Computer Science • 14 (1) 2013 http://dx.doi.org/10.7494/csci.2013.14.1.63

Mikołaj Leszczuk
Błażej Szczerba
Andrzej Głowacz
Jan Derkacz
Andrzej Dziech
Piotr Romaniak

LARGE-SCALE RESEARCH ON QUALITY OF EXPERIENCE (QoE) ALGORITHMS

Abstract *The large variety of video data sources means variability not only in terms of included content, but also in terms of quality. Therefore, quality assessment provides an additional dimension. The paper describes a comprehensive evaluation experiment on perceived video quality. Consequently, in summary,* 19 200 000 *video frames will be processed. Given the scale of the experiment, it is set up on a computer cluster in order to accelerate the calculations significantly. This work on Quality of Experience (QoE) is synchronized with that conducted by the Video Quality Experts Group (VQEG), in particular the Joint Efforts Group (JEG) – Hybrid group project.*

Keywords Video, compression, QoE, MOS, H.264

1. Introduction, motivation and objectives

The large variety of multimedia data sources means variability not only in terms of included content, but also in terms of quality. This applies in particular to video sequences, where quality is extremely variable, e.g. recordings currently made with an amateur camera at one extreme and recordings of specialized medical tests for diagnostic purposes at the other. Therefore, quality assessment provides an additional dimension, especially considering the number of places in the transmission chain where distortions can be introduced, namely (as shown in the Figure 1):

1. Original video acquisition.
2. Encoder and streaming server.
3. Network.
4. User equipment.

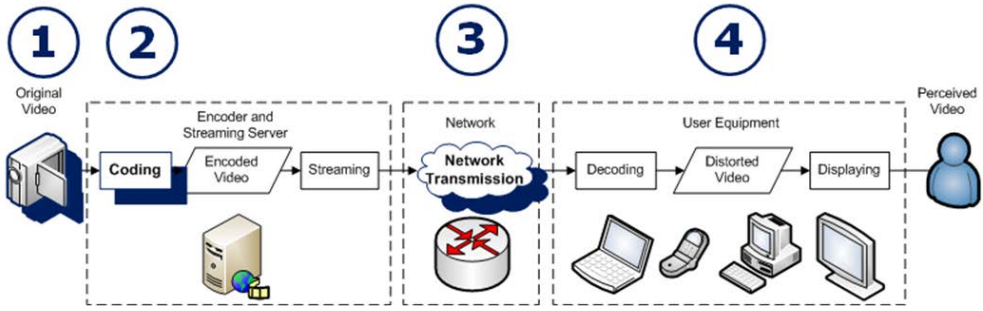

Figure 1. Transmission chain.

In terms of mechanisms delivering high-quality digital content, a common feature of existing systems is the need to ensure acceptable quality of streaming video sequences, regardless of the load on the transmission medium, the type of access network, or the end users equipment. The term "acceptable quality" is not clearly defined, and mostly depends on the scenario under consideration. Video streaming systems incapable of providing acceptable quality ceases to be attractive to potential users, and as such they are not in general use. Due to the above requirements, an important element of the research is to develop a system to ensure adequate quality of video sequences. This system should include metrics for quality assessment and quality optimization mechanisms that use information provided by the metrics.

The first objective of the work presented in the paper is to process video sequences according to Hypothetical Reference Circuit (HRC) parameter sets, consisting of various bit-rate/quantization factors, Group of Pictures (GoP) sizes and structures, structures of slices, numbers of frames per second as well as resolutions. The second objective is to calculate, for each sequence, video quality metrics and supporting video content indicators.

The rest of the paper is structured as follows. Section 2 briefly describes state of the art projects, forming the basis for this research. Section 3 discusses the innovative aspects of the research. Experiment implementation is presented in Section 4. Section 5 presents the preliminary results. The paper concludes and future work is outlined in Section 6.

2. State of the art

The Video Quality Experts Group (VQEG) has been studying quality assessment for several years [11]. For example, one of the previous large-scale evaluations from VQEG, "Multimedia Test-Plan", consisted of 43 subjective experiments with 160 sequences each.

The JEG-Hybrid Project. The work on QoE algorithms was synchronized with that conducted by the VQEG, in particular the project of the Joint Efforts Group (JEG) – Hybrid group [3]. The JEG-Hybrid Group is an open collaboration working together to develop a robust Hybrid No-Reference model that uses the bitstream and the decoded video. The JEG has developed and made available routines to create and capture bit-stream data and parse bit-streams into HMIX (Hybrid Model Input XML) files. Subjectively-rated video quality datasets with bit-stream data that can be used by all JEG researchers are currently under development. The objective of the JEG-Hybrid group is to assess the quality of video sequences based on the bit stream and decoded frames and motion, that is, information that is readily available at the decoder side. The first assessment of the JEG-Hybrid group is limited to video encoding based on the H.264 standard and compatible with packet transmission. Pre-transmission processing may include any pre- and post-processing and transcoding, provided that the last step is H.264 compatible. The ultimate goal of the JEG effort is to create an objective video quality measurement model that combines metrics developed separately by a variety of researchers [2, 5].

The SYNAT Project. One of the aims of research within the SYNAT project is developing methods of delivering high-quality multimedia content. There are also aims to implement a variety of methods for segmentating of digital objects, transcoding and adaptating multimedia content. This will include metrics and algorithms for assessing the quality of video sequences. Plans are underway to store additional information about the quality of digital objects for the materials that meet the criteria.

Metrics. For each video frame, video quality metrics need to be calculated. Peak Signal-to-Noise Ratio (PSNR) is probably the best known metric [13]. Other metrics under consideration include the Structural Similarity Index (SSIM) [12] and the Video Quality Metric (VQM) [14]. Other related metrics include TetraVQM [1], VIF [8], MOVIE [7] and JND.

3. Experiment innovation

The typical approach to learning about video quality involves conducting subjective experiments. They are time- and resources-consuming; it is also necessary to collect appropriate video content (see Figure 2 for example test sequences).

Figure 2. Example test sequences [6].

While 24 subjects (testers) are enough to test a small set of parameters (i.e. based on the VQEG test plan [10]), it is often necessary to use up to 100 testers for more parameters. The test, using the widely accepted ITU's ACR-HR (Absolute Category Rating with Hidden Reference, ITU-T P.910) [9] methodology, involves screening video sequences and collecting quality marks, one by one (see Figure 3).

Figure 3. Methodology: ITU's ACR-HR (Absolute Category Rating with Hidden Reference, ITU-T P.910 [9].

However, it is worth considering whether these experiments are necessary.

The main innovation of the proposed solution is a comprehensive evaluation of the perceived quality of video, and optimization based on information provided by the proposed metrics. The complexity of the evaluation is related to the fact that it uses distortions typical of the video acquisition process and those caused by lossy compression for streaming. Through subjective testing and application of advanced statistical methods for the analysis of data received, it will be possible to create models mapping measured quality parameters onto values for the resultant video quality, expressed in terms understandable to the user (e.g. 5-point scale of the Mean Opinion Score, MOS). Another important issue is to optimize the quality of the information provided by the metric, written in the form of metadata, or computed on demand. Optimizing the quality will take into account not only the network parameters (e.g. available bandwidth) but also the characteristics of video sequences, which is an important innovation.

4. The video processing experiment implementation

Recent activity in the SYNAT project, in parallel with the JEG-Hybrid group, resulted in the selection of sets of Source Video Sequences (SRCs) and of parameter sets – HRCs.

4.1. Collections of Source Reference Channels/Circuits (SRCs)

In this experiment, it was assumed that there are no restrictions imposed on the overall quality of the SRC source video sequences. Some noise from dim lighting conditions in a clear image, can be still considered as high quality. However, the SRC content, which is of lower quality than MOS value 4.0, needs to be treated separately. For example, for measurements in the Full Reference (FR) model (Figure 4), if a reliable prediction of subjective assessment is required, this type of content cannot be used (please see below for more details).

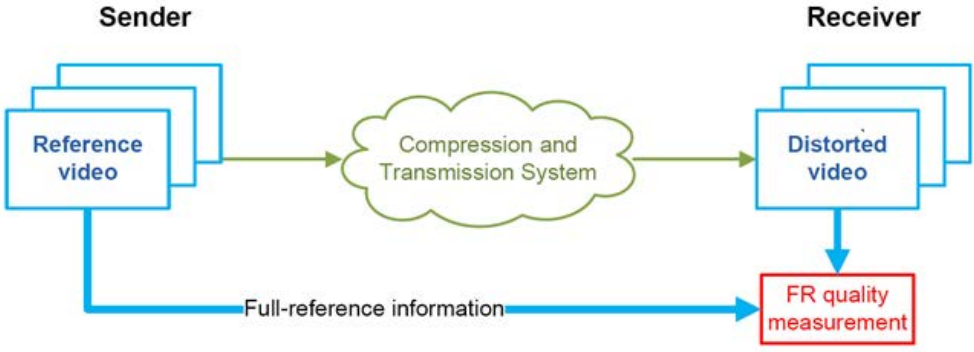

Figure 4. FR scenario.

The assumption states that No Reference (NR) models (Figure 5) tend to be less accurate than FR models, therefore the latter can serve as reference.

Figure 5. NR scenario.

This is a preliminary estimate. Certain video sequences can be repeated in some collections, which will require further verification. Many sequences originate from the database of the Consumer Digital Video Library (CDVL) [4].

SRC sequences have been divided into four classes:

- ivory sequences (subsection 4.1.1),
- bronze sequences (subsection 4.1.2),
- silver sequences (subsection 4.1.3),
- gold sequences (subsection 4.1.4).

4.1.1. Ivory sequences

This class contains all SRC sequences which are considered in the project that have a MOS value less than 4.0. They are not eligible for use as reference in FR measures.

4.1.2. Bronze sequences

This class is on the same organisational level as the Ivory SRC. Bronze sequences are those SRC which get a MOS value larger than (or equal to) 4.0. They are considered acceptable as reference input to FR measures in order to get potentially reliable MOS predictions.

4.1.3. Silver sequences

In order to identify sequences which are particularly representative for a certain type of source content such as colourfulness, spatial detail, movement or type of acquisition such as cartoon, Computer-Generated Imagery (CGI), etc. some of the bronze sequences are bundled in the silver sequence set. This sequence set may have any size that is considered useful by the JEG-Hybrid group though the number should

be limited to a reasonable value, e.g. 50. It should stay fixed for a certain period in order to allow references to this particular set. For each sequence the reason why it was considered as a silver sequence shall be described in 1–3 sentences.

4.1.4. Gold sequences

The gold sequences are unique sequences from the silver sequence set that are selected for their spread of the identified characterisation ranges. Their number is limited to 12 which allows to include all of them in one subjective experiment. They should not contain Ivory sequences. The sequence set may be updated when new sequences arrive. For each sequence the reason why it was considered as a silver sequence shall be described in 3–10 sentences.

4.2. Hypothetical Reference Circuit (HRC) parameter sets

The set of HRC parameters defines the parameter set for a video encoder which can be used for compression. Each of the proposed HRCs differs from the others in at least one of the parameters under consideration. The entire range of HRCs should cover all probable video compression scenarios, designed for digital storage in digital libraries. Each SRC should be distorted (encoded) using each of the HRCs, which will result in a number of distorted Processed Video Sequences (PVS) equal to SRC×HRC sequences.

The video compression parameters under consideration are as follows:

- Bit-rate/quantization factor, with at least 5+5 choices.
- Different GoP sizes and structures:
  - of at least 2 choices with a variable number of I frames per second,
  - of at least 2 choices with a variable number of B and P frames per second,
  - hierarchical coding – of at least 2 horizontal variants.
- Structure of slices – of at least 2 different lengths.
- Number of frames per second – of at least 2 choices (original and halved).
- Resolution – of at least 2 choices (original and halved).

4.3. Experiment scale

The effort required to generate this database is significant. In order to highlight this, the number of bit streams that will be created is calculated as follows. First of all, approximately 100 SRCs need to be selected. Each one needs to be:

- 10 seconds long,
- with 30 FPS, and
- with resolution ranging from SD ($720 \times 480/576$) to HD (1920×1080).

HRCs need to be selected next, currently giving a total of 640 combinations. Consequently, in summary, $100 \times 10 \times 30 \times 640 = 19\,200\,000$ video frames will be processed.

- Each video frame needs to be encoded (1 s is required to encode a single video frame using JM Reference Software using a single CPU core).
- For each video frame, video content characteristics need to be calculated.
- For each video frame, video quality metrics need to be calculated.

PSNR is probably the best known metric. It is a data metric, meaning that it looks at the fidelity of the signal without considering its actual content [13]. Data metrics have been an extremely popular research topic in the past, and they are widely used in image and video quality assessment. However, they always operate on the whole frame and do not consider any other important factors that can strongly affect the perceived quality (such as human visual system, HVS, characteristics). Figure 6 shows distorted pictures with the same quality according to the PSNR metrics. However, it is clear that the perceived quality is very different. The conclusion is that data metrics have a relatively low performance and low feasibility.

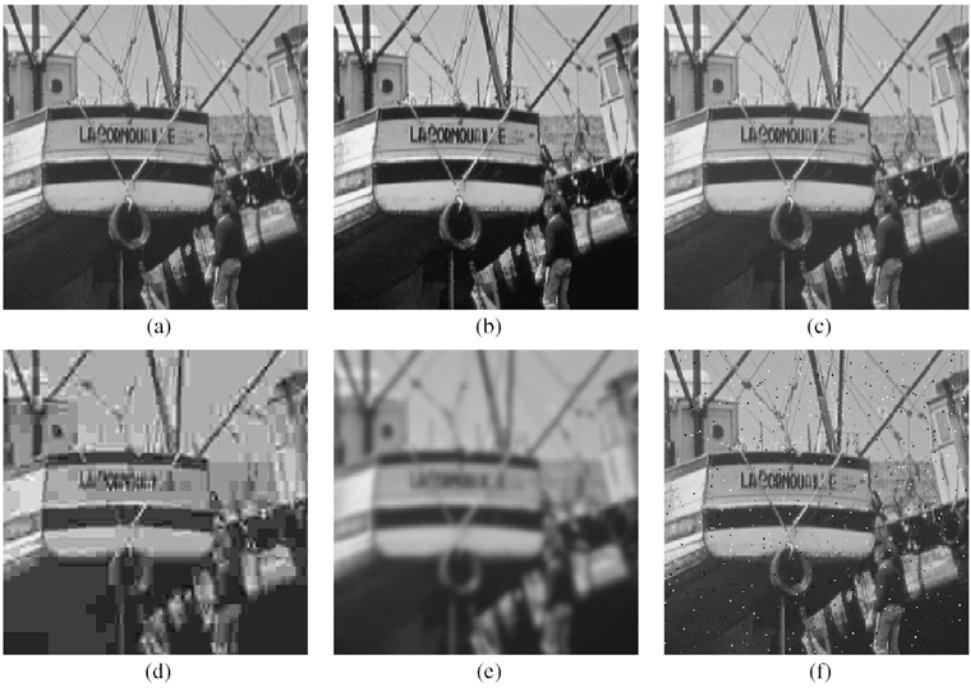

Figure 6. PSNR – is it good enough?

Consequently, other metrics under consideration include the *SSIM* [12] and the *VQM*.

The SSIM is a top-down approach (see Figure 7) using a functional model of the HVS (a more detailed description of HVS can be found in [13]).

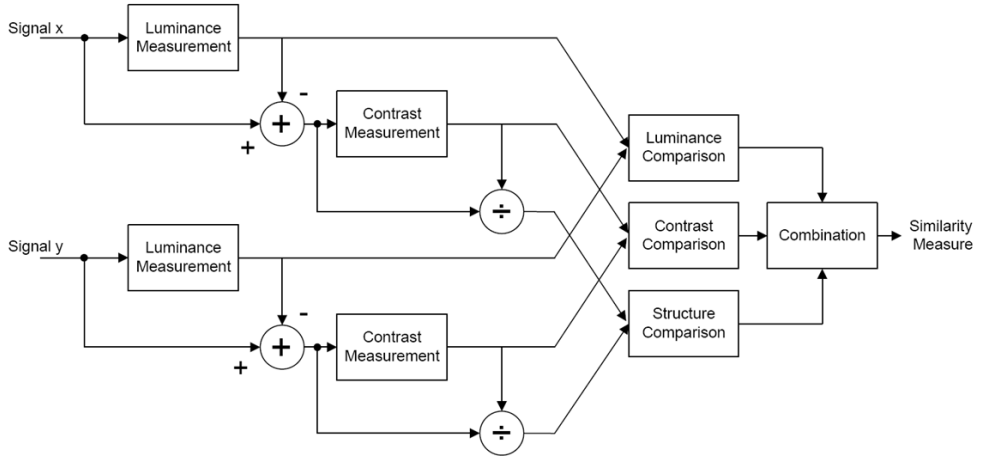

Figure 7. SSIM.

The VQM (see Figure 8) measures the perceived effects of video impairment including blurring, jerky/unnatural motion, global noise, block distortion and colour distortion, and combines them into a single metric [14].

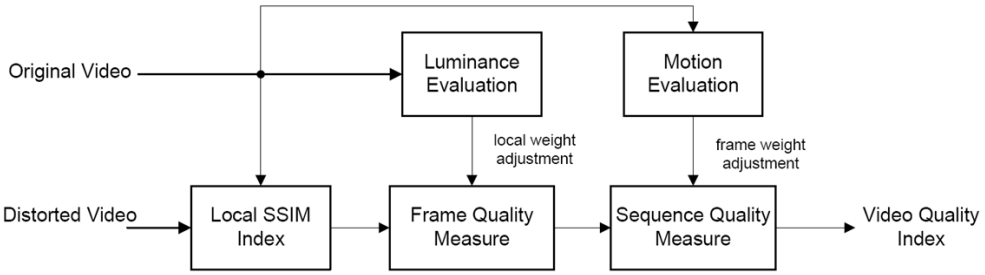

Figure 8. VQM.

Video quality metrics are supported by calculating of the following video content indicators:

- Blockiness.
- Blurriness.
- Exposure.
- Noisiness.
- Flickering.
- Spatial Activity (Intensity).
- Temporal Activity (Intensity).

- Spatial Correlation.
- Energy.
- Homogeneity.
- Variance.
- Contrast.
- Colour Layout Descriptor.
- Edge Histogram Descriptor.

4.4. Calculation time using a single core

Calculation time using a single core has been estimated as approx. 889 days. This was based on the following assumptions:

Assumption 1: 19 200 000 video frames.

Assumption 2: compression of a single video frame in 1 s.

Assumption 3: three video quality metrics to be calculated for all video frames.

Assumption 4: calculation of a single video quality metric for one frame in approx. 1 s.

Given the above, the experiment is set up on a computer cluster in order to accelerate the calculations significantly.

5. Results

The ZEUS supercomputer, located at the Academic Computer Centre CYFRONET in Kraków, was selected as the primary machine to encode a set of PVS. They were encoded using JM Reference Software version 18.3. Computational tasks are sent to the ZEUS cluster by an access machine UI with a Torque Resource Manager queuing system. To generate the HRC, appropriate configuration files need to be generated, which are the JM encoder parameter. Configuration files are created using the Bash scripting language.

Performance tests have been conducted, which helped determine the time of a single PVS coding session as 8–12 hours depending on the encoding parameters. One of the limitations that are currently being considered is the maximum number of computational tasks that are being run at the same time by a single user at the ZEUS cluster. Depending on the type of the queue, this varies between 10–1000 tasks. Currently, the test calculations are being performed for selected SRC; it is expected that numerous PVSs will be created.

Test calculations were performed for 10 SRCs. We defined six groups of video coding parameters followed by 666 HRCs.

We started with 900 HRCs; however ,certain video parameters and their levels were inadequate, as they do not significantly improve the quality differences between PVSs (i.e slice groups). The number of HRCs was calculated by dividing video coding parameters into four sets (full matrix, spatial and temporal resolution reduction,

motion estimation parameters, and I/P/B ratio) and creating these matrices. Each set is related to specific video quality indicators that could affect video qualities such as bit-rate, quantization factor, spatial resolution, temporal resolution, GoP structure, slice mode, etc. An important conclusion resulting from the analysis of encoded sequences is the fact, that certain video coding parameters should be defined during the preprocessing stage (i.e. spatial and temporal resolution) as the, encoder was unable to do it.The second conclusion was that it is necessary to use a JM decoder ,required for the correct playback and re-compression, which are not guaranteed for all decoders.

6. Conclusions and future work

Preliminary results of the research have been passed on to other members of the JEG as configuration files, JM/x264 encoded sequences and other supporting files. This will allow researchers to verify the encoding parameters of the resulting video sequences and determine the next steps of the experiments (in particular dates and resources to store PVS and uncompressed sequences). Following arrangements, new input SRCs will be prepared and new calculations conducted. Calculations resulted in the creation of a very large collection of test sequences that can be used in the subjective and objective tests, which can be used to develop appropriate algorithms for video encoding.

The resulting data will be available to the community as they may be useful in other projects as well.

In the next steps, different quality assessment algorithms will be compared and the results of the comparison will be analysed.

Acknowledgements

Work financed by The National Centre for Research and Development (NCBiR) as part of project no. SP/I/1/77065/10.

References

[1] Barkowsky M., Bialkowski J., Eskofier B., Bitto R., Kaup A.: Temporal trajectory aware video quality measure. *Selected Topics in Signal Processing*, 3(2): 266–279, 2009.

[2] Barkowsky M., Staelens N., Janowski L., Koudota Y., Leszczuk M., Urvoy M., Hummelbrunner P., Sedano I., Brunnström K.: Subjective experiment dataset for joint development of hybrid video quality measurement algorithms. In *EuroITV – 10th European Conference on Interactive TV*, pp. 254–257, Berlin, July 2012. Fraunhofer Institute for Open Communication Systems, FOKUS.

[3] Barkowsky M., Staelens N., Janowski L.: *The JEG Hybrid Group*. VQEG. `http://www.its.bldrdoc.gov/vqeg/project-pages/jeg/jeg.aspx`.

[4] CDVL. *The Consumer Digital Video Library.* http://www.cdvl.org/.

[5] Leszczuk M., Głowacz A., Derkacz J., Dziech A., Romaniak P., Szczerba B.: Large-scale video compression research work on quality of experience (QoE) evaluation for VQEG JEG-Hybrid project. In *6th International Symposium on signal, Image, Video and Communications ISIVC*, pp. 188–191, Valenciennes, France, Jul 2012.

[6] Mu M., Romaniak P., Mauthe A., Leszczuk M., Janowski L., Cerqueira E.: Framework for the integrated video quality assessment. *Multimedia Tools and Applications*, pp. 1–31, 2012. 10.1007/s11042-011-0946-3.

[7] Seshadrinathan K., Bovik A. C.: Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing*, 19(2): 335–350, 2010.

[8] Sheikh H. R.: Bovik, A visual information fidelity approach to video quality assessment. In *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp. 23–25, 2005.

[9] International Telecomunication Union: ITU-T P.910, Subjective video quality assessment methods for multimedia applications. 1999.

[10] VQEG. *Test-Plan for Evaluation of Video Quality Models for Use with High Definition TV Content.*

[11] VQEG. *The Video Quality Experts Group.* http://www.vqeg.org/.

[12] Wang Z., Lu L., Bovik A. C.: Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121–13, 2004.

[13] Winkler S.: *Digital Video Quality – Vision Models and Metrics.* John Wiley & Sons, Ltd. 2005.

[14] Wolf S., Pinson M. H.: Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system. In *Proc. SPIE*, 3845:266–277, 1999.

Affiliations

Mikołaj Leszczuk
AGH University of Science and Technology, Krakow, Poland, leszczuk@agh.edu.pl

Błażej Szczerba
AGH University of Science and Technology, ACC CYFRONET, Krakow, Poland, blazejszczerba@gmail.com

Andrzej Głowacz
AGH University of Science and Technology, Krakow, Poland, glowacz@kt.agh.edu.pl

Jan Derkacz
AGH University of Science and Technology, Krakow, Poland, derkacz@kt.agh.edu.pl

Andrzej Dziech
AGH University of Science and Technology, Krakow, Poland, adzie@tlen.pl

Piotr Romaniak
AGH University of Science and Technology, ACC CYFRONET, Krakow, Poland, romaniak@agh.edu.pl

Received: 10.09.2012
Revised: 19.10.2012
Accepted: 3.12.2012