The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics

Stefan Winkler and Praveen Mohandas

***Abstract*—This paper reviews the evolution of video quality measurement techniques and their current state of the art. We start with subjective experiments and then discuss the various types of objective metrics and their uses. We also introduce V-Factor, a "hybrid" metric using both transport- and bitstream information. Finally, we summarize the main standardization activities, such as the work of the Video Quality Experts Group (VQEG), and we take a look at emerging trends in quality measurement, including image preference, visual attention, and audiovisual quality.**

***Index Terms*—Audiovisual quality, objective metrics, standards, subjective assessment, video quality.**

I. Introduction

QUALITY OF experience (QoE) has become a term commonly used to describe the application- and user-oriented quality of video and multimedia services. QoE actually encompasses many different aspects—video quality is just one of them, arguably one of the most important [1].

Unfortunately, quality in this context is a rather ill-defined concept—we list just some of the numerous factors contributing to QoE here [2]–[4]:

- Individual interests of the viewer, such as favorite programs, which determine the level and focus of attention;
- Quality expectations of the viewer, for example a feature film screened in a cinema vs. a short clip watched on a mobile device;
- Video experience of the viewer, which also determines quality expectations (once you have seen high-definition content it's hard to go back);
- Display type (CRT, LCD, etc.) and properties (size, resolution, brightness, contrast, color, response time);
- Viewing setup and conditions, such as viewing distance or ambient/exterior light;
- Quality and synchronization of the accompanying audio;
- Interaction with the service or display device (e.g. zap time, remote control, electronic program guide).

As the wide variety and subjectivity of some of these factors indicate, the measurement (and ultimately optimization) of the quality of digital video systems is a highly complex problem. Most of today's quality metrics only account for a small subset of the factors listed above and focus on measuring the visual fidelity of the video in terms of the distortions introduced by various processing steps (mainly compression and transmission). Even if we constrain ourselves to this more well-defined problem space, two challenging issues remain:

- Video systems are complex and consist of many components, including capture and display hardware, converters, multiplexers, codecs, streamers, routers, switches. All of them process the video in some way, which can potentially affect its quality.
- Visual perception is even more complex. If we are to measure quality in a meaningful way, we need to understand how people perceive video and its quality.

These two issues and metrics addressing them are also the focus of this review.

The paper is organized as follows. Section II briefly introduces subjective quality assessment, which forms the benchmark for objective metrics. Section III discusses objective quality metrics, various classifications and some specific implementations. Section IV introduces V-Factor as an example of a hybrid metric. Section V reviews standardization activities related to video quality. Section VI takes a look at some recent trends in quality measurement, and Section VII concludes the paper.

II. Subjective Quality Assessment

The reference for multimedia quality are subjective experiments, which represent the most accurate method for obtaining quality ratings. In subjective experiments, a number of "subjects" (typically 15–30) are asked to watch a set of video clips and rate their quality. The average rating over all viewers for a given clip is also known as the Mean Opinion Score (MOS).

Since each individual has different interests and expectations for video, the subjectivity and variability of the viewer ratings cannot be completely eliminated. Subjective experiments attempt to minimize these factors through precise instructions, training and controlled environments. Yet it is important to remember that a quality score is a noisy measurement that is defined by a statistical distribution rather than an exact number.

There are a wide variety of subjective testing methods. Psychophysics provides the tools for measuring the perceptual performance of subjects [5], beginning with visibility thresholds and just-noticeable differences (JND's), which are most suitable for small impairments. The ITU has formalized direct scaling methods in various recommendations [6]–[8], which are often used in practice for larger quality ranges. They suggest standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods. Recommended testing procedures include

Manuscript received November 27, 2007; revised April 9, 2008. First published June 27, 2008; last published August 20, 2008 (projected).

The authors are with Symmetricom, San Jose, CA 95131 USA (e-mail: swinkler@symmetricom.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2008.2000733

implicit comparisons such as Double Stimulus Continuous Quality Scale (DSCQS), explicit comparisons such as Double Stimulus Impairment Scale (DSIS), or absolute ratings such as Single Stimulus Continuous Quality Evaluation (SSCQE) or Absolute Category Rating (ACR). The procedure used for a given experiment is generally selected as a function of the application, the quality range, and the viewer tasks. More details on subjective testing can be found in [9], for example.

Subjective experiments are invaluable tools for assessing multimedia quality. Their main shortcoming is the requirement for a large number of viewers, which limits the amount of video material that can be rated in a reasonable amount of time; they are neither intended nor practical for 24/7 in-service monitoring applications. Nonetheless, subjective experiments remain the benchmark for any objective quality metric.

(a) (b)

Fig. 1. Illustration of the influence of impairment type and image content on the visibility of distortions (see text for details). Both images have identical PSNR, yet their perceived quality is very different.

III. Objective Quality Metrics

Objective quality metrics are algorithms designed to characterize the quality of video and predict viewer MOS. Different types of objective metrics exist [10]. For the analysis of decoded video, we can distinguish *data metrics*, which measure the fidelity of the signal without considering its content, and *picture metrics*, which treat the video data as the visual information that it contains. For compressed video delivery over packet networks, there are also *packet-* or *bitstream-based metrics*, which look at the packet header information and the encoded bitstream directly without fully decoding the video. Furthermore, metrics can be classified into full-reference, no-reference and reduced-reference metrics based on the amount of reference information they require. These classifications are discussed next.

A. Data Metrics

The image and video processing community has long been using mean squared error (MSE) and peak signal-to-noise ratio (PSNR) as fidelity metrics (mathematically, PSNR is just a logarithmic representation of MSE). There are a number of reasons for the popularity of these two metrics. The formulas for computing them are as simple to understand and implement as they are easy and fast to compute. Minimizing MSE is also very well understood from a mathematical point of view. Over the years, video researchers have developed a familiarity with PSNR that allows them to interpret the values immediately. There is probably no other metric as widely recognized as PSNR, which is also due to the lack of alternative standards (cf. Section V).

Despite its popularity, PSNR only has an approximate relationship with the video quality perceived by human observers, simply because it is based on a byte-by-byte comparison of the data without considering what they actually represent. PSNR is completely ignorant to things as basic as pixels and their spatial relationship, or things as complex as the interpretation of images and image differences by the human visual system.

Let's look at the example shown in Fig. 1. Both images have the same PSNR, yet their perceived quality is very different—it is hard to see anything wrong with Fig. 1(a), whereas the distortions are quite obvious in Fig. 1(b). There are two main reasons for this discrepancy, both of which are closely linked to the way the human visual system processes information:

- Data metrics are *distortion-agnostic*. Distortions may be more or less apparent to the viewer depending on their type and properties. The human visual system is not sensitive to the high-frequency noise inserted into the left image. The noise in the right image is a well-localized, lower-frequency noise, whose pattern is much more apparent.
- Data metrics are *content-agnostic*. Viewer perception varies based on the part of the image or video where the distortion occurs. The noise in the left image is contained almost exclusively in the bottom region of the image, where we already have a lot of image activity from the content itself (edges, texture from the rocks and sea). The image activity masks the distortion in this region. The noise in the right image is contained in a region devoid of content activity (the smooth sky). Because little masking is present there, distortions stand out immediately.[1]

Using MSE and various modifications as a basis, a number of additional data metrics have been proposed and evaluated [11]. Although some of these metrics can predict subjective ratings quite successfully for a given compression technique, distortion type or scene content, they are not reliable for evaluations across techniques. MSE was found to be an accurate metric for additive noise, but it is outperformed by vision-based quality metrics for coding artifacts [12].

The network quality of service (QoS) community has equally simple metrics to quantify transmission errors, such as bit error rate (BER) or packet loss rate (PLR). Again, these are relevant for data links, where every bit and packet is equally important, but not for video delivery. The reasons for their popularity are similar to those given for PSNR above. Problems arise when relating these measures to perceived quality; they were designed to characterize data fidelity, but again they do not take into account the content, i.e. the meaning and thus the visual importance of the packets and bits concerned. The same number of lost packets can have drastically different effects on the video content, depending on which parts of the bitstream are affected.

[1]This is not only a spatial phenomenon; masking also occurs with high temporal activity, such as high-motion scenes or scene cuts.

B. Picture Metrics

Due to the problems with simple data metrics outlined above, much effort has been spent on designing better visual quality metrics that specifically account for the effects of distortions and content on perceived quality. The approaches in metric design can be classified in two groups, namely a *vision modeling approach* and an *engineering approach* [13].

The vision modeling approach, as the name implies, is based on modeling various components of the human visual system (HVS). HVS-based metrics try to incorporate aspects of human vision deemed relevant to picture quality, such as color perception, contrast sensitivity and pattern masking, using models and data from psychophysical experiments [14]. Due to their generality, these metrics can in principle be used for a wide variety of video distortions. HVS-based metrics date back to the 1970's and 1980's, when Mannos and Sakrison [15] and Lukas and Budrikis [16] developed the first image and video quality metrics. Later well-known metrics in this category are the Visual Differences Predictor (VDP) by Daly [17], the Sarnoff JND (just noticeable differences) metric by Lubin [18], van den Branden Lambrecht's Moving Picture Quality Metric (MPQM) [19], and the author's own perceptual distortion metric (PDM) [20].

The engineering approach on the other hand is based primarily on the extraction and analysis of certain features or artifacts in the video. These can be either structural image elements such as contours, or specific distortions that are introduced by a particular video processing step, compression technology or transmission link, such as block artifacts. The metrics look at how pronounced these features are in the video to estimate overall quality. This does not necessarily mean that such metrics disregard human vision, as they often consider psychophysical effects as well, but image content and distortion analysis rather than fundamental vision modeling is the conceptual basis for their design.

The engineering approach has gained popularity in recent years. The author's own metrics [21] look for specific spatial and temporal artifacts in the video, such as blockiness, blur or jerkiness, which are then combined into an overall quality prediction. Wang *et al.*'s Structural Similarity (SSIM) index [22] computes the mean, variance and covariance of small patches inside a frame and combines the measurements into a distortion map. Motion estimation is used for a weighting of the SSIM index of each frame in a video. Pinson and Wolf's VQM video quality metric [23] divides sequences into spatio-temporal blocks, and a number of features measuring the amount and orientation of activity in each of these blocks are computed from the spatial luminance gradient. The features extracted from test and reference videos are then compared using a process similar to masking.

C. Packet- and Bitstream-Based Metrics

While a lot of effort in video quality measurement has been devoted to evaluating compression artifacts from decoded "base-band" video, there is also a growing interest in quality metrics specifically designed to measure the impact of network losses on video quality. This development is the result of increasing video service delivery over IP networks, for example Internet streaming or IPTV.

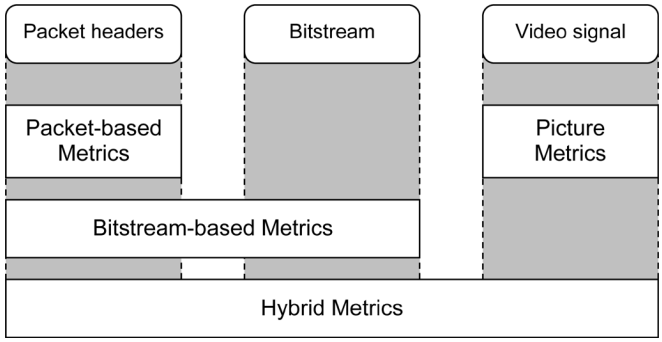

Fig. 2. Classification of packet-based, bitstream-based, picture and hybrid metrics (adapted from ITU-T).

Because losses directly affect the encoded bitstream, such metrics are often based on parameters that can be extracted from the transport stream and the bitstream with no or little decoding. This has the added advantage of much lower data rates and thus lower bandwidth and processing requirements compared to metrics looking at the fully decoded video. Using such metrics, it is thus possible to measure the quality of many video streams/channels in parallel. At the same time, these metrics have to be adapted to specific codecs and network protocols. "Hybrid" metrics use a combination of packet information, bitstream or even decoded video as input. Fig. 2 illustrates the different classes of metrics and their inputs.

Some examples of packet- and bitstream-based metrics are Verscheure *et al.* [24], who investigated the joint impact of packet loss rate and MPEG-2 bitrate on video quality, or Kanumuri *et al.* [25], [26], who used various bitstream parameters such as motion vector length or number of slice losses to predict the visibility of packet losses in MPEG-2 and H.264 video. V-Factor, the metric introduced in Section IV below, also belongs in this category.

D. Reference Information

Quality metrics are generally classified into full-reference, no-reference and reduced-reference categories based on the amount of information required about the reference video [13].

Full-reference (FR) metrics perform a frame-by-frame comparison between a reference video and the test video. They require the entire reference video to be available, usually in unimpaired and uncompressed form, which is quite a heavy restriction on the practical usability of such metrics. Furthermore, full-reference metrics generally impose a precise spatial and temporal alignment of the two videos, so that every pixel in every frame can be matched with its counterpart in the other clip. Temporal registration in particular is quite a strong restriction and can be very difficult to achieve in practice, because of frame drops, repeats, or variable delay introduced by the system under test. Aside from the issue of spatio-temporal alignment, full-reference metrics usually do not respond well to global shifts in brightness, contrast or color, and require a corresponding calibration of the videos. MSE/PSNR and HVS-based metrics typically belong to this class.

No-reference (NR) metrics analyze only the test video, without the need for an explicit reference clip. This makes them much more flexible than FR metrics, as it can be difficult or impossible to get access to the reference in some cases (e.g. video coming out of a camera). They are also completely free from alignment issues. The main challenge of NR metrics lies in telling apart distortions from content, a distinction humans are usually able to make from experience. NR metrics always have to make assumptions about the video content and/or the distortions of interest. With this comes the risk of confusing actual content with distortions (as an example, a chessboard could be interpreted as block artifacts under certain conditions). The majority of NR metrics are based on estimating blockiness [27], which is the most prominent artifact of block-DCT based compression methods such as H.26x, MPEG and their derivatives.

Reduced-reference (RR) metrics are a compromise between FR and NR metrics. They extract a number of features from the reference and/or test video, and the comparison of the two clips is then based only on those features. Examples of features are the amount of motion or spatial detail. This approach makes it possible to avoid some of the assumptions and pitfalls of pure no-reference metrics while keeping the amount of reference information manageable. Reduced-reference metrics also have alignment requirements, but they are typically less stringent than for full-reference metrics, as only the extracted features need to be aligned.

These three classes of metrics also have different operational uses. FR metrics are most suitable for offline video quality measurement such as codec tuning or lab testing, where conditions can be well controlled, and where a detailed and precise analysis of the video is more important than immediate results. NR and RR metrics are better suited for monitoring of in-service video systems, where real-time measurement and alarm triggering are essential. RR metrics still require a back-channel and access to the reference at some point.

IV. V-Factor

We now introduce a real-time video quality metric that uses the transport stream and the bitstream as input. The method does not require a reference and works at the packet level. It combines network impairments with information obtained from the video stream. The algorithm described here focuses mainly on MPEG-2 and H.264 video streaming over IP networks, but it can be adapted to other codecs and other applications such as video conferencing as well.

A compressed video stream can be viewed as a sequence of packets that are carrying video and audio information along with data. De-multiplexing such streams is required in order to identify the packets that carry video information. As an example, assessing video packet loss from IP losses directly will provide inaccurate measurements for an MPEG-2 transport stream due to the fact that a given IP packet may not contain any video data.

The V-Factor[2] metric is based on deep packet inspection of the video stream (see Fig. 3). It analyzes the bitstream in real time to collect static parameters such as picture size and frame rate as well as dynamic parameters such as the variation of quantization steps. Video quality prediction by the metric is based on the following:

- The impact of video impairments due to the content characteristics, the compression mechanism and bandwidth constraints.
- The impact of network impairments such as jitter, delay and packet loss on the video, including spatial and temporal loss propagation.

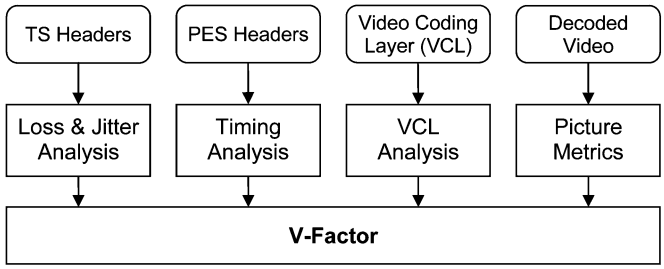

Fig. 3. V-Factor inspects different sections of the video stream, namely, the transport stream (TS) headers, the packetized elementary stream (PES) headers, and the video coding layer (VCL), in addition to the decoded video signal.

The underlying model used for the objective measurement of video impairments is based on a paper by Verscheure *et al.* [24], who proposed models for the impact of packet loss rate, MPEG-2 quantizer scale and data rate on quality using the moving picture quality metric (MPQM). We have generalized these models to state-of-the-art codecs such as H.264, and further enhanced them to take into account the complexity of the video content. Networks impairments are also analyzed in real time in order to provide the model with packet loss probability ratio (single loss, bursty loss) through a series of hidden Markov models. The models were optimized for real-time multi-channel assessment of video quality.

A. Bitstream Analysis

(the quantizer scale on a macroblock basis in an MPEG-2 video stream) provides a first approximation of how video compression affects video quality. MQUANT was shown to exhibit an approximately linear relationship to quality for MPEG-2 clips [24].

We account for spatial and temporal image coding complexity and the impact of packet loss on the spatial and temporal content at the coding layer. Video quality without any network impairments is influenced by video coding layer (VCL) complexity. The VCL complexity is modeled using quantizer values, motion vector information and intra/inter-predicted frame/slice ratios.

As an example, videos with a lot of scene changes would have a very high VCL complexity. The scene changes can be detected in different ways: a Scene Information Message, which labels pictures with scene identifiers; an instantaneous decoding refresh, where all slices are intra-coded; or intra-period changes resulting from I-slice insertion.

Fig. 4 depicts the video coding layer complexity analysis for H.264 streams. The model performs bandwidth and bandwidth variation analysis as well as an inspection of slices and macroblocks in order to analyze the variation of the quantizer, combined with a loss model.

[2]Parts of this technology are patent-pending.

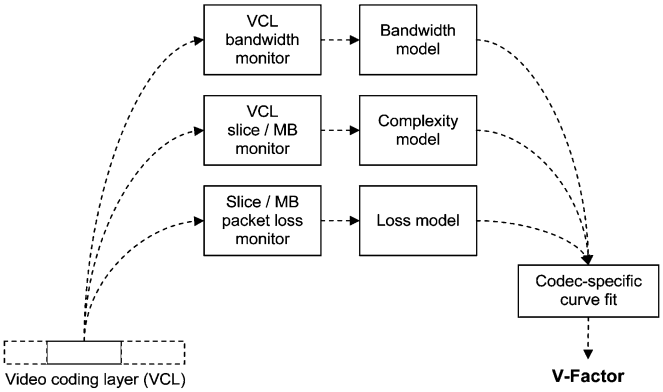

Fig. 4. H.264 video coding layer (VCL) complexity model. The model performs bandwidth and bandwidth variation analysis as well as an inspection of slices and macroblocks, combined with a loss model.

The VCL input is read from the Network Abstraction Layer (NAL)[3] or transport layer. The VCL packet size is used to compute instantaneous and average bandwidth. The bandwidth model is constructed using a 3-state (bandwidth low/average/high) Markov model.

For every macroblock, the VCL complexity model is run. Macroblock and slice quantization parameters are read from the NAL/transport layer by parsing the slice data inside the VCL. A VCL complexity quantization transition probability matrix is computed, and limiting state probabilities are computed. VCL parameters are also monitored for scene transitions and picture quality. Inter/intra macroblock types are analyzed to determine scene transitions and quantization parameter. Computing VCL complexity also follows a Markov process similar to the one for bandwidth, but limited to two states. The transition probabilities are derived from counters that are incremented each time a certain macroblock/slice type is detected.

The visual impact of packet losses on the video content is expressed by combining the video complexity model with a loss model. It includes an analysis of the part of the stream that is lost and its impact on video quality. This allows the model to distinguish between losses involving intra-/inter-predicted macroblocks, I-slices, B-slices, P-slices, high motion, scene cuts, etc.

The overall V-Factor value, which represents a MOS estimate, is computed by a codec-specific curve fit equation using inputs from the bandwidth model, the VCL complexity model and the loss model.

B. Network Losses

Losses can occur either due to IP packet loss in the network, or due to de-jitter buffer under-/over-flow. In parallel to the VCL analysis described above, the content of each packet is inspected in order to determine if the packet contains part of a reference frame or slice, or a predicted frame, slice or macroblock. This analysis is again codec-specific and produces a first statistical model of the distribution of I, B, and P (or SI and SP) frames/slices/macroblocks. Using counters from the complexity analysis, a second statistical model of the distribution of the quantizer values is produced that leads to a combined model of the bandwidth and bandwidth variation for a video stream. Furthermore, inter/intra macroblocks and motion vectors are analyzed; if high motion loss is detected, the loss factor is updated accordingly.

By tracking the inserted time stamps (depending on the encapsulation such as MPEG-2 Transport Stream or RTP) as well as the time stamp carried by some packets, and comparing the difference with a de-jitter buffer, we can produce a jitter model that is used to assess the packet loss probability due to high jitter. The system then assesses whether the computed loss probability will affect a reference or non-reference frame/slice.

C. Encryption

When the video stream is encrypted, as is often the case in commercial video distribution networks, the VCL Raw Byte Sequence Packet (RBSP) segments are not decodable. This imposes a severe limitation on computing video quality, both for traditional metrics and for hybrid metrics, as the impact of losses and loss propagation at the VCL layer cannot be measured directly.

A possible solution to this problem is to perform monitoring before encryption (e.g. at the video head-end) as well as downstream where the video is encrypted. Video timing information is obtained either from the Program Clock Reference (PCR) or the Presentation/Decode Time Stamps (PTS/DTS) from both the head-end and downstream locations; alternatively, GPS/NTP time stamping for correlating head-end and downstream information can be used. This timing information along with VCL information from before encryption and loss event/distribution information from downstream can be used in a reduced-reference manner for correlating the effects of IP packet loss on the quality of the video content even in an encrypted environment.

D. Results

Some V-Factor measurements are shown in Fig. 5 to demonstrate how the method combines transport and video stream information to compute quality. This particular example highlights how different losses and loss types (I-, P- or B-slices) have different impact on quality predictions, in addition to the dependence on video content and complexity characteristics such as the quantizer scale (MQUANT) and the number of scene cuts during loss periods.

V. Video Quality Standards

Few studies exist that compare the prediction performance of different metrics. Formal evaluations of video quality metrics on common test material have been conducted by the Video Quality Experts Group (VQEG),[4] which was established in 1997.

The first round of VQEG tests focusing on full-reference metrics for TV applications ("FR-TV") was inconclusive [28]. Nonetheless, one of the outcomes of this round was a database of test clips with associated subjective ratings that still

[3]The Network Abstraction Layer (NAL) is an intermediate layer between the video coding layer and the transport layer. It was introduced in H.264 to allow for more flexible packaging of the elementary streams.

[4]Seehttp://www.vqeg.org/ for more information.

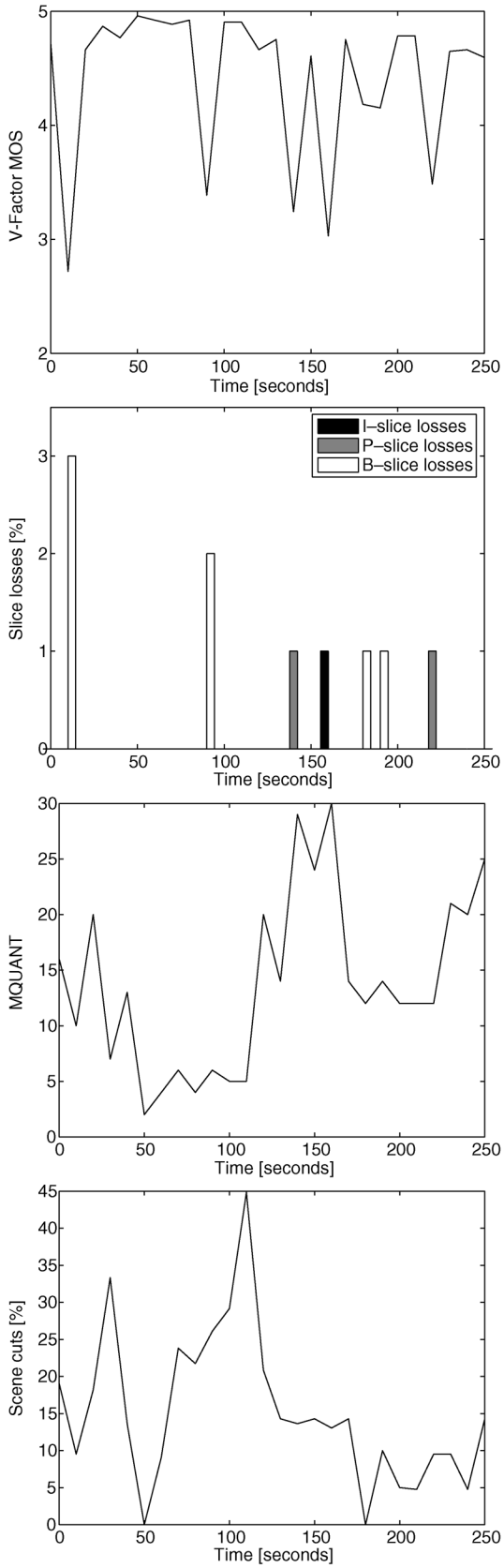

Fig. 5. V-factor MOS prediction (top plot) together with selected loss parameters and video characteristics. Different loss types (I-, P- or B-slices) have different impact on quality. Quantizer scale (MQUANT) and scene cuts are shown as examples of bitstream and content measurements. Every data point refers to a 10-second interval in the video.

represents the only such collection that is publicly available (it can be found on the VQEG web site). A follow-up test was successfully completed in 2003 [29] and has become the basis for two ITU recommendations [30], [31]. The best metrics in this second round achieved correlations as high as 94% with MOS, thus significantly outperforming PSNR with correlations of around 70%. Unfortunately, neither the test sequences nor the subjective data of the second round are public. Both rounds of tests dealt only with full-reference metrics and focused on MPEG-2 compression for digital TV broadcast, and neither included IP networks.

VQEG has also conducted an evaluation of metrics in a "multimedia" scenario, which is targeted at lower bitrates and smaller frame sizes (QCIF, CIF, VGA) as well as a wider range of codecs and transmission conditions (the final report should be available at the time of publication of this paper). Furthermore, VQEG is working on evaluations of reduced- and no-reference metrics for television ("RR/NR-TV") as well as an HDTV test. Recently the group has begun to develop tests for "hybrid" metrics, which look not only at the decoded video as in the other tests, but also at the encoded bitstream (cf. Section III-C above).

In addition to VQEG, various ITU study groups work on the standardization of video quality metrics. ITU-T Study Group 9, for example, is closely aligned with VQEG and uses the group's test results for its recommendations; it has also standardized methods for video registration and alignment [32]. ITU-T Study Group 12 is working on a non-intrusive parametric model for the assessment of multimedia streaming (P.NAMS for short), which uses packet and codec information as inputs, but explicitly excludes any payload information. It also standardized an opinion model for videophone applications [33]. Furthermore, there is an ITU IPTV Global Standards Initiative (GSI),[5] whose task is to coordinate existing IPTV standardization activities. Among other things, it is working on recommendations for performance monitoring and quality of experience requirements for IPTV.

Some other groups also look at video QoE from various angles and in different depth. The DSL Forum has published a report on QoE requirements [34]; the Video Services Forum (VSF) has recommended transport-related metrics for video over IP [35] and recently started an activity group on QoE metrics; and at ATIS, the QoS Metrics Task Force within the IPTV Interoperability Forum is looking at QoE models for video, audio, multimedia and transactions.

VI. Trends

A. Preference and Image Appeal

As mentioned earlier, many existing quality metrics have two important shortcomings:

- They measure video fidelity, i.e. how closely a test clip resembles the original.
- They measure video degradation, i.e. the test video is assumed to be of worse quality than the reference.

However, in some cases processing can improve the perceived quality of a video, even if the result looks less like the original. Video fidelity, even considering the characteristics of the human

[5]See\http://www.itu.int/ITU-T/gsi/IPTV/ for more information.

visual system, is clearly not a good quality benchmark in such situations.

For example, colorful, well-lit, sharp pictures with high contrasts are considered attractive, whereas low-quality, dark and blurry pictures with low contrasts are often rejected [36]. This is true even if the images are enhanced to the extent that they look somewhat unnatural [37]. Especially sharpness and colorfulness have been identified as relevant features in this respect.

Quantitative metrics of this "image appeal" were indeed shown to improve the quality prediction performance of video metrics [38]. Another example is the no-reference quality metric for degraded and enhanced video described by Caviedes and Oberti [39]. It is built from metrics for desirable features (sharpness, contrast, reduced artifacts) as well as non-desirable features (noise, clipping, ringing, blockiness) and accounts for the contributions of both types to video quality.

B. Attention

Another important aspect in video quality evaluation is the fact that people only focus on certain regions of interest in the video, e.g. persons, faces or moving objects. Outside the region of interest, our sensitivity to distortions is significantly reduced. Most objective quality metrics ignore this and weight distortions equally over the entire frame. Only few metrics attempt to model the focus of attention and consider it for computing the overall video quality, for example through the process of "foveation" [40], or using object/face segmentation techniques [41].

Due to the idiosyncrasies of viewer attention mentioned at the beginning of the paper, there is always the risk of viewers looking at regions that were not predicted by the metrics. While this risk may be lower for video than it is for images, understanding and modeling attention in video is still a relatively new area of research; some recent papers have addressed the issue in more detail [42]–[45].

C. Audiovisual Quality

We rarely watch video without sound. Therefore, comprehensive audiovisual quality metrics are needed that analyze both modalities of a multimedia presentation. Audiovisual quality actually comprises two factors. One is the synchronization between the two media, a.k.a. lip-sync; the other is the interaction between audio and video quality.

Various studies have been conducted regarding audio-video synchronization. In actual lip-sync experiments true to their name (showing content with a human speaker), viewers perceive audio and video to be in sync up to about 80 ms of delay [46]. There is a consistently higher tolerance for video ahead of audio rather than vice versa, probably because this is also a more natural occurrence in the real world (light travels faster than sound). Similar results were obtained in experiments with non-speech clips showing a drummer [47]; the same study found the noticeable delay to decrease with drumming frequency.

Studies on audio-video quality interactions have focused mainly on low-bitrate applications such as mobile video, where the audio stream can use up a significant portion of the total bitrate [48], [49]. In one study, we carried out subjective experiments on audio, video and audiovisual quality [50]. We focused on MPEG-4 AVC/H.264 and MPEG-4 AAC video and audio codecs to encode our test material at total bitrates of up to 72 kb/s. We found that both audio and video quality contribute significantly to perceived audiovisual quality. Audio and video quality can be evaluated individually and then combined using linear or bilinear models to predict audiovisual quality with high accuracy, as shown in Fig. 6.

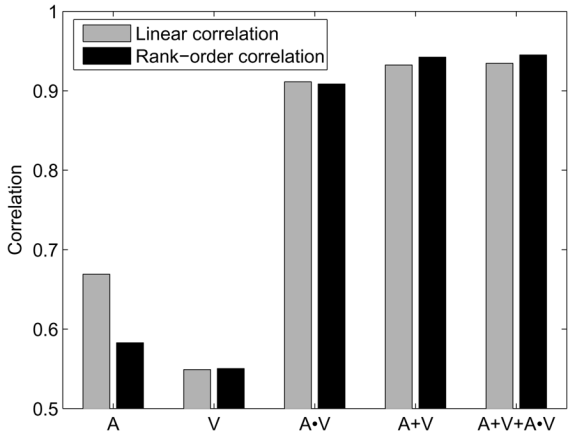

Fig. 6. Correlations of different models for audiovisual quality [50]. A: Prediction from audio quality only; V: Prediction from video quality only; $\mathrm{A}\cdot\mathrm{V}$: Multiplicative model; $\mathrm{A}+\mathrm{V}$: additive model; $\mathrm{A}+\mathrm{V}+\mathrm{A}\cdot\mathrm{V}$: Bilinear model.

Other research on audiovisual quality [51]–[54] has focused on video-conferencing applications (i.e. head-and-shoulders clips) or simulated artifacts. The test material used in these studies is quite different in terms of content range and distortions. Despite these significant differences, the models obtained match rather well in terms of coefficients and prediction performance.

We also used no-reference artifact metrics for audio and video to predict audiovisual MOS [55]. The predictions of the video metrics achieve correlations of above 90% with video MOS; the audio metrics reach 95%. When audio and video metrics are combined according to one of the models for audiovisual MOS mentioned above, audiovisual MOS can be predicted with good accuracy (about 90% correlation).

VII. Conclusions

Although data metrics such as PSNR are still widely used today, significant improvements in prediction performance and/or versatility can only be achieved by QoE metrics. While a lot of work in the past has focused on full-reference metrics, much remains to be done in the areas of no-reference and reduced-reference quality assessment. The same can be said for the quality evaluation of video transmission over networks with packet losses and bit errors. Here the development of reliable metrics is still at the beginning, and many issues remain to be solved. Bitstream- and packet-based metrics appear particularly promising for practical use due to their lower computational complexity and better scalability.

There is a need for reliable perceptual quality measurement in all video applications, and it is becoming more pressing as the number and complexity of video systems increases. As we have highlighted in this review, many interesting QoE measurement approaches and improvements have been proposed, and several

standards are in the making. Nonetheless, we are still a long way from video quality metrics that are widely applicable and universally recognized.

REFERENCES

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

[2] A. J. Ahumada, Jr. and C. H. Null, "Image quality: A multidimensional problem," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 141–148.

[3] S. A. Klein, "Image quality and image compression: A psychophysicist's viewpoint," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 73–88.

[4] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman, "Experienced quality factors—Qualitative evaluation approach to audiovisual quality," in *Proc. SPIE Multimedia on Mobile Devices*, San Jose, CA, January 28–31, 2007, vol. 6507.

[5] P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Winchester, MA: Imcotek Press, 2000.

[6] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, 2002.

[7] ITU-T Recommendation P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," International Telecommunication Union, Geneva, Switzerland, 1999.

[8] ITU-T Recommendation P.911, Subjective Audiovisual Quality Assessment Methods for Multimedia Applications International Telecommunication Union, Geneva, Switzerland, 1998.

[9] P. Corriveau, "Video quality testing," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, 2006, ch. 4.

[10] S. Winkler, "Video quality and beyond," in *Proc. European Signal Processing Conference*, Poznañ, Poland, September 3–7, 2007, invited paper.

[11] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Communications*, vol. 43, no. 12, pp. 2959–2965, December 1995.

[12] I. Avcrbaş, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 206–223, April 2002.

[13] S. Winkler, "Perceptual video quality metrics—A review," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, 2005, ch. 5.

[14] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, pp. 231–252, October 1999.

[15] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, vol. 20, no. 4, pp. 525–536, July 1974.

[16] F. X. J. Lukas and Z. L. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Communications*, vol. 30, no. 7, pp. 1679–1692, July 1982.

[17] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[18] J. Lubin and D. Fibush, "Sarnoff JND Vision Model," T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.

[19] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE Digital Video Compression: Algorithms and Technologies*, San Jose, CA, January 28–February 2 1996, vol. 2668, pp. 450–461.

[20] S. Winkler, "A perceptual distortion metric for digital color video," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, January 23–29, 1999, vol. 3644, pp. 175–184.

[21] S. Süsstrunk and S. Winkler, "Color image quality on the Internet," in *Proc. SPIE Internet Imaging*, San Jose, CA, January 19–22, 2004, vol. 5304, pp. 118–131, invited paper.

[22] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, February 2004.

[23] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, September 2004.

[24] O. Verscheure, P. Frossard, and M. Hamdi, "User-oriented QoS analysis in MPEG-2 delivery," *Real-Time Imaging*, vol. 5, no. 5, pp. 305–314, 1999.

[25] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling packet-loss visibility in MPEG-2 video," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 341–355, 2006.

[26] S. Kanumuri, S. G. Subramanian, P. C. Cosman, and A. R. Reibman, "Predicting H.264 packet loss visibility using a generalized linear model," in *Proc. International Conference on Image Processing*, Atlanta, GA, October 8–11, 2006, pp. 2245–2248.

[27] S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proc. International Symposium on Wireless Personal Multimedia Communications*, Aalborg, Denmark, September 9–12, 2001, pp. 547–552, invited paper.

[28] VQEG, "Final report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment," April 2000 [Online]. Available: http://www.vqeg.org/

[29] VQEG, "Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment—Phase II," August 2003 [Online]. Available: http://www.vqeg.org/

[30] ITU-T Recommendation J.144, "Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference," International Telecommunication Union, Geneva, Switzerland, 2004.

[31] ITU-R Recommendation BT.1683, "Objective Perceptual Video Quality Measurement Techniques for Standard Definition Digital Broadcast Television in the Presence of a Full Reference," International Telecommunication Union, Geneva, Switzerland, 2004.

[32] ITU-T Recommendation J.244, "Calibration Methods for Constant Misalignment of Spatial and Temporal Domains With Constant Gain and Offset," International Telecommunication Union, Geneva, Switzerland, 2008.

[33] ITU-T Recommendation G.1070, "Opinion model for video-telephony applications," International Telecommunication Union, Geneva, Switzerland, 2007.

[34] DSL Forum, Triple-play services quality of experience (QoE) requirements DSL Forum Architecture and Transport Working Group, Tech. Rep. TR-126, 2006.

[35] Video Services Forum, VSF Test and Measurements Activity Group 2006, Tech. Rep..

[36] A. E. Savakis, S. P. Etz, and A. C. Loui, "Evaluation of image appeal in consumer photography," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, January 23–28, 2000, vol. 3959, pp. 111–120.

[37] S. N. Yendrikhovskij, F. J. J. Blommaert, and H. de Ridder, "Perceptually optimal color reproduction," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, January 26–29, 1998, vol. 3299, pp. 274–281.

[38] S. Winkler, "Visual fidelity and perceived quality: Towards comprehensive metrics," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, January 21–26, 2001, vol. 4299, pp. 114–125.

[39] J. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. : CRC Press, 2005, ch. 10.

[40] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, March 2002.

[41] A. Cavallaro and S. Winkler, "Segmentation-driven perceptual quality metrics," in *Proc. International Conference on Image Processing*, Singapore, October 24–27, 2004, pp. 3543–3546.

[42] W. Osberger and A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, January 21–26, 2001, vol. 4299, pp. 361–372.

[43] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1304–1318, October 2004.

[44] M. T. López, M. A. Fernández, A. Fernández-Caballero, J. Mira, and A. E. Delgado, "Dynamic visual attention model in image sequences," *Image and Vision Computing*, vol. 25, no. 5, pp. 597–613, May 2007.

[45] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483–2498, September 2007.

[46] R. Steinmetz, "Human perception of jitter and media synchronization," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 61–72, January 1996.

[47] R. Arrighi, D. Alais, and D. Burr, "Perceptual synchrony of audiovisual streams for natural and artificial motion sequences," *Journal of Vision*, vol. 6, no. 3, pp. 260–268, 2006.
[48] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, and M. Rupp, "Audiovisual quality estimation for mobile streaming services," in *Proc. International Symposium on Wireless Communication Systems*, Siena, Italy, September 5–7, 2005.
[49] S. Jumisko-Pyykkö, "I would like to see the subtitles and the face or at least hear the voice: Effects of picture ratio and audiovideo bitrate ratio on perception of quality in mobile television," *Multimedia Tools and Applications*, vol. 36, no. 1–2, pp. 167–184, January 2008.
[50] S. Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, 2006.
[51] M. P. Hollier and R. Voelcker, "Towards a multi-modal perceptual model," *BT Technology Journal*, vol. 15, no. 4, pp. 162–171, 1997.
[52] C. Jones and D. J. Atkinson, "Development of opinion-based audiovisual quality models for desktop video-teleconferencing," in *Proc. International Workshop on Quality of Service*, Napa, CA, May 18–20, 1998, pp. 196–203.
[53] J. G. Beerends and F. E. de Caluwe, "The influence of video quality on perceived audio quality and vice versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, May 1999.
[54] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, December 2004.
[55] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, January 16–20, 2005, vol. 5666, pp. 139–148.

Stefan Winkler holds an M.Sc. degree in Electrical Engineering from the University of Technology in Vienna, Austria, and a Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Dr. Winkler is currently Principal Technologist for Symmetricom's QoE Assurance Division. Prior to that, he was chief scientist of Genista Corporation, which he co-founded in 2001. He has also held assistant professor positions at the National University of Singapore (NUS) and the University of Lausanne, Switzerland. Dr. Winkler has published more than 40 papers related to perceptual quality measurement and is the author of the book "Digital Video Quality". He has also been an active member of and contributor to the Video Quality Experts Group (VQEG) since it was founded in 1997.

Praveen Mohandas holds a B.Eng. degree in Computer Science and Engineering from the University of Mysore, India. He has over 17 years of experience designing systems and software for routing, switching, signaling, operating systems, multiprocessor communication and synchronization, QoS and high availability.

Mr. Mohandas is currently Software Architect with Symmetricom's QoE Assurance Division. Previously he worked at Lucent Technologies and was part of the team that was awarded the Bell Labs President's Gold Medal for outstanding level of innovation and technical excellence. He has several patents pending. His current interests include video coding, video quality assessment, multimedia systems and statistical modeling of video loss in IP networks.