Open collaboration on hybrid video quality models - VQEG joint effort group hybrid

Marcus Barkowsky, Nicolas Staelens, Lucjan Janowski

Open collaboration on hybrid video quality models – VQEG Joint Effort Group Hybrid

Marcus Barkowsky[1], Nicolas Staelens[2], Lucjan Janowski[3]

[1] *LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Rue Christian Pauc, 44306 Nantes, France,*
Marcus.Barkowsky@univ-nantes.fr

[2] *Ghent University - iMinds, Department of Information Technology, Ghent, Belgium*
nicolas.staelens@intec.ugent.be

[3] *AGH University, 30 Mickiewicza Av. PL-30-059 Krakow, Poland*
janowski@kt.agh.edu.pl

***Abstract*—Several factors limit the advances on automatizing video quality measurement. Modelling the human visual system requires multi- and interdisciplinary efforts. A joint effort may bridge the large gap between the knowledge required in conducting a psychophysical experiment on isolated visual stimuli to engineering a universal model for video quality estimation under real-time constraints. The verification and validation requires input reaching from professional content production to innovative machine learning algorithms. Our paper aims at highlighting the complex interactions and the multitude of open questions as well as industrial requirements that led to the creation of the Joint Effort Group in the Video Quality Experts Group. The paper will zoom in on the first activity, the creation of a hybrid video quality model.**

I. INTRODUCTION

Since many decades it is well-known that Peak Signal to Noise Ratio (PSNR) has severe limitations. This has been shown both theoretically as well as experimentally [1],[4]. Nevertheless it is still the most often used measurement tool. The list of alternatives is endless and outside the scope of this paper. For more information on objective video quality assessment, the interested reader is referred to [1]. The development of each such alternative objective quality metric requires time and effort. For most of them their authors showed and eventually verified that they performed better than PSNR. Few of them were validated. None of them has replaced PSNR so far.

In scientific literature, objective video quality metrics usually have a rather limited scope and applicability. Often, specific use cases are envisaged and some specific test conditions are considered for creating a number of impaired video sequences. These sequences with their limited scope are then used to set up a subjective experiment and train and verify new objective video quality metrics. As such, the validity of these metrics is usually restricted to the considered use case. Furthermore, the (impaired) video sequences and corresponding subjective quality ratings are too often kept secret. For example, the use of copyrighted video materials might prohibit the redistribution of particular video sequences.

In industry, execution speed and ease of use are often considered paramount features. Under these conditions, simple measurements as PSNR are welcomed. Several companies offer more complex solutions to the industry which have a higher prediction performance. The International Telecommunication Union (ITU) has recommended several methods for different scopes that represent typical applications such as mobile transmission or Full-HD IPTV. These methods have often been validated in an open competition approach within the Video Quality Experts Group (VQEQ) showing that they outperformed PSNR in a statistical sense.

With the start of VQEG's Joint Effort Group (JEG), a collaborative approach is now followed towards constructing novel objective video quality metrics [2]. As opposed to the competitive approach traditionally used within VQEG, JEG encourages and facilitates a free, open, and joint collaboration in subjective and objective video quality assessment. Within JEG, different members from universities, model developing industries, and video service providers join forces. By combining all available know-how during every stage of objective video quality metric design and development,

MMSP'13, Sept. 30-Oct. 2, 2013, Pula (Sardinia), Italy.
978-1-4799-0125-8/13/$31.00 ©2013 IEEE.

more in-depth and profound video quality research can be conducted as a whole.

The paper provides an overview of recent standards by the ITU and their validation process in Section 2. Advantages and limitations are highlighted. In Section 3, the newly proposed joint approach is summarized, including current and future activities. Section 4 concludes the paper.

II. ITU Standardization for Video Quality Measurement

Standardisation is considered important because of the reproducibility and the traceability of algorithms and their performance. Notably in ITU standardisation important fairness rules apply. ITU Recommendations have therefore often an important impact on the industry producing or using technology tackled in standardisation. A prominent example is video coding which allows for the efficient exchange of media between various entities regardless of manufacturer or country. The manufacturers are ascertained that they can provide the technology on equal terms with their competitors.

In subjective video quality assessment, most of the standards have been created in a collaborative manner. Discussions have led to refining standards such as ITU-R BT.500 or ITU-T P.910 to name two examples [5].

A. Competitive validation procedure for objective video quality models

For objective video quality, companies underwent a competition phase in order to determine the best algorithm. In most cases, several algorithms showed similar performance and were therefore standardized.

The competition phase often took place within VQEG in the following way:

1. The proponents developed independent objective models for a given application area.
2. A common document (test plan) was created containing the limits of the application area such as technical constraints on the degradations, the subjective assessment method for creating the validation databases, the procedure for validation including the statistical analysis, and the way in which the models would be ordered.
3. Proponents submitted executable programs and eventually encrypted source code, so called frozen models, to a member of the Independent Laboratory Group (ILG)
4. Processed Video Sequences (PVS) were generated obeying the restrictions given in the test plan on known and secret content
5. Subjective assessments were performed creating the Mean Opinion Scores (MOS) for validation
6. Based on the statistical analysis of each model's prediction performance the models were ordered
7. Standardisation in the ITU was proposed if the best model outperformed PSNR, which was eventually adapted to the requirements of the application area
8. The ITU recommended in most cases all models which performed statistically equivalent to the best performing model

In this way, ITU-T J.144 [6] was created for Standard Definition (SD) television at 50Hz and 60Hz including interlaced and progressive, ITU-T J.246 [6] and J.247 [8] for multimedia signals in VGA (640x480 pixels, Video Graphics Array), CIF (352x288, Common Intermediate Format), and QCIF (176x144, Quarter CIF) format including severe coding artefacts, transmission degradations, and pausing and skipping. J.247 requires access to the source video signal and is therefore called Full-Reference (FR) while J.246 requires only access to a signal that was extracted from the source video signal which may be transmitted as side information, a so-called Reduced Reference (RR) algorithm. ITU-T J.341 [9] allows for measuring Full-HD video signals in 1080i25, 1080i30, 1080p25, and 1080p30. J.341 and J.246 contain a single algorithm while J.144 and J.247 contain four different algorithms each.

In the process of this standardisation, various subjective databases have been created, notably the two well-known VQEG Phase 1 SDTV databases, two hidden databases for SDTV Phase 2, 14 QCIF, 14 CIF, and 13 VGA databases, and 6 databases for HDTV of which 5 are freely available via the Consumer Digital Library (CDVL) [10].

Currently, VQEG evaluates objective measurement algorithms that do not require access to the source video sequence. Instead, information in the decoded video and the bitstream as transmitted over the network is exploited. These models are called Hybrid-No-Reference algorithms.

B. Limitations

The current approach for standardising objective video quality prediction algorithms is well established, the procedure is sound, and allows for fair conditions

and equal terms. There are some drawbacks which should be considered, particularly when comparing to other standardisation activities such as video coding:

- Delay: Due to the validation procedure, the typical delay between the submission of the model and the standardisation is two to three years.
- Test conditions: Although a large number of test conditions have been evaluated in subjective experiments conducted for validation as described above, they may not suffice to cover the entire application scope and to test for robustness of the models
- Secret content: Validation requires content that is not known prior to model submission. This content usually needs to be shot particularly for each validation data set. Shooting video sequences that are balanced in terms of visual features is a time consuming and costly task.
- Exploitation: The results are often only used for the standardisation. This applies both to the analysis results because of their usage restrictions, and to the video sequences that become available only after the standardisation has finished.
- Critical model performance analysis: Evaluation of the model's performance in different parts of the targeted scope is often difficult. As an example, it may be seen that some of the models for multimedia sequences in J.247 do not measure explicitly frame rate or pauses and skips within the analysed video sequence. It has not been analysed whether this impacts on the model's performance for these particular conditions.
- Missing continuity: The standardized models do not provide a basis for future developments by the standardization group. Each proponent improves his own model for the next competition. Splitting the analysis into building blocks and comparing the performance of the algorithms' internal indicators may show advantages and weaknesses that can be exploited for the next version of the standard.
- Reproducibility: A rigorous test whether the standards contain all required information for implementing the algorithms has not been undertaken. Despite the enormous and time consuming effort of each proponent to document their algorithm, information may be missing.

III. Collaborative Approach

Most of the above mentioned limitations are a consequence of the competitive approach. Therefore, a collaborative effort was started in order to continuously work on improvements of objective video quality measurement. This includes the creation of source video databases, processed video sequences and associated hardware and software production, the development of additional tools, and the storage of side information.

A. Overview and Advantages

A graphical overview of the complexity of the approach is provided in Figure 1. The diagram mostly focuses on the engineering of an objective model. In particular it does not show the required prerequisites such as psychophysical studies. The diagram shows the steps for creating reasonable source content on the top, degrading those videos in the middle, and at the bottom the development of an objective model. Each part contains an iteration loop and the complete process requires iterations as well.

In a joint effort approach, priorities are significantly different from the competitive approach. In particular, a large database that is well balanced may reduce the requirements for independent validation as the model parameters are known. This allows, for example, avoiding overtraining by consequent analysis of fitted parameters. Following this approach, the creation of a large scale database that allows testing for robustness becomes an important factor.

Therefore, the joint effort started with the creation of this database. Several research questions have since then been posed such as the selection of different balanced subsets from a large video source set. The measurement of the uniform coverage of sequence characteristics has previously been mostly limited to simple measures such as Temporal perceptual Indicator (TI) and Spatial perceptual Indicator (SI), defined in ITU-T P.910. The notion of art work and the question of the perceived difference between professionally shot content, such as in cinema or TV broadcast, and user generated content, such as vacation videos or informal videos found on internet platforms, was tackled. The annotation of source content with meta information such as visual attention may be considered, bringing up the question whether video quality assessment as a task is realistic for watching videos of different type, i.e. entertainment or educational.

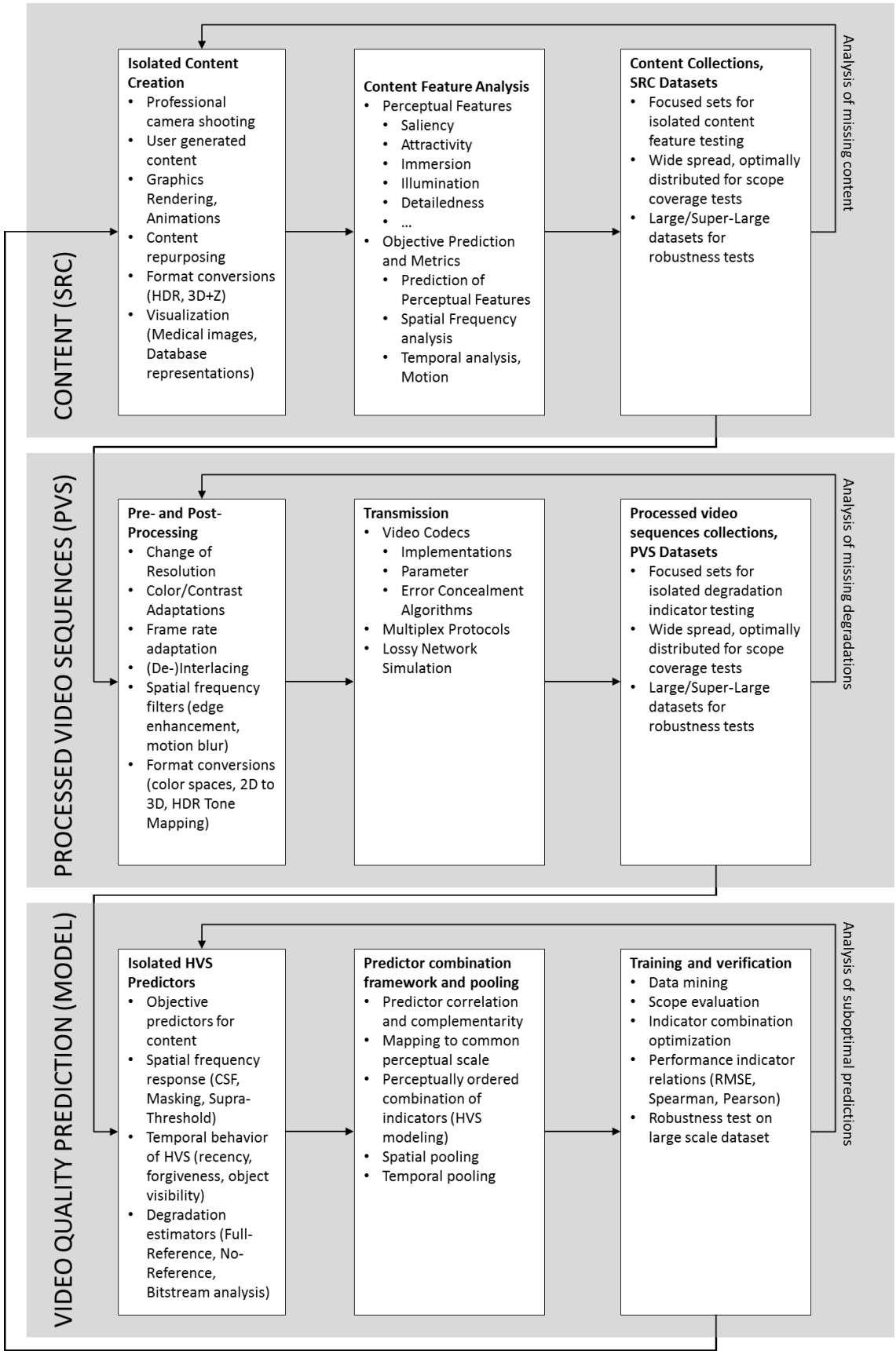

Figure 1: Overview of processing steps to develop, train and verify a video quality assessment model

The induced degradations are no longer restricted by rules available in a test plan document. Again the measurement of uniform coverage of the scope becomes important, introducing notably requirements for the definition of distinguishable perceptual degradations as different technical parameters may lead to similar perceived degradations.
A toolchain was developed that allows the automated creation of large datasets of degraded videos. Annotating the videos with video quality scores may be performed in different ways, ranging from FR measurement algorithms, over crowd sourcing strategies to formal subjective experiments in lab conditions. Combinations of several objective measurements and subjective experiments are possible and important progress is expected to be achieved by the application of data mining and machine learning algorithms.

Model construction requires the development of isolated indicators which are backed up by psychophysical studies and experimental validation on relevant parts of the large scale database. It also requires establishing a common framework for integrating the different measurements. This is particularly important because the structure of the model determines the way in which the human visual system can be modelled. Evaluating the performance of different strategies known to the perception and psychophysical community on a large amount of annotated databases may provide insight into the processing of visual information by the human observer.

B. Organisation

The joint effort has been started within the VQEG. VQEG's JEG is free and open to everyone. No subscription fees are involved for joining VQEG JEG. Contributions can be made concerning every step involved in subjective and objective video quality assessment. Furthermore, JEG encourages contributions from both academia and private industries.

Since its start, several important links have been established via Liaison Statements to other organisations and networks. An important example is the European Network on Quality of Experience in Multimedia Systems and Services, Qualinet, COST IC2003. Continuous exchange between the organisations led to several innovative publications.

C. Current status and future activities

Currently, JEG focuses on the following topics:

- Creation, analysis, and complementarity of source video sequences for a large scale database to be used freely by the research community
- Providing an easy-to-use tool chain enabling the creation of video sequences containing compression and/or transmission artefacts
- Creation of a large scale database containing degraded videos with various features
- Enabling the development of coding standard independent Hybrid Video Quality metrics by converting the encoded bit-stream data into parsed information stored in XML text files
- Researching the feasibility of measuring the video quality on more than ten thousand videos
- Developing new statistical methods for evaluating and validating metrics

In the near future, JEG plans to launch call for proposals for hybrid measurement methods evaluated on the large scale database. This will include calls for evidence to both, the industry as to the academic world.

D. Research topics and cooperation

JEG-Hybrid's preliminary results discovered new scientific questions which should be addressed by future research. We would also like to encourage the scientific community to use tools and databases already created within JEG Hybrid. For each of the following research questions, a point of contact is provided on the VQEG JEG-Hybrid homepage [2]. If you have any question and you would like to contribute, please get in contact.

Existing data base

We created a database of more than 10000 sequences. Those sequences cover numerous different compression parameters. In addition some full and non reference metrics were computed for all sequences within the database.

Combining multiple subjective and/or objective scores

A large database calls for specific model estimation. Machine learning algorithms or data mining algorithms may be applied. For some of the sequences, subjective ground truth data obtained in a formal experiment is available.

Single FR measurement scope and prediction accuracy

FR measurements computed for a large number of sequences does not necessary result in a similar quality estimation. A methodology for selecting a correct answer in case of a FR metrics disagreement is needed.

Parsed bitstream data in XML format

One of JEG Hybrid's goal is to provide a standardized XML data format combining all transmission and compression information available in the transmitted bitstream for currently wide-spread coding standards, i.e.

blockbased H.264, H.265. This greatly facilitates the construction of Hybrid Models applicable across protocols and coding standards. Initial tools for parsing network bitstreams and in RTP format and for parsing H.264 Annex-B files exist. Volunteers are welcome for the consideration of further transmission and multiplex protocols and H.265 video coding.

Including new databases

We are very much interested in including more subjective experiments' results. Even if your experiment covers a small part of the hybrid methodology (for example compression only) your results are very helpful and welcome.

Including new metrics

We are very much interested in including more metrics independent of their methodology FR, RR, NR or Hybrid. Please join the effort and test your code on the databases.

IV. CONCLUSIONS

Future developments of objective video quality measurement require algorithms with increasing complexity that need to be evaluated on large datasets in order to prove their performance and robustness in real world scenarios. While simple algorithms may provide some rapid answers, their performance is far lower than the performance of human evaluation of video quality. A collaborative approach allows for combining different research domains in order to achieve an optimized algorithm which continuously approaches the prediction performance of a group of human observers in a subjective assessment. Advantages of the joint effort have been detailed and possible future developments have been traced in this paper. Current and future VQEG-JEG activities have been summarized.

ACKNOWLEDGEMENTS

Lucjan Janowski's work was financed by The National Centre for Research and Development (NCBiR) as part of project no. SP/I/1/77065/10.

REFERENCES

[1] S. Chikkerur, V. Sundaram, M. Reisslein, L. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison", *Broadcasting, IEEE Transactions on*, *57*(2), 165–182, 2011

[2] Video Quality Experts Group. (2007) Project Homepage of VQEG Joint Effort Group - Hybrid. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/jeg/jeg.aspx

[3] B. Girod, "What's Wrong with Mean-squared Error?" In Andrew B. Watson (Ed.), Digital Images and Human Vision (pp. 207–220). Massachusetts Institute of Technology, 1993

[4] Z. Wang, A. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Transactions on Image Processing, 13(4), 600–612, 2004

[5] *Subjective video quality assessment methods for multimedia applications*. ITU-T P.910, 1997

[6] *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, ITU-T J.144, 2004

[7] *Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference*. Recommendation ITU-T J.246, 2008

[8] *Objective perceptual multimedia video quality measurement in the presence of a full reference*. Recommendation ITU-T J.247,2008

[9] *Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*, Recommendation ITU-T J.341, 2011

[10] (2013) Consumer Digital Video Library website. [Online]. Available: http://www.cdvl.org/