VQEG eLetter

Volume 1, Issue 1 • March 2014

Contents

Storyboard ……Margaret Pinson and Naeem Ramzan 2
Practice Sessions for Subjective Speech Quality Tests ……Stephen Voran 5
On Training the Crowd for Subjective Quality Studies …… Tobias Hossfeld 8
Viewer Training in Subjective Assessment ……Vittorio Baroncini 12
Training Session for Task Recognition……...Lucjan Janowski and Mikołaj Leszczuk 17
To Train or Not To Train? …Nicolas Staelens, Wendy Van den Broeck, and Filip De Turck..... 21
On viewing distance and visual quality assessment in the age of Ultra High Definition TV ……Patrick Le Callet and Marcus Barkowsky 25
Comparison of Metrics: Discrimination Power of Pearson's Linear Correlation, RMSE and Outlier Ratio ……Greg W. Cermak 31
Progress of the Monitoring of Audio Visual Quality by Key Indicators (MOAVI) Project …Mikołaj Leszczuk, Silvio Borer, and Emmanuel Wyckens 44
Information for Authors……48

Boulder VQEG meeting, January 2014

Storyboard

Margaret Pinson and Naeem Ramzan, Editors

Video Quality Experts Group (VQEG) provides an open forum where video quality experts meet to advance the field of video quality assessment. Over the years, VQEG has developed a systematic approach to validation testing and made ten subjectively rated video quality datasets available freely for research and development purposes.

In late 2013, the VQEG agreed to begin this eLetter. The goal is provide timely updates on recent developments, hot research topics, and society news in the area of video quality, including:

- Technical papers
- Summary / review of other publications
- Best practice anthologies
- Reprints of difficult to obtain articles
- Response articles

VQEG wants the eLetter to be interactive in nature. Readers are encouraged to respond to articles appearing in a prior VQEG eLetter.

Best Practices for Training Sessions

This eLetter focuses on "best practices" for training sessions during a subjective video quality test. It is the great honour of the editorial team to have five leading research groups, from both academia and industry laboratories, to report their insights on this topic.

"Practice Sessions for Subjective Speech Quality Tests" by Stephen Voran from the Institute for Telecommunication Sciences in Boulder presents the importance to prepare subjects to participate in subjective speech quality tests. The

importance of fully working test equipment and practice sessions is discussed in detail.

"On Training the Crowd for Subjective Quality Studies" by Tobias Hossfeld from University of Würzburg presents new possibilities for quality evaluation by conducting subjective studies with the crowd of Internet users. The challenges of conducting training sessions for different methods of crowd sourcing are also elaborated in the article.

"Viewer Training in Subjective Assessment" by Vittorio Baroncini of Fondazione Ugo Bordoni (FUB) questions the use of identical written instructions for all subjects. The paper advises researchers to monitor subjects and provide feedback to establish an improved understanding of the subjective scale.

"Training Sessions for Task Recognition" by Lucjan Janowski and Mikołaj Leszczuk from AGH University present lessons learned from different task recognition tests. This paper explores the dramatically different impact of a specialized audience and task on training.

"To Train or Not To Train?" by Nicolas Staelens, Wendy Van den Broeck, and Filip De Turck from University of Ghent, discusses the differences between the results obtained in controlled labs and real-life environments. The article questions whether or not practice sessions are appropriate, depending upon the purpose of the experiment.

While this VQEG eLetter issue is far from delivering complete coverage on this exciting research area, we hope that these invited letters give the audience a taste of the main activities in this area, and provide them an opportunity to explore and collaborate in the related fields.

Margaret H. Pinson is a researcher at the Institute for Telecommunication Sciences (ITS) in Boulder, Colorado. She joined the Video Quality Program in 1989. Mrs. Pinson is a Co-Chair of the Audiovisual HD project, the Independent Lab Group, and the new VQEG eLetter. She is an Associate Rapporteur of Questions 2 and 12 in ITU-T Study Group 9. Mrs. Pinson administers the Consumer Digital Video Library (CDVL, www.cdvl.org).

Dr. Naeem Ramzan, FHEA, SMIEEE, MIET, received M.S in Telecom from Brest, France and PhD in Electronics Engineering form Queen Mary University of London. From 2008 to 2012 he worked as senior researcher on different EU projects. Currently, he is an Assistant Professor in Visual Communication in the University of the West of Scotland. He has been a chair/co-chair of number of special sessions and International workshops. He has served as Guest Editor of a special issue of the Elsevier Journal Signal Processing: Image Communication and IEEE COMSOC E-Letter. He is co-chair of VQEG UltraHD group and editor-in-chief of VQEG E-Letter.

Technical Papers and Reprints

In addition to the main topic, this eLetter contains three invited articles on other topics.

"On Viewing Distance and Visual Quality Assessment in the Age of Ultra High Definition TV" was contributed by Patrick Le Callet and Marcus Barkowsky of IRCCyN, Polytech Nantes, Université de Nantes, LUNAM Université. This paper reflects upon the viewing distance choices for ultra-high definition television subjective tests.

"Comparison of Metrics: Discrimination Power of Pearson's Linear Correlation, RMSE and Outlier Ratio" by Greg W. Cermak is a reprint of a VQEG contribution from 2008. This article is a comparative analysis of the metrics correlation (Pearson's R), Root Mean Square Error (RMSE), and outlier ratio metrics in the context of video quality evaluation.

"Progress of the Monitoring of Audio Visual Quality by Key Indicators (MOAVI) Project," was written by the three MOAVI co-chairs: Mikołaj Leszczuk of AGH University, Silvio Borer of SwissQual, and Emmanuel Wyckens of Orange Labs. This article lists the technical progress made by the MOAVI committee through 2013.

Finally, we would like to thank all the authors for their great contributions.

Practice Sessions for Subjective Speech Quality Tests

Stephen Voran

Motivation

Subjective testing requires careful design and execution. We must do a large amount of work before the first subject participates. This often includes pre-testing and test refinements to ensure that the test can indeed capture the required information. By the time the test is launched we have frozen the test design and all test procedures to ensure consistency. All that remains are multiple trials with multiple subjects aided by our test administrator and written instructions.

Every subject brings his or her own prior experiences, assumptions, strengths, and weaknesses to the test. The ability to include this diversity is one strength of subjective testing. But we cannot let this diversity derail the testing or jeopardize the capture of the required information.

Thus, we usually start a test with a practice session. The goals often include to:

- Verify that all equipment is working as expected and results are properly recorded
- Allow the subject to become familiar and comfortable with the test equipment and procedures
- Allow adjustments to be made if permitted (e.g., adjust volume to preferred listening level)

In some cases an additional goal is to expose the subject to some or all of the speech quality levels that are in the test.

Discussion

Depending on the details of the test, we typically include 5 to 15 trials in the practice session. A practice session is necessary, but we don't wish to use up too much of a subject's valuable time with practice. If the session is not going well we will interrupt it, resolve the issue, and then start the practice session again from the beginning.

During practice we periodically encounter subjects who need some coaching in order to effectively interact with a touch screen. We often find subjects use the practice session to adjust seating and screen positions to find the most comfortable and functional configuration. We always invite questions after the practice session. Procedural questions are addressed in full detail. Common questions involve the number of trials or the expected duration of the test. Practice makes the task at hand very concrete!

On the other hand, questions about the content or inner workings of the test or our expectations for subjects' responses are always deferred (e.g., "We can discuss that after the test is completed.") It is critical that we not influence any subject by providing information beyond the standard information that is provided to all subjects through written or scripted instructions.

Our practice sessions are not intended to "calibrate" a subject. Each subject's perceptions and opinions are valid as is, and there is no feedback path designed to influence those. However, practice sessions sometimes provide an opportunity for subjects to calibrate themselves, if they wish. Some subjects seem to conclude that the full quality scale presented should be used and thus may use the practice session to learn that range and associate it with the different points along the quality scale. If the practice session does not cover the full range of quality levels in the test, these "self-calibrating" subjects may become frustrated when encountering previously

unheard quality levels in sessions that follow the practice session. Thus we often seek to present the full speech quality range (but not necessarily every speech quality level) in the practice session

Conclusion

A practice session is important to prepare subjects to participate in subjective speech quality tests. In addition to verifying that the test equipment is working, the practice session serves to familiarize the subject with the environment and the mechanics of the test procedure. This helps to ensure that during the test itself subjects are focused on the quality assessment task and we are capturing the information we need to fulfill the purpose of the test.

Stephen Voran studies applications of signal processing to quality assessment, coding, transmission, and enhancement of speech and audio signals. He has been with the Institute for Telecommunication Sciences in Boulder, Colorado since 1990.

On Training the Crowd for Subjective Quality Studies

Tobias Hossfeld

Experiences: Crowd vs. Lab Tests

Crowdsourcing enables new possibilities for quality evaluation by conducting subjective studies with the crowd of Internet users. The advantages are the large number of test subjects, fast turn-around times, and low reimbursement costs of the participants. Further, crowdsourcing allows easily addressing additional features like diverse populations or real-life environments.

Moving the evaluation task into the Internet, however, generates additional challenges and differences from lab studies in conceptual, technical and motivational areas (Hossfeld & Keimel, 2014). Due to the remoteness of the test participants, reliability of test results requires advanced test design including consistency checks, content questions, etc. as well as statistical analysis methods such as outlier detection, as not all test conditions will be typically assessed by all subjects in crowdsourcing. Hossfeld et al. (2014) provided best practices for test design and analyzed statistical methods that lead to similar subjective results for crowdsourcing and laboratory studies, e.g. for initial delays and stalling of online video streaming (Hossfeld et al., 2012). Nevertheless, quality tests of videos compressed with H.264/AVC at different bitrates and transmission errors differed absolutely for lab and crowd studies (Hoßfeld & Keimel, 2014). The reasons for the difference may be hidden influence factors in crowdsourcing due to heterogeneous hardware like subjects' screens or improper training sessions.

Training Sessions in Crowdsourcing

The conceptual differences arise mainly from the fact that crowdsourcing tasks are usually much shorter (5-15 min.) than comparable laboratory tests and due to the lack of a test moderator. The user is guided via the web interface through the tests, including an explanation about the test itself, what to evaluate and how to express the opinion. The training of subjects is mostly conducted by means of qualification tests. Nevertheless, in case of any problems with understanding the test, uncertainty about rating scales, sloppy execution of the test, or fatigue of the test user, appropriate mechanisms or statistical methods have to be applied. Therefore it is more difficult to ensure subjects have fully understood the training, in particular as no direct feedback between supervisors and subjects is possible. Due to the short task duration in crowdsourcing, demo trials to familiarize the subject with the test structure and practice trials not included in the analysis significantly decrease the efficiency of a test and increase the costs. Hossfeld and Keimel (2014) show that without any worker training and reliability questions the results are significantly worse than with lab or advanced crowdsourcing designs. Training phases must be included in the task design!

Integrate a Feedback Channel

In general, all questions from the subjects should be answered. A feedback channel can be implemented, e.g., via comments, a contact form or forums. For allowing direct feedback, a communication chat (e.g., via social network apps) is possible, but only for small or short tests, as crowdsourcing users conduct the test whenever they want until the number of required subjects is reached.

As a side effect, this helps to increase the reputation of the test administrator, as participants tend to gather in virtual communities and share their experiences with certain tests and tasks.

Two-stage Test Design

Hoßfeld et al. (2014) propose a general recommendation for crowdsourcing quality tests, the two-stage test design. The first stage includes a simple and easy to do task which tests the reliability of users, gathers a huge subject pool, gathers (demographic) information about the users, is very short (less than 1 min.) and low paid. Also, the training session including demo and practice trials is performed in this stage. This creates a pseudo reliable, trained group of users, who will be later invited to the actual quality test, which presents the second stage, as illustrated in Figure 1. In our experiments, creating this pseudo-reliable panel increases the overall efficiency by more than 60 % in terms of costs and reliable results, which is the major argument for introducing the first stage. Nevertheless, reliability mechanisms in the second stage and post-screening are required to ensure a reliable data set. This design only works with same pool of participants gathered in first stage. Hence, a series of tests should be done in a reasonable time frame, otherwise the training session may be useless.

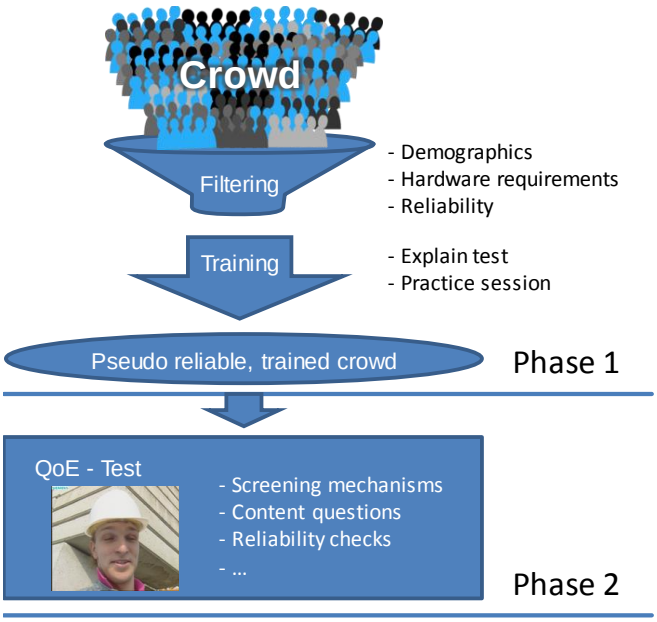


Figure 1. Two-stage design for crowdsourcing subjective studies.

In Momento Methods

Another possibility to cope with efficiency and costs of training sessions compared to actual quality tests in crowdsourcing is introduced by Gardlo et al. (2014). The basic idea of the *in momento* approach is that users are shown an *in momento* verification of their reliability and that users decide whether to stop or to continue the test, but only if a reliability

threshold is exceeded. Users who want to increase their earnings are allowed to perform additional tasks, while users who intentionally only came for one short task assignment should not be overstressed. Users may also be allowed to continue a test session after a certain time, but an upper limit needs to be specified so as not to lose the effect of the training session. As a result of their approach, the performance of the crowd in their study was significantly increased with lower overall costs and more reliable results. Nevertheless, this approach requires automated reliability mechanisms and advanced statistical output analysis of the user ratings which are even more complex than for the two-stage design.

References

Hossfeld, T., Egger, S., Schatz, R., Fiedler, M., Masuch, K., & Lorentzen, C. (2012, July). Initial delay vs. interruptions: between the devil and the deep blue sea. Proceedings of Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012).

Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., & Tran-Gia, P. (2014). Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing. *IEEE Transactions on Multimedia*, 16(2), 1-18. doi: 10.1109/TMM.2013.2291663.

Hossfeld, T., & Keimel, C. (2014). Crowdsourcing in QoE Evaluation. In S. Möller & A. Raake (Eds.), *Quality of Experience: Advanced Concepts, Applications and Methods* (pp. 323-336). Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0.

Gardlo, B., Egger, S., Seufert, M., & Schatz, R. (2014, June). Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing. Proceedings of the IEEE International Conference on Communications (ICC 2014).

Tobias Hossfeld is heading the FIA research group "Future Internet Applications & Overlays" at the Chair of Communication Networks, University of Würzburg. He finished his PhD in 2009 and his professorial thesis (habilitation) "Modeling and Analysis of Internet Applications and Services" in 2013. He has been visiting senior researcher at FTW in Vienna with a focus on Quality of Experience research. He has published more than 100 research papers in major conferences and journals and received the Fred W. Ellersick Prize 2013 (IEEE Communications Society) for one of his articles on QoE.

Viewer Training in Subjective Assessment

Vittorio Baroncini

A long time ago

When for the very first time I participated in a subjective assessment trial, it was in one of the most famous and referenced test laboratories belonging to a well-known and highly considered European Broadcaster. The impression was a little shocking to me, in that I was sitting there, 3H from the professional grade 1 studio monitor and in a carefully controlled environment, i.e., silent room with low ambient lights and no noise from outside. Everything was perfect other than a small close to irrelevant detail: what were we to do? Then we were ready to go and an old technician (wearing a white smock) came in and read to us in a grave and formal voice a short text saying "This is a visual quality experiment … … thank you for coming and good work". The text read was exactly the one reported in the ITU-R Recommendation BT.500-2. Then he went out, closing the door and the display began to show the video to assess.

In this short paper the importance of good training is stressed as mandatory practice to obtain the best and most stable results possible. Training the subjects immediately before they run a test session has the main goal of letting them understand what they have to look at and how to properly do the scoring. But not only. Participating in a subjective test is not the same as going to a cinema and watching a feature film; you might even be a little upset at having accepted a long and tedious task that will keep you there doing a "stupid task." So viewing subjects must be put in a psychologically favorable disposition. And in this, emotional involvement may certainly help. This can be achieved by explaining the importance of the experiment, to let the subjects feel that they are going to do something special for you and, most of the time, important for the whole scientific community or for the introduction of new services in the digital world.

We were seven people: three seated at 3 times and four at 4 times the height of the screen. During the test, no one was

taking notice of whether we were filling out the scoring sheets properly or commenting on the video on the screen or even joking among ourselves! This was a really shocking experience for me, but you must also consider that at those times the scores collected from 3H and 4H viewers and from all the test sequences were all put together to compute a "grand mean."

Today no one would consider this behavior a "best practice" in subjective assessment. Scores collected from viewers seated at different distances are considered separately, as are scores coming from different video clips. Though this seems quite obvious and easily sharable, not that much has been done so far to harmonize the "instructions to the viewers." All the relevant recommendations suggest to read a text to the viewers that briefly explains what they are to do and how they are to do it. The use of a training session is also recommended, as well as the use of test material that is different from that used for the test.

Now let me disagree with both of the above.

Reading a text certainly has the advantage of providing all the viewers the same information; but it may be that not all the subjects understand the text in the same way, and in any case this tends to result in an aseptic relationship between the test manager and volunteers participating as viewers. As mentioned above, testing is often boring, due to the fact that the same four (may be five) video clips are seen by the viewers so many times and sometimes with very little differences in quality among them. This demands a lot of attention of the viewing subjects, and it almost always causes a lot of frustration as the test sessions progress ("Always the same flowers!" or "Always the same train running beneath a calendar going up and down!").

So it's a "must" that the test manager engage the viewers! But the point is: how?

Certainly paying them is a good start, but it may be not sufficient. In most cases people come to your laboratory without knowing anything about testing, so you must make them feel comfortable about performing the task. But it might be useful to involve them emotionally, telling them that "this is an important experiment, this is the first time such an experiment is being done, and many industries world-wide are waiting for the results," and so on. Motivation works very well for people entering a laboratory for the first time. It happened to me that not a few of my subjects asked for a "participation certificate." So at first talk about the importance and the meaning of the test, and then begin to explain what to watch and what to do with their scoring sheet (or buttons on the screen).

This is another crucial task that, if well done, will allow you to obtain better and more stable results. This is the main role of the practice session (also called training session). You will pick some of the lowest, middle and highest quality video clips of the ones you are using for the test; these video clips will be presented using the same presentation that will be used for the actual test, to simulate the task the subjects will face.

Why do I select video clips that are part of the test set? Because it is important to show to the naïve subjects where it is preferable to look, picture by picture. This allows all the naïve subjects to respond in a more homogeneous way to the stimuli (i.e., to the impairments or improvements in the video clips). I know this is against what is written in most of the traditional literature, but it works! What you have to avoid in the editing of the training session is to show the same video clips in the same order they will be shown at the beginning or during the actual test. Also the training session will consist of not less than five but not more than eight Basic Test Cells.[1]

[1] A Basic Test Cell (BTC) is the sequence of messages and video clips presentations that allows to evaluate a single test condition (also called "test point" - TP)

Before the training session begins, explain the meaning of each grade in the scale selected for your experiment. It is important to provide to them a mental anchor for each grade. As an example for a five level scale ACR test you may explain that 5 is used when no impairment is seen and the video looks perfect, 4 is used when they see or even think they see some artifact, but in any case impairment is difficult to see, 3 is used when some artifact is easily visible, 2 is used when many artifacts are seen and they are clearly visible, and 1 is used when the picture quality is really poor.

During the training session you will stay close to the subjects, verifying that they vote at the appropriate time (not before the video clips are finished and not too late) and helping them to properly understand the meaning of a score; they must not be scared to use the full quality or impairment range available. Furthermore, I strongly recommend that you intervene when you see that a subject scores a perfect image (e.g. a source) with a "3" (or lower rate) or when a very poor quality video clip is scored with a "4" or a "5". You can also provide such comments when the training session is completed, revising the scores they have entered together with the subjects. If you see that one or more subjects made several errors, it is recommended that the training session be played again.

Remember that humans are "really different" from each other, and also what is obvious to you may be hard for people not in your industry to see. I know so many people who told me that after having participated in a test in my laboratory, they were no longer able to fully enjoy a TV program because they were seeing a lot of "impairments" and did not feel as comfortable as before when watching TV!

Last, but not least, the current literature describes how to screen viewing subject for their vision. Well, the training phase allowed me in many cases to screen the subjects for their behavior during the training. Some people who appeared "normal" clearly revealed their psychology when asked to

perform a task that was revealed as too complicate for them to complete. This comes out clearly when you see all the scores flattened to the lower or to the upper grades

Let me conclude by saying that human subjects are one of the main tools a test manager needs to have. You have to select them carefully but mainly you have to train them in the best possible way to avoid getting unstable or even unusable results.

Vittorio Baroncini is a senior researcher at Fondazione Ugo Bordoni a research institute in Rome. He is the responsible for Multimedia and TV quality assessment area. Chair of the MPEG Test Group, Vice-Chair of ITU-R WP6C and co-founder of the Video Quality Expert Group, he is author of many paper to conferences and scientific journals. He is also co-author of two books on MPEG.

Training Session for Task Recognition

Lucjan Janowski and Mikołaj Leszczuk

Task Recognition Specificity

In many applications the video quality is not as important as the ability to accomplish a specific task for which the video was created. Typical examples are: surveillance systems; a camera installed in a car helping to park; or a remote medical consultation system. A general idea behind the quality tests for task recognition is to find a threshold at which the task can be achieved with a certain probability or accuracy. Therefore, instead of the quality evaluation, the subjective experiment is focused on a task performance measurement. For example, the test might measure the probability that a license plate number is accurately recognized, a car is parked correctly, or a correct diagnosis is made. Therefore, the training session is focused on clearly describing the task description and familiarizing a subject with the test's interface. One can think that explaining a task is especially easy. A task can be described as simply as: "Please type the license number," "Please park a car," or "Please recognize if an organ is healthy," following the previous examples.

Problems of quality measurements for task-based video are partially addressed in an ITU-T Recommendation (P.912, "Subjective Video Quality Assessment Methods for Recognition Tasks," 2008) that mainly introduces basic definitions, methods of testing and psycho-physical experiments. Section 7.4 of ITU-T P.912 ("Instructions to subjects and training session") says that "The subject should be given the context of the task before the video clip is played, and told what they are looking for or trying to accomplish. If questions are to be answered about the content of the video,

the questions should be posed before the video is shown, so that the viewer knows what the task is." We followed such simple training guidance, but still some problems were found. Here we point to some errors we made running subjective task recognition experiments. Possible solutions are proposed.

Examples

The first example is a license plate recognition task. The task was: "Please write all characters which you are able to read in the text box below" [1]. Analyzing results we found that this description was not precise enough. Some subjects understood that if a character is difficult to read it cannot be read, others try hard to read all characters. As a consequence some subjects recognized just the most obvious characters and others many more of them. Of course, we cannot be sure if better training would change the results much, since we are used to observe different subjects engagement. Nevertheless, a clear training session containing a video with difficult, but possible to read, characters would make it clear for subjects that if they are not sure they should still try their best. More details about errors made by users in this and another recognition tasks experiments can be found in [2].

The second example is an experiment in which goal was object recognition. NTIA ITS performed the object recognition tests with different groups and interfaces [4] [5]. The same experiment was repeated by AGH [3]. For all those experiments, a training session showed a short video with each object and the object name. The experiments' results demonstrate that only one subject misunderstood the training and marked a radio as a mobile phone. With a large number of subjects (164 in total) a training session cannot be blamed. Therefore for a simple object recognition experiment, a simple training session seems to be enough.

The third example is a subjective test of remote ultrasonography conducted by project [6]. The task was to recognize an organ and decide if there were any problems with it. Since the quality of ultrasonography is strongly dependent on the person who is conducting the examination, the test had to be interactive. We explained why the provided quality was low and where the system was to be used (remote places with limited Internet access like Mali in Africa). Nevertheless, in a typical examination, additional information about a patient is available. In cases where there were problems conducting the examination, many other ideas about how to proceed in a real life situation were proposed by doctors. It made the experiment very chaotic. We also noticed that an examination cannot be too long and the tasks (what should be investigated) cannot be very similar, or a doctor will likely lose interest in the experiment. The training session has to include a very clear and detailed explanation of the experiment. The best would be to consult with a doctor to frame the explanation. Also, we should be prepared to give additional information as to why all scenarios had to be conducted to obtain the results needed by the project or algorithm development. Motivation is one of the most crucial parts.

Each task recognition experiment is different. Even if a task description looks easy, we advise that a small preliminary test be run. Not only the results, but also interviews with the subjects taking part in a preliminary test, help to prepare an experiment description and training set that not only explains what to do but also motivates the subjects to perform the task correctly.

References

[1] Mikołaj Leszczuk, Lucjan Janowski, Piotr Romaniak, Andrzej Głowacz, and Ryszard Mirek. Quality assessment for a licence plate recognition task based on

a video streamed in limited networking conditions. In *Multimedia Communications, Services and Security*, pages 10–18. Springer Berlin Heidelberg, 2011.

[2] Task-based subject validation: reliability metrics, L Janowski Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop ...

[3] Mikołaj I Leszczuk, Artur Kon, Joel Dumke, and Lucjan Janowski. Redefining ITU-T P. 912 recommendation requirements for subjects of quality assessments in recognition tasks. In *Multimedia Communications, Services and Security*, pages 188–199. Springer Berlin Heidelberg, 2012.

[4] VQiPS: Video quality tests for object recognition applications. Public Safety Communications DHS-TR-PSC-10-09, U.S. Department of Homeland Security's Office for Interoperability and Compatibility (June 2010)

[5] VQiPS: Recorded-video quality tests for object recognition tasks. Public Safety Communications DHS-TR-PSC-11-01, U.S. Department of Homeland Security's Office for Interoperability and Compatibility (June 2011)

[6] http://www.qol.unige.ch/research/TeleUSG.html

Lucjan Janowski is an assistant professor at the Department of Telecommunications (AGH University of Science and Technology). He worked on malicious traffic analysis in CNRS-LAAS in France at a post-doc position in 2007. In 2010/2011 he spent half a year on a post-doc position in University of Geneva working on QoE for health applications. His main interests are statistics and probabilistic modelling of subjects and subjective rates used in QoE evaluation.

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 as a steering committee member or researcher in several national and European projects, including: INDECT, BRONCHOVID, GAMA, e-Health ERA, PRO-ACCESS, Krakow Centre of Telemedicine, CONTENT, E-NEXT, OASIS Archive, and BTI. 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.

To Train or Not To Train?

Nicolas Staelens, Wendy Van den Broeck and Filip De Turck

The way standards do

It is generally known that subjective (audio)visual quality assessment experiments need to be conducted in stringent controlled environments, as detailed in ITU-T Rec. P.910 and ITU-R Rec. BT.500. This facilitates experiment repeatability, enables comparing results obtained from different experiments conducted at different locations, and minimizes the influence of contextual factors during quality evaluation.

As part of assessing (audio)visual quality by means of subjective experiments, specific instructions are provided on how to evaluate and rate the different video sequences. Also, a training session is used to familiarize the observers with the experiment and the type (and range) of impairments they can expect. As such, observers 'know' what to look for and what to expect. However, what about the influence of context and user expectations on quality perception?

Several subjective testing methodologies have already long been standardized. Notwithstanding their specific application domains (e.g. video, speech, conferencing, recognition), they all share common ground and require, amongst other things, that test subjects be properly informed about the experiment and the task at hand.

Prior to the start of the experiment, detailed instructions are provided to the observers explaining the intended application under test, the overall trial structure, and the quality rating mechanism. Furthermore, a training phase is incorporated in preliminary trials in order to illustrate the type and range of quality impairments that can occur during the experiment. Consequently, observers are primarily focused on active (audio)visual quality evaluation.

These methodologies are widely used in video quality research and ongoing VQEG projects to measure the video's *technical* quality as perceived by the test subjects.

What about Quality of Experience?

But what about measuring Quality of Experience (QoE), a buzzword associated with terms like *delight*, *user expectations*, *enjoyment*, *personality*, *service*, *content*, and *context of use*? To what extent are user expectations and context of use taken into account during standardized subjective quality assessment? Can the existing subjective quality assessment methodologies be used to measure QoE?

> **"Quality of Experience** is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state."
>
> Le Callet P., et al (2012), "Qualinet White Paper on Definitions of Quality of Experience," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.2, March 2013.

The mandatory training phase prepares subjects for the experiment and informs them about what to expect. By informing them, their focus is aimed directly at evaluating the video quality as such. So, when we want to assess QoE, should we at all incorporate a training phase as part of our experiments? Or should we try to mimic realistic viewing behavior as much as possible?

Contextualized subjective experiments

With respect to our QoE research, we have conducted several contextualized subjective quality assessment experiments by integrating the everyday life context (Staelens et al., 2012; Van den Broeck et al., 2012). These experiments were conducted in complement to controlled lab tests in order to enable results comparison and study the influence of more ecologically valid

testing environments on quality perception. These studies have highlighted the importance of immersion and primary focus during subjective video quality assessment.

In one of our studies, users were asked to watch a full length DVD movie in their most natural environment, i.e. at home on their own device (Staelens et al., 2010). Users were not informed about the possible presence of visual artifacts during playback. Hence, primary focus shifted to watching the actual content of the movie. Compared to controlled lab experiments, impairments were less noticed (see Figure 1). However, despite the fact that blockiness is more easily detected compared to frame freezes, subjects indicated that freezes are more disturbing during DVD playback. Freezes tend to break the natural flow of the movie and users feel their *immersive experience* is hampered. It is important to note that, due to the restrictions imposed by the ITU recommendations, the feeling of immersion cannot be (re)produced during controlled lab experiments, also because the duration of the video sequences is limited.

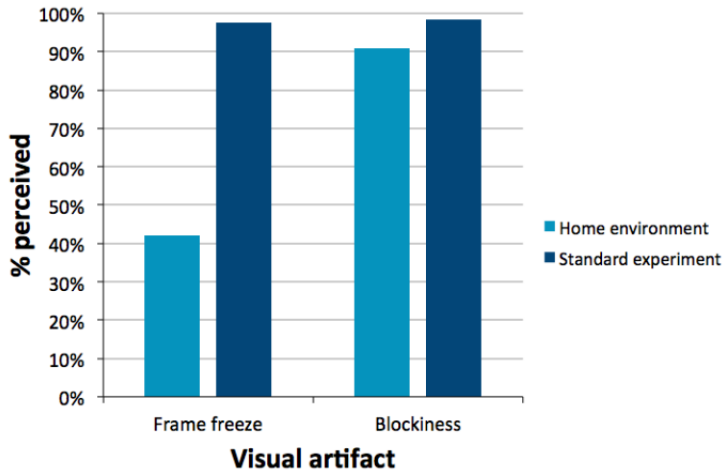


Figure 1. Influence of primary focus on impairment visibility.

Controlled lab or real-life?

Matulin and Mrvelj (2013) also state that the most accurate QoE evaluations include real-life experiments in the typical environments where the services are used, without subjects actively being focused on (audio)visual quality assessment. Based on a comprehensive summary of QoE experiments conducted in real-life environments, the authors conclude that there are substantial differences between the results obtained in controlled labs and real-life environments. In general, users

Nicolas Staelens obtained his PhD degree in Computer Science Engineering in 2013 from Ghent University, Belgium. His research activities focus on assessing the influence of network impairments on perceived audiovisual quality. As part of his research, he has gained experience in conducting subjective experiments in real-life environments.

Wendy Van den Broeck is a senior researcher at iMinds-SMIT, Brussels, Belgium. She is active in the user Empowerment research unit and conducted research in different projects concerning the domestication of new media technologies in a home context. Her recent projects are related to multi-screen user practices, QoE research on video streaming, and research related to iDTV, 3D-TV, second screen applications and HDTV.

Filip De Turck leads the network and service management research group at the Department of Information Technology of the Ghent University, Belgium and iMinds (Interdisciplinary Research Institute in Flanders). His main research interests include scalable software architectures for network and service management, design and performance evaluation of novel QoE-aware multimedia delivery systems.

are more forgiving of quality degradations in real-life environments.

Thus, conducting experiments in real-life environments without really informing test subjects might yield more representative results in the case of investigating end-users' QoE.

In this respect, implementing methodologies like the Experience Sampling Method (ESM) (Hektner et al., 2007) might be the way to go in order to get a better understanding of QoE in real-life.

So …

… "*To train or not to train (test subjects), that is the question.*" And for sure, the answer will depend on what we really want to assess.

References

Hektner, J.M., Schmidt J.A., Csikszentmihalyi M. (2007),"Experience Sampling Method: Measuring the Quality of Everyday Life", California: Sage Publications Inc.

Matulin M., Mrvelj Š. (2013), "State-of-the-Practice in Evaluation of Quality of Experience in Real-Life Environments", Promet – Traffic & Transportation, Vol. 25, No. 3, pp. 255-263.

Staelens N., Moens S., Van den Broeck W., Mariën I., Vermeulen B., Lambert P., et al. (2010), "Assessing Quality of Experience of IPTV and Video on Demand Services in Real-life Environments", IEEE Transactions on Broadcasting, Vol. 56, Issue 4, pp. 458-466.

Staelens N., Van den Broeck W., Pitrey Y., Vermeulen B., Demeester P. (2012), "Lessons Learned during Real-life QoE Assessment", in EuroITV – 10th European Conference on Interative TV.

Van den Broeck W., Jacobs A., Staelens N. (2012), "Integrating the Everyday-life Context in Subjective Video Quality Experiments", Fourth International Workshop on Quality of Multimedia Experience (QoMEX).

On viewing distance and visual quality assessment in the age of Ultra High Definition TV

Patrick Le Callet, Marcus Barkowsky

Viewing distance and Quality assessment

The consumer video market is largely driven by the introduction of new formats (e.g., new pixel resolution). Each time, the story remains the same: what is the optimal viewing distance? Ultra High Definition TV is not an exception. This simple question is of crucial importance when it comes to the issue of quality and the added value of a new technology. In this letter, we revisit the topic, starting from best practices and then raising open questions.

Ultra High Definition (UHD) TV is following the tradition of enhancing Quality of Experience in consumer video. It notably offers the prospect of attaining a large field of view while fulfilling the limits of the Human Visual System (HVS) in terms of spatial and temporal contrast sensitivity. This should lead to a higher level of immersion which may reduce the influence of disturbing context influence factors by decoupling the observer from his environment. In order to ensure the adoption of the new technology by consumers, it is necessary to identify the conditions and limits under which the Quality of Experience is sufficiently increased. In this context, subjective experiments are useful to learn about the influence factors and provide meaningful guidelines. Visual distance, due to its close relationship with viewing field and immersion, is a key influence factor. In particular, as quality judgment might differ from one observer to another, well-defined experimental conditions are preferable, allowing for reproducibility from one individual to another or from one test environment or test lab to another. The viewing distance must be controlled and set under ad hoc rules.

Viewing distance and ITU recommendations: a (his)story of resolution

The ITU (International Telecommunication Union) has produced over the last decades numerous recommendations for the different parameters and conditions needed to conduct subjective quality assessment experiments. Usual controlled factors are the viewing distance, general ambiance (lighting, color of the walls...) and the display screen. Traditionally, the room setup and the display are chosen such that the detection of artifacts is as easy as possible for the observer.

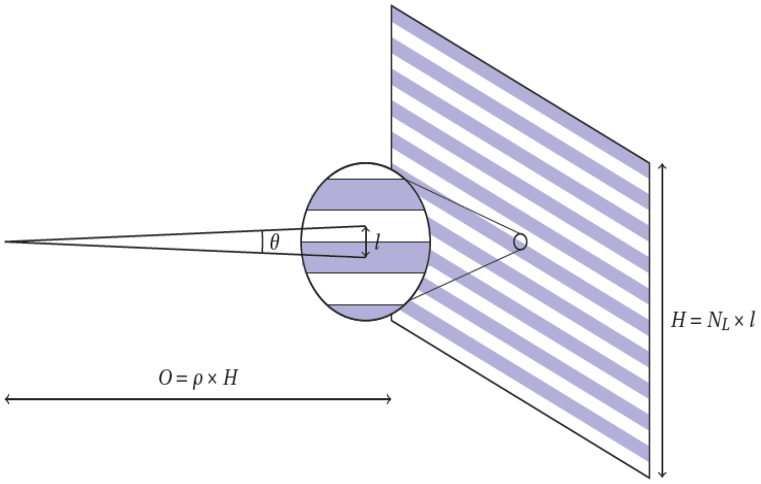


Figure 1. Viewing distance O and its related physical parameters

Historically, the ad hoc viewing distance depends on the number of lines of the image. To take maximum advantage of the resolution, the optimal position for an observer should correspond to the limit of visual discrimination between two lines. Discrimination power of a regular (normal vision) observer is on average one minute of arc, which corresponds to a critical pattern frequency of 30 cycles per degree (cpd). The angle between two lines as represented in Figure 1, can be computed using the equation:

$$\theta = 2.\arctan\left(\frac{1}{2.\rho.N_L}\right) \quad (1)$$

with N_L being the number of lines and ρ the ratio between the viewing distance and the physical height of the active screen

area. Consequently, in the case of Standard Definition TV with 576 lines, one should be at a distance corresponding to:[2]

$$\rho = \frac{1}{2x576x\tan\left(\frac{1'}{2}\right)} = 5.98 \quad (2)$$

which is around 6 times the image height. For 1080 line HDTV, this value is reduced to around three times the image height. This distance has a direct impact on the extent of the visual field that is covered by the image as reported in Table 1. The horizontal viewing angle α can be obtained as:

$$\alpha = 2.\arctan\left(\frac{Np}{N_L}\frac{1}{2.\rho}\right) \quad (3)$$

with N_P the number of pixels on a line.

Table 1. Relative viewing distance and corresponding horizontal viewing field for different resolutions.

Resolution	Relative viewing distance (to the image height)	Horizontal Viewing Field (in degree)
SDTV (576 lines) [3]	5.98	11,93
HDTV (1080p) [4]	3.18	31.27
UHDTV (2160 lines) [5]	1.59	52.87

The critical frequency of 30 cpd can be considered as a lower bound for a usual observer. This value tends to increase depending on the contrast of the pattern, its speed, and the surrounding conditions (60 cpd can be considered as a higher bound).

[2] In (2) the unit of the input of the tan function is in minutes of arc.
[3] Aspect ratio (number of pixels per line/number of lines) is 1.25:1.
[4] Aspect ratio is 1.78:1.
[5] Aspect ratio is 1.78:1.

Figure 2 shows the relationship between the diagonal of the display, measured in inches, and the viewing distance in

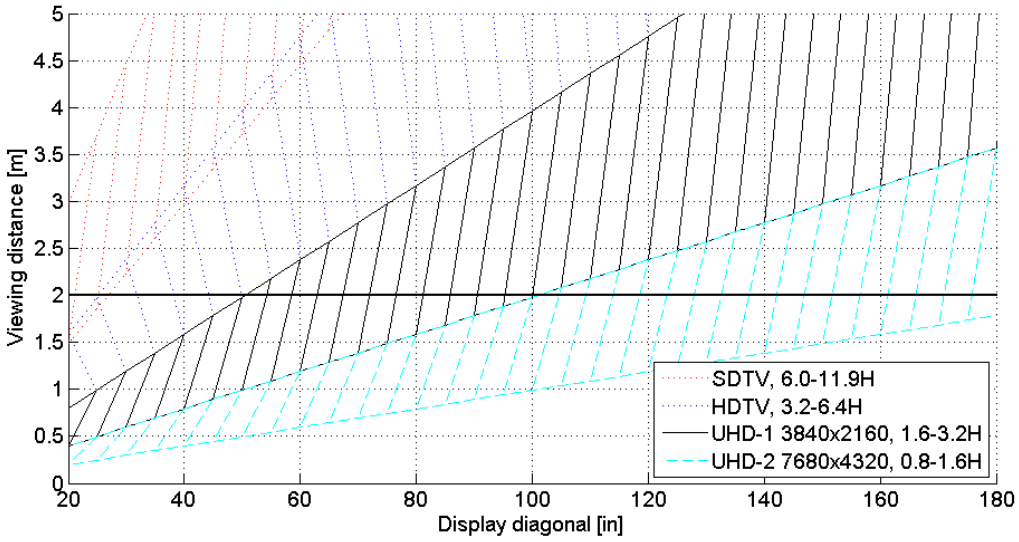


Figure 1. The relationship between absolute viewing distance in meters and the display diagonal in inch for the three resolutions HDTV, UHD1, and UHD2 when considering a range of resolution of the human eye of 30cpd to 60cpd. In home viewing, a typical absolute viewing distance may be considered as 2m. In case of line interleaved 3D displaying, the vertical resolution is halved, thus the next lower resolution applies.

meters for the four resolutions SDTV, HDTV, UHD1, and UHD2. The upper limit of the area provides the highest spatial contrast sensitivity that the HVS may support (60 cpd), notably when objects with a high-contrast texture at the critical frequency are moving at an average speed of about 0.15 degrees per second.[6] The lower bound of the area is calculated for 30 cpd, a retinal frequency that still avoids seeing the pixel grid in most cases. It has been previously used, for example in the case of HDTV[7] [3]. The diagram shows that for a typical viewing distance of 2 m in a living room, the size of the display needs to be significantly enlarged, i.e. up to 100 in (2.54 m) for UHD-1.

[6] Daly, S. Engineering Observations from Spatiovelocity and Spatiotemporal Visual Models. In IS&T/SPIE Conference on Human Vision and Electronic Imaging III., SPIE Vol. 3299, pp. 180-191, January 1998.

[7] Cermak, G., Thorpe, L., & Pinson, M. (2009). Test Plan for Evaluation of Video Quality Modelsfor Use with High Definition TV Content. Video Quality Experts Group (VQEG)

Viewing distance and UHD TV: revisiting the history?

Higher resolution offers a reduction in viewing distance and an increase in viewing angle, implying better immersion and better Quality of Experience. To what extent is the last part of this statement valid?

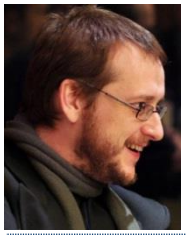
When targeting higher resolution and consequently lower viewing distance and larger excited visual field, factors other than discrimination between lines might come into play and affect the comfort of the observer, especially when the perceived quality of the media is not sufficient. It has been observed[8] when comparing standard definition and high definition conditions that larger viewing field has a positive effect at high quality while it exhibits clearly negative effects at mid quality levels (standard definition is then preferred compared to high definition). More generally, the focus may shift from pure video quality evaluation to Quality of Experience (QoE),[9] which can lead to the concept of preferred viewing distance.

For instance, it should be noted that for smaller display sizes, observers prefer larger viewing distances. This is partly due to the accommodation effort that is required when the viewing distance is inferior to 1 m, a distance that may even imply focusing difficulties for senior viewers. It has also been shown recently[10] that illumination conditions may influence the

[8] S. Péchard, M. Carnec, D. Barba, et others, « From SD to HD television: effects of H. 264 distortions versus display size on quality of experience IEEE International Conference on Image Processing, ICIP, 2006, p. 409–412.

[9] a term which aims at evaluating the overall satisfaction of the user. It has been recently defined as "…the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state". Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds , "Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003),., Lausanne, Switzerland, Version 1.2, March 2013

[10] Lee, D. - S., & Shen, I. - H. (2012). Effects of illumination conditions on preferred viewing distance of portable liquid-crystal television. Journal of the Society for Information Display, 20(7), 360–366.



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. He is currently serving as associate editor for IEEE transactions on Circuit System and Video Technology, SPRINGER EURASIP Journal on Image and Video Processing, and SPIE Electronic Imaging.

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

preferred viewing distance as well, which may be explained by the fact that the contrast sensitivity increases with higher illumination.

Moreover, while a higher level of immersion or presence is usually perceived as advantageous, it may also introduce discomfort issues. Because of the larger field of view, simulator sickness may occur due to the decoupling of the visual stimulus with the sense of balance. This is particularly true for fast camera movements.

As UHD content is currently not very widespread, and the habits of consumers nowadays include watching online available content that is often only available at lower resolutions and reduced quality, the optimal viewing distance may vary with the usage condition in the home environment, i.e., smaller viewing distance when watching high quality UHD content and larger viewing distance when watching low quality web content. In some conditions, it may also prove advantageous to reduce the active screen size in order to avoid visual discomfort issues such as simulator sickness. While one could stick to the original ITU methods, optimal guidelines on viewing distance might need to be developed both for lab experiments as well as for the home environment, in particular for large UHD displays.

Comparison of Metrics: Discrimination Power of Pearson's Linear Correlation, RMSE and Outlier Ratio

Greg Cermak

Editor's note: This article by former ILG Co-Chair Greg W. Cermak appeared in the VQEG reflector in June, 2008 under the title "Comparison of Metrics: VQEG Multimedia Data." The article is reprinted with permission from Greg W. Cermak. Bracketed text and footnotes indicate clarifications by the editor.

The graphs and tables below show three things:

- [Pearson's linear] correlation, RMSE,[11] and outlier ratio all measure essentially the same thing.
- RMSE is better at discriminating between models.
- The advantage of RMSE over correlation increases as the number of video samples decreases, and vice versa.

These conclusions were also true in FRTV2.[12]

RMSE is better than Pearson's linear correlation and outlier ratio at discriminating between objective video quality models.

This note is organized into three parts. Part 1 shows the interrelationship of the metrics correlation (Pearson's R), RMSE, and the outlier ratio. Part 2 shows the performance of the metrics correlation, RMSE, and outlier ratio for the VQEG MM[13] data set. Part 3 shows the actual performance of R and RMSE for the FRTV2 and MM data sets and for hypothetical data from an experiment with 20 PVSs.[14]

[11] Root mean square error, Ed.
[12] VQEG's full reference television validation test, phase II, Ed.
[13] VQEG's multimedia validation test, phase I, Ed.
[14] Processed video sequence, Ed.

Part 1. Explanation

Each of the plots below is based on the FR metrics[15] for the 13-14 tests across 5 proponents; therefore 65-70 data points per plot. The metrics are highly correlated with each other.

Below the plots, for each resolution, is the output of a Principal Components factor analysis on the same data. The highlighted number labeled "proportion" is the proportion of variance in the 3 metrics across the 13-14 tests and 5 models that is accounted for by a single factor. That proportion of variance (an R2 measure) is always around 0.9, and the proportion accounted by any other factor is tiny. That is, each of the metrics is measuring essentially a single underlying factor, although in slightly different ways.

Following the graphs and factor analyses (Part 1) are the results of doing significance tests comparing each model to the best-performing model according to each type of metric, for each resolution (Part 2). These results are presented as tables of 1's and 0's. A '1' means that a model is tied with the top-performing model in the sense that it is not statistically significantly different. The more 1's in a table, the more ties. The more ties, the poorer the discrimination of the metric. Counting up the 1's, RMSE outperforms both correlation and outlier ratio in discriminating between models.

This analysis was critical in VQEG's decision to use RMSE to measure significant differences between objective video quality models in the HDTV validation test.

Correlation and outlier ratio have their advantages. Correlation is good for a simple summary of results. Outlier ratio is good for diagnosing model performance in order to improve the model's performance. When it comes to distinguishing between models, RMSE does the best job.

[15] Full reference metrics, Ed.

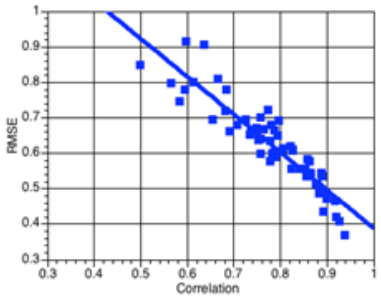


Figure 1. VGA data, RMSE plotted against Pearson's R. Linear fit R2 = 0.854, R = 0.92

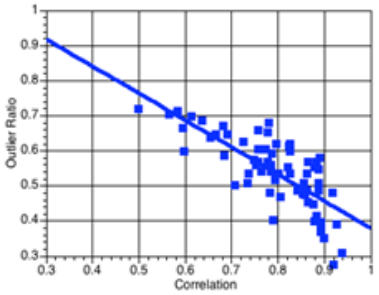


Figure 2. VGA data, Outlier Ratio plotted against Pearson's R. Linear fit R2 = 0.577, R = 0.760

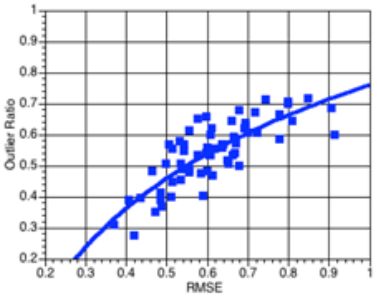


Figure 3. VGA data, Outlier Ratio plotted against RMSE. Log fit R2 = 0.653, R = 0.808

Principal Components analysis of the three metrics for VGA 1
10:04 Monday, May 5, 2008

The PRINCOMP Procedure

Observations	65
Variables	3

Simple Statistics

	corr	rmse	outratio
Mean	0.7901538462	0.6132461538	0.5412153846
StD	0.1000312812	0.1156449349	0.1012611865

Correlation Matrix

	corr	rmse	outratio
corr	1.0000	-.9240	-.7595
rmse	-.9240	1.0000	0.7858
outratio	-.7595	0.7858	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.64838753	2.37151340	0.8828	0.8828
2	0.27687412	0.20213577	0.0923	0.9751
3	0.07473835		0.0249	1.0000

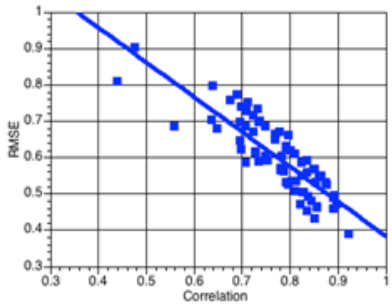


Figure 4. CIF data, RMSE plotted against Pearson's R. Linear fit R2 = 0.721, R = 0.849

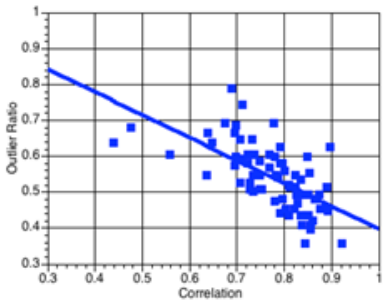


Figure 5. CIF data, Outlier Ratio plotted against Pearson's R. Linear fit R2 = 0.405, R = 0.636

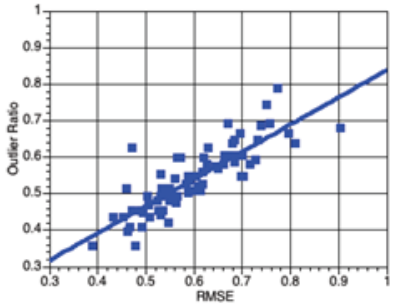


Figure 6. CIF data, Outlier Ratio plotted against RMSE. Linear fit R2 = 0.707, R = 0.841

Principal Components analysis of the three metrics for CIF 1
10:26 Monday, May 5, 2008

The PRINCOMP Procedure

Observations	70
Variables	3

Simple Statistics

	corr	rmse	outratio
Mean	0.7722857143	0.6013857143	0.5422428571
StD	0.0907168142	0.1022225893	0.0907040794

Correlation Matrix

	corr	rmse	outratio
corr	1.0000	-.8494	-.6362
rmse	-.8494	1.0000	0.8382
outratio	-.6362	0.8382	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.55313907	2.18923473	0.8510	0.8510
2	0.36390434	0.28094776	0.1213	0.9723
3	0.08295658		0.0277	1.0000

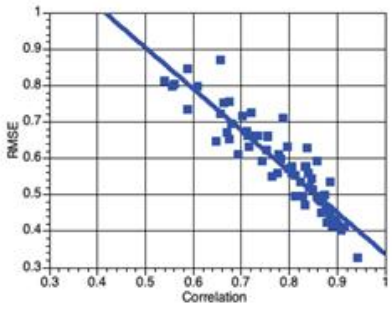


Figure 7. QCIF data, RMSE plotted against Pearson's R. Linear fit R2 = 0.853, R = 0.924

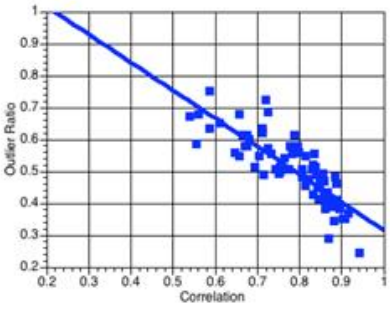


Figure 8. QCIF data, Outlier Ratio plotted against Pearson's R. Linear fit R2 = 0.702, R = 0.838

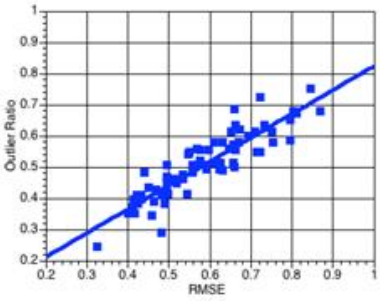


Figure 9. QCIF data, Outlier ratio plotted against RMSE. Linear fit R2 = 0.805, R = 0.897

Principal Components analysis of the three metrics for QCIF 1
10:48 Monday, May 5, 2008

The PRINCOMP Procedure

Observations	70
Variables	3

Simple Statistics

	corr	rmse	outratio
Mean	0.7816857143	0.5848142857	0.5069142857
StD	0.0997379963	0.1224948733	0.1043994201

Correlation Matrix

	corr	rmse	outratio
corr	1.0000	-.9235	-.8378
rmse	-.9235	1.0000	0.8972
outratio	-.8378	0.8972	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.77288905	2.60847722	0.9243	0.9243
2	0.16441183	0.10171272	0.0548	0.9791
3	0.06269911		0.0209	1.0000

Part 2. Performance of the metrics Correlation, RMSE, and Outlier Ratio for the VQEG MM data set[16]

Editor's note: Rows contain objective video quality models. Columns contain subjective video quality datasets (e.g., V01, V02). The table title indicates the metric used to calculate statistical equivalence: Pearson linear correlation, RMSE, or outlier ratio.

Table 1. VGA data, correlation metric. FR Models

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	Total
Psy_FR	1	0	1	1	0	1	1	1	1	1	1	1	1	**11**
Opt_FR	0	1	1	1	1	1	1	0	1	0	1	1	1	**10**
Yon_FR	1	0	0	1	1	1	1	1	1	1	1	0	1	**10**
NTT_FR	1	0	1	1	1	1	0	1	0	0	0	1	1	**8**
PSNR DMOS	1	0	1	1	0	0	0	0	0	0	0	0	0	**3**

Table 2. VGA data, RMSE metric. FR Models

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	Total
Psy_FR	1	0	1	1	0	1	0	1	1	1	1	1	1	**10**
Opt_FR	0	1	1	1	1	1	1	0	1	0	1	0	0	**8**
Yon_FR	0	0	0	1	1	1	1	1	0	1	0	0	0	**6**
NTT_FR	1	0	1	1	0	0	0	1	0	0	0	0	0	**4**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

[16] To assist in the readability of the tables in this reprint, (1) label "Total=" was replaced with "Total", (2) label "PSNR_DMOS" was replaced with "PSNR DMOS" and (3) the tables were transposed. As a consequence of the transposition, the label in the upper-left box ("Test") became incorrect and was omitted. See section 9 of the Multimedia Phase I ILG Data Analysis for these tables in their original format, Ed.

Table 3. VGA data, outlier ratio metric. FR Models

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	Total
Psy_FR	1	0	1	1	1	1	1	1	1	1	1	1	1	**12**
Opt_FR	1	1	1	0	1	1	1	0	1	1	1	1	1	**11**
Yon_FR	1	0	0	1	0	1	1	1	1	1	0	0	1	**8**
NTT_FR	1	0	1	1	1	0	1	1	1	0	0	1	1	**9**
PSNR DMOS	1	0	1	1	0	0	0	0	1	0	0	0	0	**4**

Table 4. CIF data, correlation metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**14**
Opt_FR	1	1	0	1	1	1	1	1	1	1	1	1	1	1	**13**
Yon_FR	1	0	0	1	1	1	1	1	1	1	1	1	0	0	**10**
NTT_FR	0	1	1	1	1	1	1	0	0	1	0	1	0	0	**8**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Table 5. CIF data, RMSE metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	1	1	1	1	1	1	0	**13**
Opt_FR	1	1	0	1	0	1	1	1	1	0	1	1	0	1	**10**
Yon_FR	1	0	0	0	1	1	1	1	1	1	1	1	0	0	**9**
NTT_FR	0	1	1	1	0	1	1	0	0	0	0	1	0	0	**6**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Table 6. CIF data, outlier ratio metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	0	1	1	1	1	1	0	**12**
Opt_FR	1	1	0	1	1	1	1	1	1	1	1	1	1	1	**13**
Yon_FR	1	0	1	1	1	1	1	1	1	1	1	1	0	0	**11**
NTT_FR	1	1	1	1	0	1	1	0	1	1	0	1	1	0	**10**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**1**

Table 7. QCIF data, correlation metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	1	1	0	1	1	1	0	**12**
Opt_FR	0	0	1	1	1	0	1	1	1	1	1	1	1	1	**11**
Yon_FR	1	0	0	0	1	0	0	1	0	0	1	0	0	0	**4**
NTT_FR	1	1	1	1	1	0	0	1	0	0	1	0	1	1	**9**
PSNR DMOS	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**1**

Table 8. QCIF data, RMSE metric. FR Models

	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Total
Psy_FR	1	1	1	0	1	1	1	1	1	0	1	1	1	0	**11**
Opt_FR	0	0	1	1	1	0	0	1	1	1	1	1	1	1	**10**
Yon_FR	0	0	0	0	0	0	0	1	0	0	1	0	0	0	**2**
NTT_FR	1	0	1	1	0	0	0	1	0	0	1	0	1	1	**7**
PSNR DMOS	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**1**

Table 9. QCIF data, outlier ratio metric. FR Models

	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Total
Psy_FR	1	1	0	1	1	1	1	1	1	1	1	1	1	0	**12**
Opt_FR	0	1	1	1	1	0	1	1	1	1	0	1	1	1	**11**
Yon_FR	1	1	0	1	1	1	1	1	0	0	1	0	0	0	**8**
NTT_FR	1	1	1	1	1	0	0	1	0	1	1	0	1	1	**10**
PSNR DMOS	0	0	1	0	0	0	0	0	0	1	1	0	0	1	**4**

Part 3. Comparison of R and RMSE for data sets of different sizes

(Thanks to Steve Wolf for a close reading and suggestions about this section.)

Consider the data from FRTV2. In FRTV2 there were two experiments, one for 525-line video and one for 625-line video. From FRTV2 we have, for six models that were in any kind of contention, correlation (Pearson's R) and RMSE scaled to a 5-point scale:

Table 10. FRTV2 data for 525 and 625 experiments, correlation and RMSE metrics.

R, 525 data	RMSE, 525 data	R, 625 data	RMSE, 625 data
0.937	0.37	0.898	0.395
0.935	0.375	0.886	0.415
0.856	0.55	0.884	0.42
0.836	0.585	0.87	0.445
0.756	0.695	0.779	0.565
0.682	0.775	0.703	0.64

In FRTV2, there were 64 PVSs in each experiment. In the MM experiments there were in excess of 150. For purposes of the following analyses, we consider experiments with 150, 64, and (hypothetically) 20 PVSs.

First, the critical difference in R required to declare two models different is given in sections 8.4.1 and 8.5.1 of the MM Final Report draft 1.4.1. The R to Z transform is applied, then the critical difference in Z-scores is computed; this critical Z depends on R and N, the number of PVSs in the test. Using a handy spreadsheet designed by Jamie DeCoster & Anne-Marie Leistico, we can determine that if R = 0.85 and N=150, then the critical R difference = 0.08. In Table 11 below, we also compute the critical R difference for R = 0.85 and N = 64 (the number of PVSs in FRTV2) and for N = 20.

The corresponding critical "RMSE difference" is actually a ratio of mean squared errors (MSEs) for any two models being compared. Given N, the critical F ratio is available from published tables or can be calculated in spreadsheets. We use critical F at the 95% confidence level for N = 150, 64, and 20.

Next we determine the corresponding RMSE's. We have empirical relationships between RMSE and R in Table 10 and in Figures 1, 4, and 7 above. Since there is not a single, unique relationship in our empirical data, we do computations for three different RMSE-R relationships given below (the one for VGA is very similar to the ones for CIF and QCIF):

- VGA: RMSE = -1.07*R + 1.46
- FRTV2 525: RMSE = -1.62*R + 1.91
- FRTV2 625: RMSE = -1.26*R + 1.53

Using these empirical relationships, and assuming that the target range of R's that are of interest is around 0.85, plus or minus some, we go through the following steps. These steps are based on being able to calculate critical R differences based on the Z-transform and known relations between N and Z; transforming from R to RMSE given the empirical relations

above; calculating critical RMSE's from F-tables (based on corresponding MSEs and N); and transforming back to the familiar R scale using the empirical relations above. We then can compare critical differences in the data required for significance using R and RMSE. The steps:

1. For a given N (column 1 in Table 11), calculate the critical R difference (column 2 in Table 11). I used the spreadsheet by DeCoster & Leistico; in this example it is 0.08.
2. Using one of the empirical relationships above, find the corresponding RMSE. In the case of both VGA and FRTV2 525 it turned out to be 0.550.
3. Square the RMSE to find MSE.
4. For the given N, find the critical F value (for 95% confidence). For N = 150, that turns out to be 1.31 (column 3 in Table 11).
5. Find the critical MSE for the second model; in this example it is ((0.550)**2)*1.31 = 0.396.
6. Convert back to RMSE by taking the square root; in this example, it is 0.630 (column 4 in Table 11).
7. Find the corresponding R value using the empirical relationship above (for the VGA example given, this would be (0.630 – 1.46) / (-1.07) = 0.776.
8. Take the difference between the starting R (0.85) and the critical R (0.776); in the VGA example it is approximately 0.07 (column 6 in Table 11).

Following these steps, we get Table 11.

Table 11. Differences in Pearson's R required for statistical significance for three values of N, the number of PVSs, and corresponding RMSE differences (scaled in terms of R).

1	2	3	4	5	6	7	8
N	**R diff**	**F (.95)**	**Critical RMSE for VGA and FRTV2 525**	**Critical RMSE for FRTV2 625**	**Estimated R diff for VGA**	**Estimated R diff for FRTV2 525**	**Estimated R diff for FRTV2 625**
150	0.08	1.31	0.63	0.525	0.07	0.06	0.05
64	0.14	1.51	0.676	0.564	0.12	0.09	0.08
20	0.32	2.12	0.801	0.669	0.23	0.17	0.17

Table 11 shows that as N gets smaller, the critical R difference (column 2) and the corresponding critical RMSE (columns 4 and 5) both get larger, as we expect. However, the difference in sensitivity between R and RMSE also gets larger as the sample size decreases (compare column 2 with columns 6, 7, 8). Or, the other way around, if N gets very large, then the sensitivity of R and of RMSE probably converge. Also, we first noticed the advantage of RMSE in FRTV2 where the N was smaller than the recent MM project, and the consequent advantage in sensitivity for RMSE was more obvious.

Theoretical

Clearly, RMSE and R both depend on N and on the empirical distribution of discrepancies between model predictions and the observed MOS or DMOS scores[17] (called Perror in the MM Final Report). Presumably, one could write out the relationships between R and N and Perror, and between RMSE and N and Perror. Then it might be obvious when the critical R difference and the critical RMSE difference should differ from each other. I have not tried this exercise yet. Also, the empirical relationships between RMSE and R given above

[17] Mean opinion score (MOS) and differential mean opinion score (DMOS), Ed.

are certainly just estimates of some theoretically "true" relationship. Steve Wolf and I have made different guesses about what this relationship might be, but we are not quite ready to say what those guesses are.

Reference

"Draft final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I," Version 1.4.1 April 15, 2008. ©2008 VQEG.

Greg W. Cermak performed the ILG's official data analysis for the VQEG Full Reference Phase I and II validation tests, the VQEG Multimedia validation test, and the VQEG RRNR-TV validation test. Greg was a Co-Chair of the VQEG Independent Lab Group (ILG) until he retired from Verizon in 2010.

Progress of the Monitoring of Audio Visual Quality by Key Indicators (MOAVI) Project

Mikołaj Leszczuk, Silvio Borer, and Emmanuel Wyckens

The VQEG MOAVI project is an open collaborative for developing No-Reference models for monitoring audio-visual service quality. The goal is to develop a set of key indicators that describe service quality in general and to select subsets for each potential application. MOAVI models predict the presence or absence of these key indicators, not the overall quality.

The MOAVI project has accomplished the following tasks from inception through 2013.

- Implementation of 7 metrics for the following artifacts:
  - Blockiness – the probability of correct classification: 98.48%
  - Blur – the probability of correct classification: 80.52%
  - Exposure time distortion
  - Noise
  - Framing
  - Freeze
  - Blackout
- Initial values of thresholds for particular metrics were settled
- Development of metrics for audio artifacts (mute and clipping) in a MATLAB® environment
- Development of metrics for block loss and interlace artifacts in a MATLAB environment
- Preliminary tests of subjective opinion with the purpose of improving the approach to thresholds
- Design and construction of the website where the metrics are publicly available (vq.kt.agh.edu.pl)
- Writing a paper about the MOAVI project for the SIGCOMM conference in Hong-Kong and VPQM conference in Arizona
- SIGCOMM and VPQM conferences reviewers have provided some feedback comments that should be analyzed and taken into account for future steps of the MOAVI project. The most important weakness detected is the lack of any presentation of

actual results in the articles, although there is a set of metrics of artifacts ready.

- Therefore, a set of video and audio files has been created to test the metrics developed in previous months (Mute, Clipping and the Voice Activity Detector). These results of the metrics on those videos are ready to be compared with some ground truth determined by the researchers or eventually the results of subjective tests.
- In the case of the Voice Activity Detector in particular, its accuracy in detecting the voice activity in the audio clips extracted from the database has been measured by comparing the results obtained from the detector with the ground truth determined by both the observation of the waveforms and listening to the sound.
- The metric to detect Lip Activity in the videos has been enhanced during this month and the results of the temporal activity in the region of the mouth for the videos in the database have been stored for future analysis. The main goal of the latter is the establishment of a threshold for considering the video frame as "lip active" or not.
- A set of test videos has been created with the following characteristics:
  - Frontal view of talking faces.
  - Duration around 20 s.
- Real delay introduced to make the tests compared with the delay detected by the metric:
  - Average deviation = 130 ms.
  - The metric discriminates positive and negative delays.
- For the supercomputing cluster calculations we had to move the Temporal Activity and Spatial Activity metrics to C++, which we think may also contribute to the small progress in the MOAVI project.
- Also just creating all the databases with the results of the MOAVI project metrics required the use of the project applications, which can be considered as a solid test (for a total of more than 7500 videos).

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

Setting metric threshold values

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

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

$$accuracy(th_{TEMP}) = \frac{number\ of\ matching\ results}{number\ of\ results} \quad (1)$$

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

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

***Mikołaj Leszczuk**, PhD, is an assistant professor at the Department of Telecommunications, AGH University of Science and Technology (AGH-UST), Krakow, Poland). His current activities are focused on e-Health, multimedia for security purposes, P2P, image/video processing (for general purposes as well as for medicine), the development of digital video libraries, particularly video summarization, indexing, compression and streaming subsystems.*

***Silvio Borer** received his Master degree in mathematics from the University of Zurich and the Ph.D. degree in science from EPF Lausanne, Switzerland, in 2004. He is currently with SwissQual AG, where he is a Senior Research Engineer. His research interests include video and audio quality estimation algorithms.*

***Emmanuel Wyckens,** research engineer for operational audiovisual services, graduated in Electronic Engineering at Valenciennes University (France), specializing in the design of audiovisual broadcasting solutions.*

In 2000 he joined Orange labs, being involved in the improvement of MPEG-2/4 codec video quality and the optimization of audiovisual chain tools. Mr Wyckens' current activities are in the field of subjective, objective audiovisual quality in multimedia applications, and for high definition television domains. He participated in the work of EBU/ITU projects for standardizing subjective methodologies.

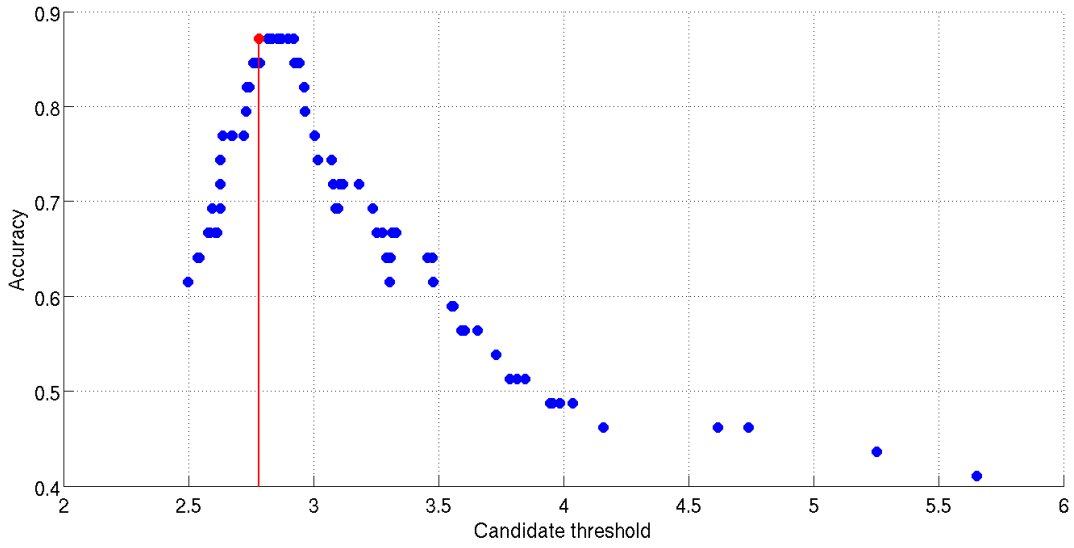


Figure 1. Blur metric threshold determination. Points represent the relation between candidate thresholds and accuracy. The line is drawn at the best candidate, which is chosen to be the metric threshold.

Key indicators verification

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

Table 1. Key Indicators verification – probability of distortion detection.

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

Information for Authors

VQEG wants the eLetter to be interactive in nature. Readers are encouraged to respond to articles on this eLetter's topic: practice sessions for subjective quality experiments. Response articles can be used to ask questions or provide feedback that could improve the other author's work. Response articles may give knowledge from a related field of study, identify alternate solutions, or provide evidence that a particular technique is unreliable. Please submit response articles to the eLetter editors, Margaret H. Pinson mpinson@its.bldrdoc.gov and Naeem Ramzan Naeem.Ramzan@uws.ac.uk.

A future VQEG eLetter will contain an anthology of articles on objective model validation. Interested authors should contact issue editor Kjell Brunnström Kjell.Brunnstrom@acreo.se. The article submission deadline is June 27, 2014. See the VQEG eLetter webpage for author instructions and the eLetter template.

Also of interest is the 6th International Workshop on Quality of Multimedia Experience (QoMEX) . Their paper deadline is May 4, 2014. Authors of recent journal papers can submit a proposal to present an overview poster.