Objective Video Quality Assessment - Towards large scale video database enhanced model development

IEICE TRANS. COMMUN., VOL.E98–B, NO.1 JANUARY 2015

INVITED PAPER *Special Section on Quality of Diversifying Communication Networks and Services*

Objective Video Quality Assessment — Towards Large Scale Video Database Enhanced Model Development

Marcus BARKOWSKY[†a)], **Enrico MASALA**[††], **Glenn VAN WALLENDAEL**[†††], **Kjell BRUNNSTRÖM**[††††,†††††], **Nicolas STAELENS**[†††], ***and*** **Patrick LE CALLET**[†], ***Nonmembers***

SUMMARY The current development of video quality assessment algorithms suffers from the lack of available video sequences for training, verification and validation to determine and enhance the algorithm's application scope. The Joint Effort Group of the Video Quality Experts Group (VQEG-JEG) is currently driving efforts towards the creation of large scale, reproducible, and easy to use databases. These databases will contain bitstreams of recent video encoders (H.264, H.265), packet loss impairment patterns and impaired bitstreams, pre-parsed bitstream information into files in XML syntax, and well-known objective video quality measurement outputs. The database is continuously updated and enlarged using reproducible processing chains. Currently, more than 70,000 sequences are available for statistical analysis of video quality measurement algorithms. New research questions are posed as the database is designed to verify and validate models on a very large scale, testing and validating various scopes of applications, while subjective assessment has to be limited to a comparably small subset of the database. Special focus is given on the principles guiding the database development, and some results are given to illustrate the practical usefulness of such a database with respect to the detailed new research questions.

key words: *video quality assessment, large scale database, reproducible research*

1. Introduction

Despite several decades of research, the algorithmic prediction of Quality of Experience (QoE) for video services has not yet been widely adopted, neither by the industry nor by other research domains such as video coding. The root cause is the complexity of the task. Two sources of complexity may be distinguished. First, the complexity of the human's decision taking process for video QoE which involves the human visual system, cognitive influence factors, social and environmental influence factors etc. Second, the complexity of today's video services, starting from the content creation process, the capturing, encoding with continuously increasing dimensionality of parameters, stored and sent over chains of digital transmission channels, decoded, error concealed, postprocessed, and rendered on any unspecified display. Most research activities have focused on the first part and significant progress has been obtained, notably concerning the modeling of the human visual system in specific conditions. Psychophysical experiments have been conducted, for example to measure and model the spatial contrast sensitivity function [1], [2], to learn about the influence of non-fluent playback conditions [3], or the sensations due to visual discomfort while watching moving objects in 3D video [4]. This selection is meant to illustrate the narrow scope that may be covered in psychophysical studies and may give a hint about the difficulty to model the human response as a whole. It becomes evident that even the measurement algorithms that have been validated and were recommended by standardization bodies, may have limited accuracy when confronted with stimuli or viewing conditions that were not foreseen at development time. Some of these conditions have been revealed during independent validation which examined the algorithms by subjectively assessing and comparing selected video samples of the video services that they are meant to measure.

In this approach, the complexity of the service, as explained above cannot be fully exploited. Several major disadvantages become apparent. First, the algorithm's performance cannot be estimated for a particular use case, service or measurement situation. Second, as the dimensionality of influence parameters when designing such algorithms is larger than the validation database, conclusions on reasons for model failures are difficult to draw.

The ease of analysis whether an unexpected outcome originates from the system under test or from the measurement algorithm itself is one of the main reasons why Peak Signal to Noise Ratio (PSNR) is still popular. Despite decades of research, most researchers consider still PSNR as a valid ground truth. Compared to other measurements, PSNR has the advantage of fulfilling the criteria of a mathematical metric, providing reasonable rank ordering when used on the same content, and ease of mathematical exploitation by minimizing squared errors. Several different implementations exist, taking into consideration ITU-R BT.601 constraints or limiting the PSNR value to reasonable finite digital representations, i.e. 8 bit [5]. In terms of popularity, SSIM comes close today, emphasizing more on local structure similarity [6] but nevertheless staying closely linked to PSNR as shown in [7].

Many algorithms with their own indicators have been developed. They are typically targeted towards a certain measurement situation such as near-lossless quality, or low-bitrate scenarios, or packet loss situations. Depending on whether the full reference video is available for measure-

Manuscript received June 10, 2014.
Manuscript revised August 21, 2014.
[†]The authors are with University of Nantes, France.
[††]The author is with Politecnico di Torino, Italy.
[†††]The authors are with Ghent University - iMinds, Belgium.
[††††]The author is with Acreo Swedish ICT AB, Sweden.
[†††††]The author is with Mid Sweden University, Sweden.
a) E-mail: Marcus.Barkowsky@univ-nantes.fr
DOI: 10.1587/transcom.E98.B.2

ment, only an excerpt of it, or no information about the reference at all, the algorithms are categorized into Full Reference (FR), Reduced Reference (RR), or No Reference (NR). When the compressed and transmitted bitstream is analyzed only, the algorithm is called a bitstream model. Exploiting the bitstream information together with the decoded video is a comparably new approach, called Hybrid-FR, Hybrid-RR, or Hybrid-NR respectively.

A recent overview of video quality algorithms including a sophisticated categorization can be found in [8]. In [9], the authors focus on identification of internal indicators used by some well-known algorithms such that they may be exploited independently to widen the scope of application. Indicators should measure correctly when changes inside their scope of application appear but should stay neutral when the changes do not concern their measurement specificity, for example, a measure for annoyance of irregular frame skips should not be affected by a reduced frame rate.

It becomes evident that developing and training indicators has mostly been tackled with respect to their in-scope application but rarely concerning their out-of-scope (neutral) behavior. In addition, fusion and training algorithms are limited by small and biased training sets.

A possible solution is to create a large scale database of contents and conditions which can be reproduced at any time. Algorithms may be compared to each other prior to subjective testing. The algorithms can then be enhanced focusing on optimal accuracy in all cases. This enables successive elimination of outlier conditions by continuous improvement of the algorithms. Positive side effects are expected such as feedback on missing or incompletely modeled properties of the human visual system, requiring further psychophysical analysis.

The paper details this approach with the following structure: Sect. 2 motivates further the creation of the large scale database by enumerating the source of inaccuracies in the current development process. Section 3 gives an overview of the currently available algorithms and databases for training and verification of algorithms. The requirements and advantages of a large scale database will be detailed in Sect. 4. First exemplary performance analysis results of simple and medium complexity, well-known video quality measurement algorithms will be presented in Sect. 5, motivating the discussion on future research benefits and open questions in Sect. 6. The paper ends with a conclusion in Sect. 7.

2. Motivation - Inaccuracies in Current Algorithm Development

The development and the structure of most existing algorithms can be divided into several steps which introduce inaccuracies as follows.

1. Psychophysical experiments are conducted, limitations on the number of observations and reproducibility issues affect the precision.
2. Models are fitted to the obtained datapoints which are often a first order approximation of the obtained results.
3. Computational algorithms are developed to automatically measure the parameters which were distinctively selected in step 1, introducing measurement noise and often exceeding the narrow scope of the psychophysical study.
4. A selection of such algorithms returning indicators is implemented to measure a larger extent of effects, adding measurement noise due to the insufficient complementarity of the individual algorithms.
5. The results are summarized, often in three dimensions: Space, time, and indicators.
6. Training is performed against a limited number of video databases, which were obtained in particular conditions and which add further inaccuracies due to observation errors. Verification experiments often reveal that the chosen models are too simplistic and overfitting is a common issue.

In order to learn about the amount of inaccuracy in a specific algorithm, validation experiments are required. These are sparse in scientific publications and require joint efforts for standardization. The same limitations as for the training apply, i.e. the obtained ground truth data is noisy due to the chosen test conditions, the limited number of observations, and observation errors.

3. Available Databases for Training and Verification

Many video databases were created and published in recent years to serve for various purposes. Within the Qualinet action (www.qualinet.eu), a list of references to most publicly available video databases was collected [10]. In most database publications, the evaluation of a well-identified scientific question such as the influence of a coding or transmission parameter, or the comparison between video coding standards in a specific application scenario was targeted. The second largest effort was probably dedicated to the validation of objective video quality measurement algorithms, notably within the VQEG in preparation of the recommendations published by the International Telecommunication Union. Table 1 lists the application scope, provides a link to the final report, the resulting recommendation, and lists how many subjectively assessed video sequences have been used and published in each evaluation. It may be observed that for the validation of the recommended algorithms a large database was deemed necessary. Unfortunately, such a large database is difficult to obtain and in earlier VQEG phases,

Table 1 Overview of the databases published by VQEG.

Evaluation Scenario	Report	ITU Rec.	Database size used	published
Standard Television I	[11]	-	340	340
Standard Television II	[11]	J.144	128	-
Multimedia	[12]	J.246-247	5320	-
High Definition TV	[13]	J.341	888	744
Hybrid Bitstream	[14]	tbd	>1760	tbd

the missing availability of freely distributable content prohibited the publication of the database.

It should be noted that both sources of databases may be used for training and verification of newly created algorithms while independent validation is prohibited because the databases have been published. Typical subjective experiment databases contain 100 to 200 degraded video sequences in order to avoid observer fatigue when evaluating the perceived quality. Therefore, training and verifying the performance of an algorithm requires the combination of several databases, which has been tackled in the literature with common sequences in all sets [15], [16]. The combination of datasets without a common set of sequences introduces additional fitting parameters into the training procedure and poses problems when the perceptual scales of the databases differ significantly, notably in the case of different types of degradations or large differences in the evaluated quality range.

4. Large Database - Creation and Properties

The flowchart of an optimal setup for the development of objective algorithms is depicted in Fig. 1. It focuses notably on reproducibility. A source video sequence taken from a freely accessible database is encoded using a parameter set that is entirely stored in a database. Packet losses are introduced based on stored packet loss patterns that may either stem from models or from measured network data. A robust decoder simulation guarantees that packet losses have the same effect for any source video sequence, encoder parameter, and packet loss. This is achieved by removing information that would not be present at the decoding side due to packet loss inside the reference decoder rather than treating missing packets at the entrance to the decoder. In order to facilitate model development, bitstream information of any supported video coding standard is parsed and output in a common XML format. The decoded videos undergo objective measurement using FR measurements, but also subjective assessment may be conducted on parts of the sequences. Several quality indicators, both bitstream based or derived from the decoded video, are calculated and fused in an algorithm with respect to the known scope of the configuration used to create the sequence. Finally, a video quality estimation algorithm can be created, improved, and recommended in different versions, similar to the continuous advancements seen in the video coding community. The following subsections will briefly detail the process and provide an estimation of the size of such a database.

4.1 Content

A large variety of content is required featuring different properties such as natural video sequences containing slow and fast moving objects with or without camera movement, cartoons or screen casts containing unnaturally sharp edges [17]. The initial quality of the natural content should reflect

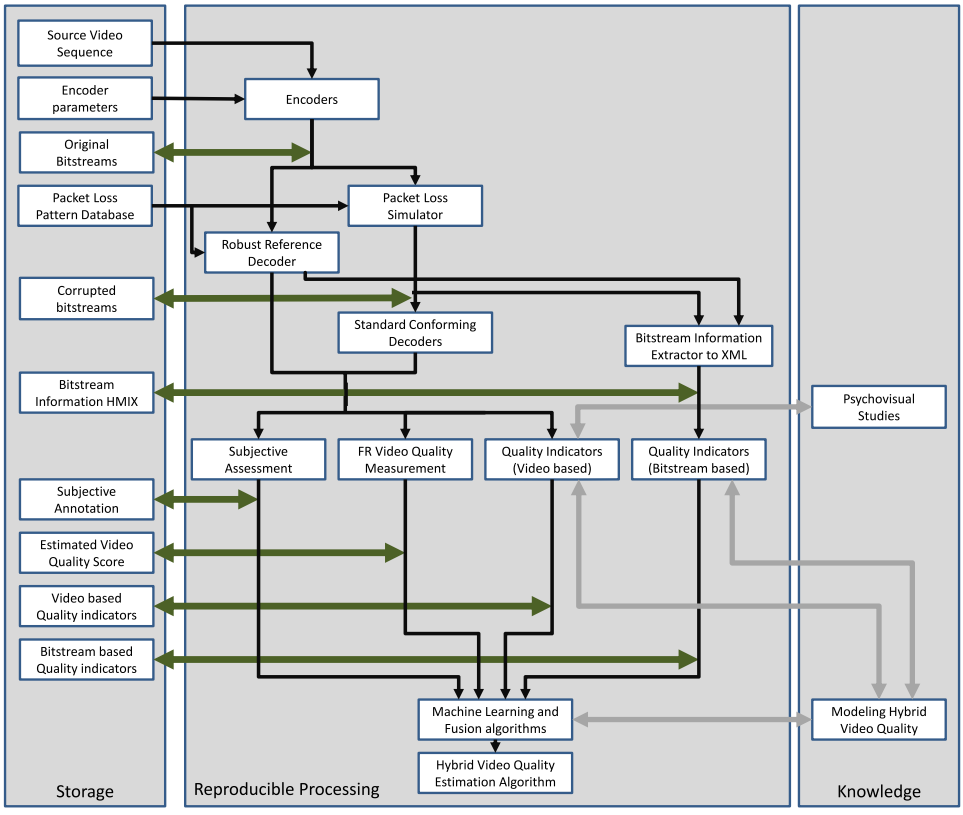

Fig. 1 Processing steps for a large database creation towards development of a reliable Hybrid Model.

the conditions of usage and range from electronic cinema productions and professional television shootings to consumer produced videos with mobile phones. Special scenarios should be taken into consideration such as cloud gaming applications or remote desktop applications. A suitable input database may be found in the Consumer Digital Video Library (www.cdvl.org) which provides a large collection of freely available content for research purposes.

4.2 Video Coding Standards

When considering video quality assessment algorithms, the scope of the algorithm is generally restricted to a certain video compression standard or a family of standards. The most well-known compression standards are the ones codeveloped by ITU-T and ISO/IEC, namely H.262/MPEG-2, H.264/AVC [18], and H.265/HEVC [19]. While the MPEG-2 standard is still widely used because it enabled the entire broadcasting industry to go digital, it gets more and more replaced by its successors, namely H.264 and HEVC. Because of the decoding complexity involved with H.264/AVC decoding, a lot of devices got support for hardware accelerated decoding for this codec resulting in wide acceptance of this standard up to now. Individual standardization organization or individual companies have also been working on compression technology and the result of their effort can be found in compression algorithms like VC-1 (SMPTE 421M), MPEG-4 visual, ITU-T H.263, and Google VP9.

4.3 Coding Parameters

Recent video coding standards are highly configurable allowing for a wide variety of applications, ranging from low resolution previews to high fidelity reproduction in home cinemas or even lossless reproduction. Higher resolutions come with higher decoding requirements and will therefore exclude slower devices from the application scope.

While the resolution determines the principal range of bitrates for an application, the most effective encoder's rate distortion control parameter is the Quantization Parameter (QP). How this QP is controlled depends on the chosen bitrate strategy. For optimal compression and a constant visual quality, a constant QP parameter is preferred. As fast moving or high detailed complex parts of the video will require a lot more bitrate to get compressed at constant QP, the term Variable BitRate coding (VBR) is assigned to this type of video streams. On the contrary, Constant BitRate (CBR) coding limits variations in bitrate over time. A videoconferencing application is a typical example in which there is no time to buffer a lot of the video stream and the video should closely match the available network bandwidth.

Every application requires a different random access and error robustness strategy, which can be controlled using the Random Access Period (RAP). This RAP is usually implemented by special types of intra frames, where prediction between consecutive frames is stopped which enables a decoder to start decoding the video stream from that frame on. The random access property can also be considered a robustness property because an error that has been introduced in the video stream is stopped by these intra frames. As an alternative to inserting an intra frame, a moving set of intra blocks that swipes over the screen, for example from left to right, can be used for random access purposes. This is called intra refresh, and using this technique, the high cost of intra blocks is spread over different frames, simplifying a CBR strategy. Visually, intra frames or intra blocks may annoy the observer if the QP is not chosen carefully.

The error robustness aspect can further be improved by the concept of slices. Slices divide the frame in parts which can be decoded independently. When a network packet belonging to a certain slice gets lost, the entire slice cannot be decoded anymore, but the other slices of the same frame still can. Additionally, slices are also used in ultra-low latency applications in which it is necessary to send a portion of the frame over the network before the entire frame is encoded. Notably in case of packet losses, slices may still enable intelligibility of the content even if the video quality itself is strongly degraded.

In recent standards like H.264/AVC or H.265/HEVC, frames can predict from any previous frame allowing hierarchical referencing structures and enabling temporal scalability. With such a structure, frames can be removed on the fly, reducing the frame rate of the sequence without any drift effects on succeeding frames but reducing significantly the QoE.

4.4 Transmission Influence

During the transmission of video over packet-based best-effort IP networks (e.g. the Internet), network impairments, such as packet loss, can also deteriorate end-users' QoE. As a general term, Quality of Service (QoS), is used to denote the quality of the (delivery) network and is typically measured in terms of bandwidth, packet loss, delay, and jitter. A lot of research has already been conducted towards mapping QoS measurements to QoE prediction [20]–[23].

Based on network monitoring and video packet analysis, QoS measurements can be combined with information on the video encoding and video content [24]. Even further, in the case of deep packet inspection (DPI), detailed video information up to the level of macroblocks, motion vectors, and quantization coefficients can be even used to further improve quality estimation [25], [26]. This more in depth analysis is needed as the packet loss rate is not enough to reliably estimate QoE. For example, the impact of packet loss will be much smaller when the losses occur in B-pictures compared to losses occurring in I-pictures [27], [28]. This is similar in the case of random or bursty losses [29].

Depending on the transport mechanism used to deliver the video content to the end-user, network impairments can result in different kinds of perceivable distortions in the video. For example, in the case of real-time video streaming

using the Real-time Transmission Protocol (RTP)[†], network impairments will most likely result in an irrevocable loss of information. Hence, the original video stream cannot be entirely reconstructed. This will be perceived by the end-users as, what is called, slicing errors or random block patterns [30]. When using more reliable transport mechanisms such as progressive download or HTTP Adaptive Streaming (HAS)[††], lost video packets are automatically retransmitted. Still, in this situation, severe network impairments can result in video packets being delivered out of time. In this case, video playback will be interrupted resulting in a video stalling [31]–[33]. It is clear that this results in different kinds of visual degradations to the end-users [34].

The result of network monitoring can be stored in the form of packet traces which allow simulating, in a reproducible way, the impairments caused on the compressed video bitstream. Trace repositories are available for researchers to simulate different environments [35], [36]. Traces typically include information about lost packets and packet arrival time, the latter is particularly useful to simulate different playout deadlines applied at the receiver by discarding packets that would be too late for decoding.

While many repeated captures from real network may cover quite different set of conditions in terms of packet loss and delay, models have also been developed to provide better flexibility in recreating particular situations on-demand. Models can range from the very common 2-state Markov chain [37], especially useful to simulate consecutive packet losses, to more complicated ones such as hierarchical models [38]. Clearly, captured traces have the great advantage of being very realistic, while models can be fine-tuned to cover the test conditions in the best possible way, not mentioning that an arbitrary number of channel realizations can be generated by using them.

Regardless of the method chosen to recreate the network conditions under test, when packet losses are present in a video, decoding software robust to the corruption of the compressed bitstream, should be employed. Ideally, when a data loss is encountered, the decoder should be able to recover as fast as possible by resynchronizing itself with the compressed data, so that the amount of wasted data due to the error is minimized, hence the distortion in the reconstructed video sequence. Unfortunately, video decoding software often crash in this condition, particularly when the amount of lost data is large or affect consecutive elements. Complex modifications are typically needed to make the software robust to any loss pattern. These are only available in commercial products, which are undesirable since inhibit reproducible research.

However, a generic publicly available decoder such as the standard reference one [39] can be made robust if the internal state is not modified, apart from the content of the decoded picture buffer (DPB). This can be achieved by always processing the original, uncorrupted, bitstream and simulating loss events (i.e., dropped NAL units) through the application of the concealment technique within the decoder itself. This procedure is able to reproduce concealment artifacts on both the current and subsequent frames, and it can handle any loss pattern. Therefore, this could be used as a reference decoder for the case of corrupted bitstreams (note that how to handle this situation is out of the scope of the normative parts of the standard). For completeness, it should be noted that in very few cases there might be a slight misalignment between the reference decoder and an actual decoder due to few, particular encoding modes that form predictions on the basis of data that should be considered not available by the reference decoder.

The parameter space for network impairments comprises of the network protocols, the packet loss pattern, the retransmission scheme, etc. It is therefore huge and currently available databases are scarcely evaluating more than one of the dimensions and seldom with more than a few samples. A reproducible large scale database may however reproduce any condition at any time.

4.5 Typical Application Scope Examples

Application scopes often limit the combinations of coding and network parameters, here they will be clustered depending on their latency constraints.

On one side of the spectrum, there are applications with large latencies like Video on Demand (VoD) which tend to use HAS nowadays, dividing the video in independently decodable segments which are requested one by one using the HTTP protocol. Except for the RAP present at the start of every segment, no additional random access or error robustness features should be applied to the video stream because transmission takes place over a reliable TCP connection. Encoding of the entire segment needs to be finished before it can be made available to the client, therefore a lot of frames can be buffered and a predictive structure using a lot of bi-predicted frames can be applied. Additionally, a constant bitrate is only considered on segment level such that more efficient VBR coding can be used within the segment. Consequently, with HAS used for VoD, because the network takes care of error resilience and because latency is not a big issue, the codec can be configured in an optimal way.

When reducing the end-to-end delay between source and receiver from several seconds to several milliseconds, applications like video conferencing and remote desktop appear. With videoconferencing, the video stream generally gets multiplexed and transported by the RTP protocol. For obtaining such low delays, complex prediction structures using bi-predicted frames cannot be applied anymore. Additionally, with lower latencies it becomes more and more important to maintain a constant bitrate. With the remote desktop application, the source content has very different statistics compared with natural content and therefore other considerations should be made. For example,

[†]RTP is delivered using the unreliable User Datagram Protocol (UDP).

[††]In the case of progressive download and HAS, video is delivered using the Transmission Control Protocol (TCP).

where 4:2:0 chroma subsampling is acceptable for natural images, screen content gets deformed noticeably by such color downscaling. Additionally, compression tools [40] have been developed for such content, so even more options become available for the codec developer.

A large database collection could be separated into partially overlapping segments that allow to evaluate the performance of video quality measurement algorithms with respect to a certain application.

4.6 Estimated Size

Although in the previous section the different parameters got strictly clustered around the different applications depending on latency, in reality other parameter combinations are possible as well. Therefore, to estimate the ideal size of a possible large scale database these strict clusters should be avoided. As a minimal size of such database consider at least six different resolutions [41] for sizes up to HD plus 4K resolutions. Six bitrate points for every resolution leads to 42 combinations. As discussed earlier, bitrate can be distributed through a VBR or a CBR strategy, the latter one being employed on block, slice, frame, or segment basis. Allowing these five variations gives 210 possible encoding configurations. The amount of RAPs depends on the application, but in any application either the random access feature or the error robustness feature of RAPs is useful. In error prone environments, an intra frame every eight frames is reasonable and in practice this amount is lowered until one RAP every 10 seconds for HAS. In all these cases, three types of RAPs can be used, namely open-GOP, closed-GOP, or intra refresh. Taking a limited set of 10 RAP-period values combined with these three types of RAPs results in 6300 possible configurations. For slicing the frame, the two main uses for slicing should be considered, namely parallel encoding and robust transmission. For parallel encoding, from our own experience it can be stated that up to around eight separate software threads can still work with insignificant overhead for threading, so up to eight slices can be considered. For error robustness purposes, 1500 byte slices should be added to the parameter space resulting in 56700 configurations. Finally, the most difficult parameter to cover a large scope of encoders and application scenarios is prediction structure. Encoder implementations range from static prediction structures and reference frame settings to dynamic structures with intelligently chosen reference frames depending on the source content. For each prediction structure type (static or dynamic), four different ratios of B-frames to the number of P-frames should at least be considered to be able to cover the encoders in the market, bringing the total set of configurations up to 453600. With such a set, not the entire encoder market is covered, but this number should make clear that it is a big undertaking to evaluate or create a quality metric able to cope with the large scope of variability present in the encoder market. It should be added that this number only considers the encoder and does not talk about the impact of the network, the decoder or the display device. Please also note that this would apply for a single video content. It becomes evident that such a database will probably never exist as a video sequence collection, it requires too much calculation time and disk space.

4.7 Evaluating Perceived Quality

Evaluating such a large database subjectively is not feasible. Research needs to show whether objective algorithms may be employed in order to obtain a selection of database points that may be estimated with sufficient precision. Many FR algorithms exist, ranging from low complexity such as PSNR, SSIM, to higher complexity algorithms such as VQMT, VIF, VQM or PVQM [42]–[46]. Their application requires processing power but allows for measuring their agreement in order to obtain information about their prediction performance. This assumes that agreement between fundamentally different approaches of objectively measuring perceived quality may indicate their reliability in predicting subjective quality. The remaining need to undergo subjective assessment but iterative re-evaluation after improvement of the objective algorithms may be applied. First work in this direction has already been started [47].

4.8 Ongoing Large Scale Database Efforts

Within the VQEG-JEG, two databases have been made publicly available which were also annotated with objective scores[†]. One database contains 12960 objectively annotated H.264 sequences [48], the other has been designed for HEVC evaluations with the parameter selection provided in Table 2. The database consists of 5952 different encoding configurations per sequence. This is a subset of the described parameters in Sect. 4.6, selected because of processing limitations. This subset contains three different resolutions, at four VBR rates and 12 CBR settings. Additionally, four prediction structures, four slicing variations, four intra periods, and only open-GOP or closed-GOP RAPs are considered as shown in Table 2. With a limited set of 10 source sequences, 59520 encoded video streams have been generated taking approximately 35.7 computing years to generate.

Table 2 HEVC compression parameters.

Parameter	Values
Rate control algorithm	VBR: Fixed QP=26, 32, 38, 46 CBR at frame level: rate=0.5, 1, 2, 4, 8, 16 Mbps CBR at CTU level: rate=0.5, 1, 2, 4, 8, 16 Mbps
Random access	Closed-GOP intra refresh (IDR), Open-GOP intra refresh (CRA)
Intra period	8, 16, 32, 64
Resolution	1920x1080, 1280x720, 960x544
Slices	Count: 1, 2, 4; Size: 1500 byte
GOP structure and size	GOP 1 (IPPPPPPPPP), GOP 2 (IBPBPBPBP) GOP 4 (IBBBPBBBP), GOP 8 (IBBBBBBBP)

[†]see ftp://ftp.ivc.polytech.univ-nantes.fr/VQEG/JEG/HYBRID

5. Exemplary Analysis Results for Full Reference Quality Measurements

This section shows some examples of the results that can be achieved by analyzing the quality metrics already in the database. First, scatter plots are used to visualize the correlation between them. For instance, Fig. 2 shows the VQM value as a function of the corresponding PSNR value, for all the sequences in the database. Each sequence is characterized by its own color. Several considerations can be done by observing the figures, however we focus our attention on some characteristics that we deem important from the point of view of the database usefulness and future developments.

Some points have fairly low PSNR values (hence low quality according to PSNR) and at the same time quite good quality for VQM (i.e., a low value): such sample points are highlighted in the figure by an arrow. These situations of strongly contrasting metrics are important for the database since they might be an indication of the need of further investigation, for instance in terms of subjective quality experiments. Another observation shows that for some sequences, such as *seq08*, values sharply rise when the PSNR variation is relatively low (points are highlighted by the black ellipse in the figure). This again indicates that the sequence as a whole has some peculiar characteristics which make it different from the others: VQM has large variations which are not observed in the PSNR domain.

Figure 3 directly compares the VQM and PVQM measures, showing that there is a good correlation for certain conditions, shown by the high point density arranged into a straight line shape at the bottom of the figure, but a very weak correlation for other conditions. Unfortunately, due to the high number of points, it is not easy to identify a pattern which relates well with certain values of the coding parameters.

However, in the following we attempt such an analysis showing that some trends can be highlighted. The objective is to measure how much the change of a given coding parameter (such as the GOP size and the video resolution) affect each measure. The analysis is based on measuring the standard deviation of the metric of interest on a set of data obtained as the average of all the measures which share the same coding parameters, except the one under test and the rate control algorithm. The latter one is treated independently since it has an obvious influence on the quality regardless of the employed measure. To achieve consistency between the metrics, first we normalize the values in the range 0 to 1 using a standard linear mapping. For each metric, all the measures in the database have been considered for the normalization operation.

Due to lack of space, only some sample results are shown, drawing them from the most interesting cases. Nevertheless, we believe that this type of results can effectively demonstrate the utility of having such a large database to investigate the peculiar behaviors of video quality metrics in some particular conditions. Figure 4 shows the dependency

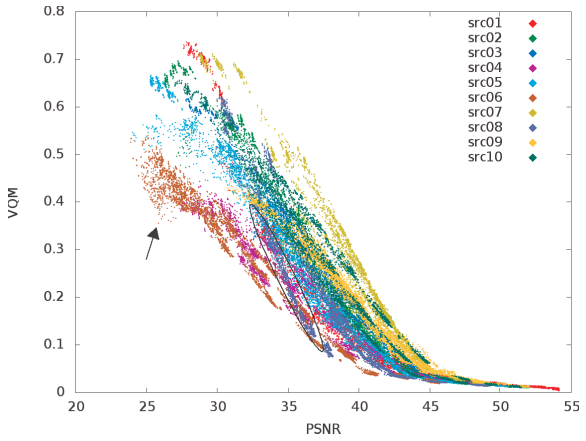

Fig. 2 Scatter plot of the VQM value versus the PSNR value for all the sequences in the database.

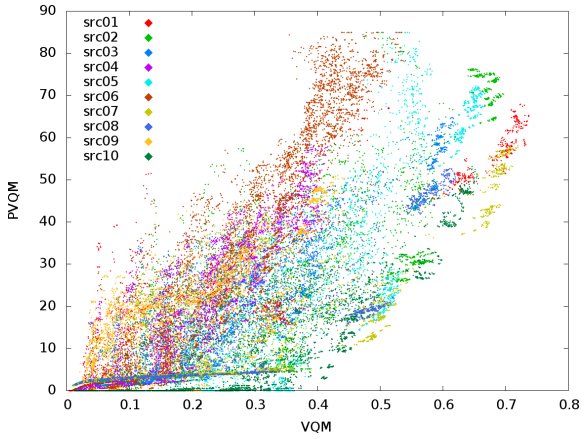

Fig. 3 Scatter plot of the PVQM value versus the VQM value for all the sequences in the database.

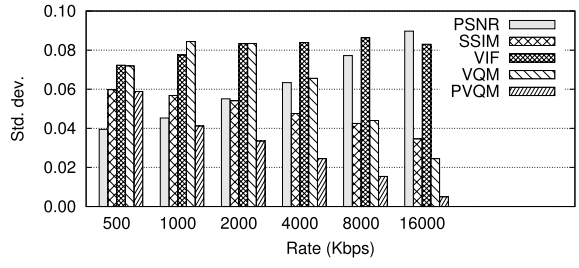

Fig. 4 Standard deviation of the quality measures when different resolution values are considered.

of the metrics on the resolution, for the case of rate control with bitrate allocated at the frame level. As the bitrate increases, the behavior of the metrics change depending on the metric itself. For instance, the PSNR tends to slowly increase. Others such as VIF are not significantly influenced by the bitrate variation. Other results considered in Fig. 5 show that when the GOP size parameter is considered, the situation varies. In particular, the absolute value of the standard deviation is much lower than the previous case, and the drop effect is already visible at 8 Mbps. Finally, to better visualize the previous data, Fig. 6 shows a subset of the values

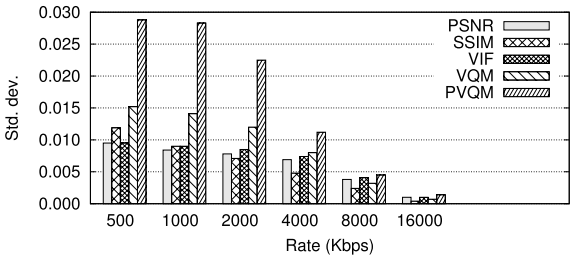

Fig. 5 Standard deviation of the quality measures when different GOP size values are considered.

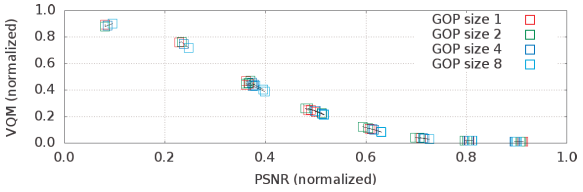

Fig. 6 Average values of the VQM and PSNR metrics when different GOP sizes are considered. All points with the same rate control configuration are connected by a line. Sequence *src02*.

for sequence *src02*, resolution equal to 960x544 and for the PSNR and VQM metric. The points represent the average value of the metrics for the same conditions except the GOP size and the rate control parameter. It is possible to notice that, depending on the rate control parameter, the points are closer or more distant, and the amount of variation of the two metrics is different.

6. Future Research Domains

The main motivation of this large database effort is to deepen the knowledge concerning the interaction between the human perception of video presentations and the bandwidth reserved for video transmissions. It is obvious that many factors of QoE are not yet taken into account, such as immersion or viewing comfort. Nonetheless, many questions may be tackled from a different view point when changing the perspective from a well designed but isolated subjective experiment to an exhaustive (or at least large) dataset of conditions. A few examples will be given here.

First, the development of objective measurement algorithms can be more rigorously structured. Indicators of perceived degradations are nowadays often only tested within their scope of application, i.e. when changing the perceived degradation in a controlled way, but seldom for their behavior when the degradation is not present or constant. Verification can be performed in a reproducible way, remaining inaccuracies can be signaled and discussed on subsets of the communicated large database. Besides the scope question, the linearity of the indicators response may be improved. The analysis of a large scale database using objective metrics will also give an indication about which features contribute most to the compression efficiency of a video codec. Until now, these features have mainly been evaluated on small sets of source sequences under a restricted set of encoder parameter combinations and only using PSNR. Second, the combination of indicators can be tackled independently from their design and improvement. Machine learning techniques and verification techniques may be applied and improved for the specific problem of the combination of indicators that may behave nonlinearly or incorrectly when used out of scope and that may provide partially overlapping responses when used within scope. Third, correlation and accuracy analysis may be refined when several different measurements are compared. Knowing when a measurement significantly outperforms another in terms of prediction performance is a crucial information. Currently only very basic statistic tools can be used due to the low number of subjectively evaluated video samples. Fourth, by including impairment effects of the network, this database will be able to reveal valuable information about actual error robustness of a combination of features. Up until now, these robustness features are selected by experts based on common knowledge, but large scale video databases could give scientific indication of well tuned robustness under a wide variety of circumstances. Using the results that will come out of this large database research, best practices for adaptive streaming can be formulated. These best practices will on their turn result in improved encoder settings for such servers and improved segment selection procedures for the adaptive streaming clients in such environments. Fifth, it may be considered that the database is large enough such that the validation may be performed on a subset of the database that was not used for training or even verification may be sufficient if the training and verification is performed on several million video sequences.

7. Conclusion

This work presented a large scale, reproducible and easy to use database aimed at advancing the current research efforts in the development of new video quality assessment algorithms. Several well-known quality measurement outputs are available at the frame-level granularity to enable researchers to perform statistical analysis of video quality measurement algorithms. While this is an ongoing effort as the database is continuously updated and enlarged, some interesting phenomena, which cannot be observed on smaller scale databases, can already be noticed as shown in the sample results section. The Joint Effort Group of the Video Quality Experts Group (VQEG-JEG) hopes that this will be the seed for significant advances in the development of new, innovative quality metrics.

Acknowledgments

The research activities described in this paper were partially funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union. Some aspects of this work were carried out using the STEVIN Supercomputer Infrastructure at Ghent University.

References

[1] F. Campbell, R. Carpenter, and J. Levinson, "Visibility of aperiodic patterns compared with that of sinusoidal gratings," The Journal of Physiology, vol.204, pp.283–298, 1969.

[2] A.B. Watson and J. Ahumada, "A standard model for foveal detection of spatial contrast," Journal of Vision, vol.5, pp.717–740, 2006.

[3] Q. Huynh-Thu and M. Ghanbari, "Impact of jitter and jerkiness on perceived video quality," Proc. Workshop on Video Processing and Quality Metrics, 2006.

[4] J. Li, M. Barkowsky, and P. Le Callet, "The influence of relative disparity and planar motion velocity on visual discomfort of stereoscopic videos," International Workshop on Quality of Multimedia Experience, pp.155–160, 2011.

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

[6] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, pp.600–612, 2004.

[7] R. Dosselmann and X. Yang, "A comprehensive assessment of the structural similarity index," Signal, Image and Video Processing, vol.5, pp.81–91, 2011.

[8] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," IEEE Trans. Broadcast., vol.57, pp.165–182, 2011.

[9] M. Barkowsky, I. Sedano, K. Brunnström, M. Leszczuk, and N. Staelens, "Hybrid video quality prediction: Reviewing video quality measurement for widening application scope," Multimedia Tools and Applications, pp.1–21, 2014.

[10] K. Fliegel and C. Timmerer, eds., "WG4 Databases White Paper v1.5: QUALINET Multimedia Database enabling QoE Evaluations and Benchmarking," http://dbq-wiki.multimediatech.cz, March 2013.

[11] I.T.U. ITU-T, "Objective perceptual assessment of video quality: Full reference television," Technical papers and tutorials, 2004.

[12] I.S.G. 9, "Final Report of VQEG's Multimedia Phase I Validation Test," TD 923, 2008.

[13] G. Cermak, L. Thorpe, and M. Pinson, "Test plan for evaluation of video quality models for use with high definition TV content," Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/media/5871/vqeg_hdtv_testplan_v3_1.doc, 2009.

[14] J. Berger, C. Lee, D. Hands, N. Staelens, Y. Dhondt, and M. Pinson, eds., "Hybrid perceptual bitstream test plan 2.11," Video Quality Experts Group (VQEG), http://www.its.bldrdoc.gov/media/36068/VQEG_hybrid_testplan_v2_11.doc, 2012.

[15] M. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," SPIE Video Communications and Image Processing Conference, pp.8–11, 2003.

[16] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet, "Aligning subjective tests using a low cost common set," QoE for Multimedia Content Sharing, Lisbonne, Portugal, 2011.

[17] M. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," EURASIP Journal on Image and Video Processing, vol.2013, p.1, 2013.

[18] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," IEEE Trans. Circuits Syst. Video Technol., vol.17, no.9, pp.1103–1120, Sept. 2007.

[19] J.-R. Ohm and G.J. Sullivan, "High efficiency video coding: The next frontier in video compression," IEEE Signal Process. Mag., vol.30, no.1, pp.152–158, Jan. 2013.

[20] F. Battisti, M. Carli, and A. Neri, "No reference quality assessment for MPEG video delivery over IP," EURASIP Journal on Image and Video Processing, vol.2014, no.1, pp.1–19, 2014.

[21] J. Song and F. Yang, "Real-time quality monitoring for networked H.264/AVC video streaming," Proc. 3rd International Conference on Multimedia Technology (ICMT 2013), ed. A.A. Farag, J. Yang, and F. Jiao, Lecture Notes in Electrical Engineering, vol.278, pp.237–245, Springer Berlin Heidelberg, 2014.

[22] M. Alreshoodi and J. Woods, "Survey on QoE\QoS correlation models for multimedia services," International Journal of Distributed and Parallel Systems (IJDPS), vol.4, no.3, pp.53–72, May 2013.

[23] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," IEEE Trans. Circuits Syst. Video Technol., vol.23, no.8, pp.1322–1333, Aug. 2013.

[24] J. Joskowicz, R. Sotelo, and J. Lopez Arado, "Comparison of parametric models for video quality estimation: Towards a general model," IEEE International Symposium onBroadband Multimedia Systems and Broadcasting (BMSB), pp.1–7, June 2012.

[25] S. Argyropoulos, M.N. Garcia, M. Salem, D. List, R. Schleicher, and A. Raake, "Objective no-reference prediction of saliency changes in the presence of packet loss," Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-13), pp.6–11, Jan. 2013.

[26] N. Staelens, G. Van Wallendael, K. Crombecq, N. Vercammen, J. De Cock, B. Vermeulen, R. Van de Walle, T. Dhaene, and P. Demeester, "No-reference bitstream-based visual quality impairment detection for high definition H.264/AVC encoded video sequences," IEEE Trans. Broadcast., vol.58, no.2, pp.187–199, June 2012.

[27] A. Reibman, S. Kanumuri, V. Vaishampayan, and P. Cosman, "Visibility of individual packet losses in MPEG-2 video," International Conference on Image Processing (ICIP), pp.171–174, Oct. 2004.

[28] A. Reibman, V. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," IEEE Trans. Multimed., vol.6, no.2, pp.327–334, April 2004.

[29] Y. Liang, J. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.5, pp.684–687, April 2003.

[30] M.N. Garcia, S. Argyropoulos, N. Staelens, M. Naccari, M. Rios-Quintero, and A. Raake, "Video streaming: Advanced concepts, applications and methods," in Quality of Experience, ed. S. Möller and A. Raake, T-Labs Series in Telecommunication Services, pp.277–297, Springer International Publishing, 2014.

[31] P. Ameigeiras, A. Azcona-Rivas, J. Navarro-Ortiz, J. Ramos-Munoz, and J. Lopez-Soler, "A simple model for predicting the number and duration of rebuffering events for YouTube flows," IEEE Commun. Lett., vol.16, no.2, pp.278–280, Feb. 2012.

[32] A. Gouta, C. Hong, D. Hong, A.M. Kermarrec, and Y. Lelouedec, "Large scale analysis of HTTP adaptive streaming in mobile networks," IEEE 14th International Symposium and Workshops on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp.1–10, June 2013.

[33] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, "To pool or not to pool": A comparison of temporal pooling methods for HTTP adaptive video streaming," Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp.52–57, July 2013.

[34] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of IPTV and Video on Demand services in real-life environments," IEEE Trans. Broadcast., vol.56, no.4, pp.458–466, Dec. 2010.

[35] CRAWDAD project, "A community resource for archiving wireless data at Dartmouth," http://crawdad.cs.dartmouth.edu, 2014.

[36] M. Ellis, C. Perkins, and D.P. Pezaros, "End-to-end and network-internal measurements on real-time traffic to residential users," Proc. ACM Multimedia Systems, pp.111–116, San Jose, CA, USA, Feb.

2011.
[37] E. Elliot, "Estimates on error rates for codes on burst-noise channels," Bell Syst. Tech. J., vol.42, pp.1977–1997, Sept. 1963.
[38] S.A. Khayam and H. Radha, "Markov-based modeling of wireless local area networks," Proc. ACM Intl. Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems (MSWIN), pp.100–107, 2003.
[39] JCT-VC, "HEVC test model (HM) reference software, v. 12.1," Nov. 2013.
[40] G.J. Sullivan, J.M. Boyce, Y. Chen, J.-R. Ohm, C.A. Segall, and A. Vetro, "Standardized extensions of High Efficiency Video Coding (HEVC)," IEEE J. Sel. Top. Signal Process., vol.7, pp.1001–1016, Dec. 2013.
[41] A. Inc., "Best Practices for Creating and Deploying HTTP Live Streaming Media for the iPhone and iPad," TN2224, April 2014.
[42] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol.13, no.4, pp.600–612, April 2004.
[43] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," IEEE Trans. Image Process., vol.15, no.2, pp.430–444, Feb. 2006.
[44] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans. Broadcast., vol.50, no.3, pp.312–322, Sept. 2004.
[45] A.P. Hekstraa, J.G. Beerends, D. Ledermann, F.E. de Caluwe, S. Kohler, R.H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss, "PVQM — A perceptual video quality measure," Signal Processing: Image Communication, vol.17, no.10, pp.781–798, Nov. 2002.
[46] P. Hanhart and R. Hahling, "Video quality measurement tool (VQMT)," http://mmspg.epfl.ch/vqmt, Sept. 2013.
[47] A. Reibman, "A strategy to jointly test image quality estimators subjectively," Image Processing (ICIP), 2012 19th IEEE International Conference on, pp.1501–1504, 2012.
[48] M. Leszczuk, L. Janowski, and M. Barkowsky, "Freely available large-scale video quality assessment database in full-HD resolution with H.264 coding," Proc. IEEE Globecom 2013, p.1, Atlanta, États-Unis, 2013.

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 on 3D displays using psychometric and medical measurements. He currently co-chairs the VQEG "3DTV" and "Joint Effort Group Hybrid" activities.

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

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

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

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

Patrick Le Callet is currently a Full Professor at Ecole Polytechnique de l'Université de Nantes. He has been teaching as an Assistant Professor from 1997 to 1999 and as a full-time Lecturer from 1999 to 2003 in the Department of Electrical Engineering, Technical Institute of University of Nantes (IUT). Since 2003, he has been teaching at Ecole Polytechnique de l'Université de Nantes (Engineering School) in the Electrical Engineering and the Computer Science Departments. In 1997, he joined the Image and Video Communication group at CNRS IRCCyN. Since 2006, he has been the head of this group that includes ten permanent professors, two assistant professors, 18 postdoctoral and Ph.D. students, and five research engineers. His research focuses on better understanding of the human visual system and designing and applying HVS models in image and video processing. Current topics of interest are DTV, image, and video quality assessment, watermarking techniques, and visual attention modeling and applications. He is co-author of more than 70 publications/communications an co-recipient of six international patents.