 1. **The overshoot effect**: adjusting the strength for an enhancement processing (e.g. sharpening) may lead to quality improvement until a point that overshoot is reached, leading to a decrease of quality as the strength of the processing increase. Optimal strength is hard to estimate as it is often content dependent, as illustrated in this plot.

evaluating the effects of post processing, such as image enhancement, that should lead to an improvement in perceptual quality over the original image. Major challenges lie in trying to solve this problem, the primary one being that of content-dependency. No-reference predicted scores tend to be biased by content. Two images or videos of similar qualities should ideally get similar scores even if the content is highly different (for instance a low frequency sky scene versus a high frequency forest scene). The overshoot effect (see Figure 1) is another issue that needs to be accounted for by a blind quality evaluator. However, the decoupling of content from quality is a very challenging issue in blind quality assessment that still needs to be resolved.

Further, while certain methods can achieve relatively high correlations on databases that span a wide range of qualities from extremely bad (not necessarily always representative of what consumers encounter in real life) to excellent, how well proposed methods perform on a narrower range of qualities (typically a range in the higher quality end) is important for more realistic predictions on consumer content. This is illustrated in Figure 2, and is referred to as "the range effect". Addressing this issue might require revisiting subjective testing methodologies for NR model evaluation. For instance, usual subjective test methodologies such as ACR or DSCQS require a very large number of observers per condition before exhibiting statistically significant differences. Pair comparison methods might be good alternatives to achieve better sensitivity.

In full- and reduced-reference problems, the question addressed is essentially that of fidelity: how close is a test image/video to a reference one. In blind assessment on the other hand, prior to predicting quality, one needs to define what is meant by pristine or perfect quality.

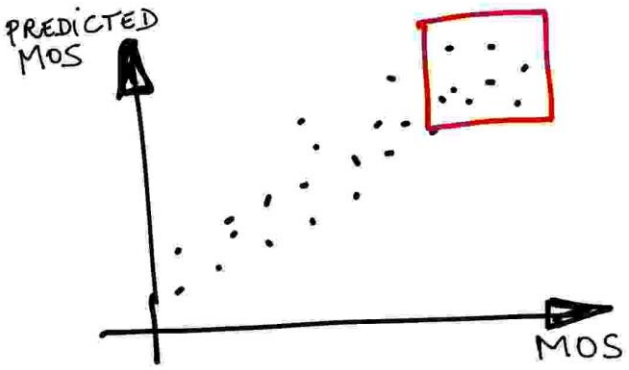


Figure 2. **The range effect**: on the overall quality range, MOS and predicted MOS seem well correlated; focusing on a particular range (e.g .points within red box), the correlation is lower.

Color is another domain where perceptual modeling for the purpose of quality assessment is lacking. The complexity of the human visual system's processing of color information has made understanding the effect of color aberrations (as opposed to only structural ones) difficult to model and predict in the no-reference quality prediction space.

Towards user profile

When it comes to image capture, another important factor has to be taken into account: the intention of the image taker. Blur, for instance, which is typically considered a distortion in image quality assessment, is often introduced on purpose by professional photographers. The distinction between artistic and undesirable effects of blur is a higher level problem that needs to be better understood and modeled. This also applies to other types of artistic effects such as film grain and motion blur.

What makes a good picture is highly subjective and very unique. Trying to capture all these effects through the prism of "king MOS" or a general quality metric may lead to a "grey car effect" (a situation in consumer science where simply averaging opinions may lead to a trade-off (but faulty) conclusion that only grey cars should to be produced since this averages out preference for black and white cars!) With this in mind, one might consider blind image quality tools tuned to specific user profiles or needs: one could imagine parameterized measures instead of general agnostic tools.

These are just a few of the issues pertaining to images. All of the mentioned "unanswered questions" hold for blind video quality assessment. Video on the other hand, exponentially increases the complexity of the search space. Motion modeling has for a long time now been an open area of research and

Michele A. Saad is a Senior Engineer and Researcher in perceptual image and video quality assessment at Intel. She received her Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2013, the B.E. degree in computer and communications engineering from the American University of Beirut, Lebanon, in 2007, and the M.S. degree in electrical and computer engineering from the University of Texas at Austin in 2009. Her research interests include statistical modeling of images and videos, motion perception, design of perceptual image and video quality assessment algorithms, and statistical data analysis and mining and machine learning.

Patrick Le Callet is full professor at Ecole Polytechnique de l'Université de Nantes. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. He is currently co-chairs the "3DTV" activities and the "Joint-Effort Group", driving mostly High Dynamic Range topic in this latest.

Philip J. Corriveau is a Principal Engineer in Experience Development and Assessment in SMG at Intel. Philip received his Bachelors of Science Honors at Carleton University, Ottawa Canada in 1990. He immediately started his career at the Canadian Government Communications Research Center performing end-user subjective testing in support of the ATSC HD standard for North America. In January 2009 he was awarded a National Academy of Television Arts & Science, Technology & Engineering Emmy® Award for User Experience Research for the Standardization of the ATSC Digital System. He now directs a team of human factors engineers conducting user experience research across Intel technologies, platforms and product lines. Philip is currently on the board of directors for the School of Computing at Clemson University, on the UF CISE Industrial Advisory Board and an Adjunct Professor at Pacific University. He was a founding member of and still participates in VQEG.

understanding its effect on perceptual quality is yet to be better understood and modeled. Similar to the problem of image quality assessment, a few approaches have been proposed to assess the quality of video, but the generalizability of these approaches still has a way to go before we reach a solution generalizable enough to be standardized.

A new group within the Video Quality Experts Group has been formed so that advances can be made in these identified challenge areas. The focus is on understanding and driving solution spaces with blind and no-reference models. You all are encouraged to join, follow, and contribute to moving the needle on creating, validating, and standardizing these new models.

References

Mittal A. et al. (2013), "Making a 'Completely Blind' Image Quality Analyzer" in *IEEE Signal Process. Lett*., vol. 21, no. 3, pp. 209-212.

Ponomarenko N. et al. (2008, 2013), "Tampere Image Database", [online]: http://www.ponomarenko.info/tid2013.htm.

Saad M.A. et al. (2012), "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality" in *IEEE Trans. Image Process*., vol. 21, no. 8, pp. 3350-3364.

Sheikh H. R. et al. (2005), "LIVE Image Quality Assessment Database Release 2", [online]: http://live.ece.utexas.edu/research/quality.

Tang H., et al. (2011), "Learning a blind measure of perceptual image quality," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*., pp. 305-312.

Ye P. et al. (2011), "No-reference image quality assessment using visual codebook," *Proc. IEEE Int. Conf. Image Process*., pp. 3089-3092.

Meeting & Conference Announcements

VQEG's 3DTV group held a special session meeting December 8-10 in Nantes, France. Edits were marked on all three draft new Recommendations on 3D. This special session meeting was very important for VQEG's work, and the participants identified such a large amount of edits that there was not sufficient time to insert all needed new text and changes.

Participants agreed to further improve identified parts of the recommendations until the February VQEG meeting. We hope to finalize the three documents at that point and forward our recommended edits in a liaison to ITU-T SG9.

The next full VQEG meeting will be http://www.its.bldrdoc.gov/vqeg/meetings/santa-clara-california-usa-february-23-27-2015.aspx, in California, USA.

The 9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) will be February 5-6, 2015, in Chandler, Arizona, USA.

The 7th International Workshop on Quality of Multimedia Experience (QoMEX) will be held May 26-29, 2015, in Costa Navarino, Messinia, Greece. The paper submission deadline is February 20, 2015.