VIDEO QUALITY AND BEYOND

Stefan Winkler

Symmetricom, QoE Division
San Jose, CA 95131, USA
http://qoe.symmetricom.com
swinkler@symmetricom.com

ABSTRACT

This paper gives a brief overview of the current state of the art of video quality metrics and discusses their achievements as well as shortcomings. It also summarizes the main standardization efforts by the Video Quality Experts Group (VQEG). It then looks at recent trends and developments in video quality research, in particular the emergence of new generations of quality metrics (compared to those focused on compression artifacts), including comprehensive audiovisual quality measurement.

1. INTRODUCTION

Video quality has many different aspects, and the evaluation (and ultimately optimization) of the quality of digital video systems has turned out to be a highly complex problem. The reasons are two-fold [28]:

- Video systems are complex and consist of many components. These include capture and display hardware, converters, codecs, networks, all affecting quality in their own way.
- 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.

Many approaches to this problem have been proposed, some of which will be briefly discussed in this paper, yet there is still no well-recognized reliable method for video quality assessment. The reference to date are still subjective experiments; unfortunately, they are time-consuming in setup and execution, and severely limited in the scope of the video material that can be evaluated. When it comes to monitoring hundreds of hours/channels of video, objective video quality metrics are the only practicable solution.

The paper is organized as follows. Section 2 introduces the basics of quality assessment and metrics classification. Section 3 discusses developments in standardization. Section 4 takes a look at the latest trends in video quality research. Section 5 concludes the paper.

2. BASICS

Numerous factors contribute to what a viewer perceives as "video quality" [1, 11, 18]. These include individual interests, quality expectations, display type and properties, viewing conditions etc. The wide variety and subjectivity of some of these factors are indicators of the complexity of the quality measurement problem. Indeed, most of today's metrics remain oblivious to the majority of these factors and focus instead on measuring the visual fidelity of the video in terms of the distortions introduced by various processing steps, in particular lossy compression.

The measurement of these distortions can be done in two ways. *Data metrics* look at the fidelity of the signal without considering its content. *Picture metrics* treat the data as the visual information that it contains. Metrics can also be classified based on the amount of reference information they require. These classifications are discussed next.

2.1 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. 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. Over the years, video researchers have developed a familiarity with PSNR that allows them to interpret the values immediately. Minimizing MSE is also very well understood from a mathematical point of view. Last but not least, there are no other metrics as widely recognized as these two, and the lack of video quality standards does not help either.

Despite their popularity, MSE and PSNR only have an approximate relationship with the video quality perceived by human observers, simply because they are based on a byte-by-byte comparison of data without knowing what the data actually represents. They are completely oblivious 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. One of the best examples that PSNR cannot work well is the dithering of images with reduced color depth. Dithering adds noise to the image to remove the perceived banding caused by the color reduction. As a result, PSNR is reduced, yet perceived quality improves.

A number of additional pixel-based metrics have been proposed and tested [7]. It was found that although some of these metrics can predict subjective ratings quite successfully for a given compression technique or type of distortion, they are not reliable for evaluations across techniques. MSE was found to be a good metric for additive noise, but is outperformed by more complex HVS-related techniques for coding artifacts [3].

The network quality-of-service community has equally simple metrics to quantify transmission error effects, such as packet loss rate or bit error rate. 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 importance of the packets and bits concerned. The same number of lost packets can have drastically different effects depending on which parts of the bitstream are affected.

2.2 Picture Metrics

Due to the problems with these simple data metrics, much effort has been spent on designing better visual quality metrics that quantify 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* [29].

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 [25]. Due to their generality, these metrics can in principle be used for a wide range of video distortions. HVS-based metrics date back to the 1970's and 1980's, when Mannos and Sakrison [15] and Lukas and Budrikis [14] developed the first image and video quality metrics, respectively. Later well-known metrics in this category are the Visual Differences Predictor (VDP) by Daly [6], the Sarnoff JND (just noticeable differences) metric by Lubin [13], van den Branden Lambrecht's Moving Picture Quality Metric (MPQM) [20], and the author's own perceptual distortion metric (PDM) [26].

The engineering approach is based primarily on the extraction and analysis of certain features or artifacts in the video. These can be either structural elements such as contours, or specific distortions that are introduced by a particular video processing step, compression technology or transmission link. The metrics look for the strength of these features 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 analysis rather than fundamental vision modeling is the conceptual basis for their design. The engineering approach has gained popularity in recent years, and several metrics that fall into this category will be discussed in in Section 4 below.

2.3 Reference Information

Quality metrics are generally classified into the following categories based on the amount of information required about the reference video [29].

Full-reference (FR) metrics perform a frame-by-frame comparison between a reference video and the video under test. They require the entire reference video to be available, usually in uncompressed form, which is quite an important 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 reference clip. Temporal registration in particular is quite a strong restriction and can be very difficult to achieve in practice. 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.

No-reference (NR) metrics analyze only the video under test, without the need of an explicit reference. This makes them much more flexible than FR metrics, as it can be next to impossible to get access to the reference (e.g. video captured by a camera). They are also completely free from alignment issues. The main difficulty 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 (e.g. a chessboard may be interpreted as block artifacts in the extreme case). The majority of NR metrics are based on estimating blockiness, which is the most prominent artifact of block-DCT based compression methods such as H.26x, MPEG and their derivatives. The author has developed NR metrics for blockiness and various other artifacts [30, 31].

Reduced-reference (RR) metrics are a compromise between FR and NR metrics. They extract a number of features from the reference video (e.g. the amount of motion or spatial detail), and the comparison with the video under test is then based only on those features. This makes it possible to avoid some of the 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.

3. VIDEO QUALITY STANDARDS

Few studies exist that compare the prediction performance of different metrics. Formal evaluations of video quality metrics have been conducted by the Video Quality Experts Group (VQEG), [1] which was established in 1997.

Due to the complexity and scale of this task, the first round of tests was inconclusive [22]. Nonetheless, one of the outcomes of this round was a database of test clips with subjective ratings that still represents the only publicly available such collection. A follow-up test was successfully completed in 2003 [23] and has become the basis for two ITU recommendations. The best metrics in the second round achieved correlations as high as 94% with the 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 applications ("FR-TV").

VQEG is currently conducting 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. Furthermore, VQEG is working on evaluations of reduced- and no-reference metrics for television ("RR/NR-TV") as well as HDTV. In the latest meeting, the group also proposed to study "hybrid" metrics, which look not only at the decoded video as in the other tests, but also at the encoded bitstream.

4. TRENDS

4.1 Preference and Image Appeal

An important shortcoming of existing metrics is that they measure image fidelity instead of perceived quality. The accuracy of the reproduction of the original on the display, even considering the characteristics of the human visual system, is not the only quality benchmark. For example, colorful, well-lit, sharp pictures with high contrasts are considered attrac-

[1] See the VQEG web site for more information on its activities and reports: http://www.vqeg.org/.

tive, whereas low-quality, dark and blurry pictures with low contrasts are often rejected [18]. Especially sharpness and colorfulness have been identified as relevant features. Quantitative metrics of this "image appeal" were indeed shown to improve quality prediction performance [27].

Another example of this approach is the perceptual video quality measure (PVQM) by Hekstra et al. [9]. It uses a linear combination of three specific features, namely the loss of edge sharpness, the color error normalized by the saturation, and the temporal variability of the reference video. The PVQM was one of the best metrics in the VQEG FR-TV Phase I test [22].

4.2 Structural Information

Wang et al. [24] presented a video quality assessment method based on structural similarity (SSIM). It 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. Despite the relative simplicity of the metric, it performed well on the public VQEG FR-TV Phase I database (see Section 3). One strong point of SSIM is that it also has also been show to work for some artifacts that are not directly related to compression, such as added noise.

Wolf and Pinson [34] designed a video quality metric (VQM) that uses effective low-level video features, which were selected empirically from a pool of candidate features. The test sequences are divided 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 are then compared with the reference using a process similar to masking. The VQM was among the best metrics in the VQEG FR-TV Phase II evaluation [23].

4.3 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 some moving objects. Outside of the region of interest, our sensitivity is significantly reduced. Most quality assessment systems ignore this and weight distortions equally over the entire frame. Some recent metrics attempt to model the focus of attention and consider it for computing the overall video quality [5, 12, 16]. However, understanding and modeling attention in video is still a challenge. Due to the idiosyncrasies of viewer attention, there is always the risk of viewers looking at regions that were not predicted by the metrics.

4.4 Delivery Quality

While a lot of effort in video quality measurement has been devoted to compression artifacts and decoded video, there is also a growing interest in metrics specifically designed to measure the impact of network losses on video quality. Because losses directly affect the encoded bitstream, such metrics are often based on parameters that can be extracted from the bitstream with no or only partial decoding. This has the added advantage of lower data rates and thus lower processing power and speed requirements compared to metrics looking at the decoded video.

For example, Verscheure et al. [21] investigated the joint impact of packet loss rate and MPEG-2 bitrate on video quality. Kanumuri et al. [10] used various bitstream parameters such as motion vector length or number of slice losses to predict the visibility of packet losses in MPEG-2 video.

4.5 Audiovisual Quality

We rarely watch video without sound. Therefore, comprehensive audiovisual quality metrics are needed that analyze both parts of the 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 [19]. There is a consistently higher tolerance for video ahead of audio rather than vice versa. This is confirmed by experiments with non-speech clips showing a drummer [2], which found the noticeable delay limit to decrease with drumming frequency.

In a study on audio-video quality interactions, the author carried out subjective experiments on audio, video and audiovisual quality [33]. Our main interest was mobile video transmission. We focused on MPEG-4 AVC/H.264 and MPEG-4 AAC to encode our test material at very low bitrates. We found that both audio and video quality contribute significantly to perceived audiovisual quality; the product of audio quality and video quality is an effective model of audiovisual quality, and so is the linear combination of the two.

Other research has focused on video-conferencing applications (i.e. head-and-shoulders clips) or simulated artifacts; the material used in our test was quite different in terms of content range and distortions. Despite these significant differences, the additive and multiplicative models are similar to previous results [4, 17, 8].

We also used non-reference artifact metrics for audio and video to predict audiovisual MOS [32]. The predictions from 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 the models for audiovisual MOS mentioned above, audiovisual MOS can be predicted with good accuracy (about 90% correlation).

5. 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 perceptual quality metrics. While a lot of work has focused on full-reference metrics for TV/broadcast applications, 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 low-bitrate video and transmission error artifacts. Here the development of reliable metrics is still at the beginning, and many issues remain to be solved.

Applications for reliable perceptual quality measurement can be found in all video systems. As some of the examples in Section 4 show, interesting improvements and approaches have been proposed. Nonetheless, we are still a long way from universally accepted video quality metrics.

REFERENCES

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

[2] R. Arrighi, D. Alais, D. Burr: "Perceptual synchrony of audiovisual streams for natural and artificial motion sequences." *J. Vision* **6**(3):260–268, 2006.

[3] İ. Avcıbaş, B. Sankur, K. Sayood: "Statistical evaluation of image quality measures." *J. Electronic Imaging* **11**(2):206–223, 2002.

[4] J. G. Beerends, F. E. de Caluwe: "The influence of video quality on perceived audio quality and vice versa." *J. Audio Eng. Soc.* **47**(5):355–362, 1999.

[5] A. Cavallaro, S. Winkler: "Segmentation-driven perceptual quality metrics." in *Proc. ICIP*, pp. 3543–3546, Singapore, 2004.

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

[7] A. M. Eskicioglu, P. S. Fisher: "Image quality measures and their performance." *IEEE Trans. Comm.* **43**(12):2959–2965, 1995.

[8] D. S. Hands: "A basic multimedia quality model." *IEEE Trans. Multimedia* **6**(6):806–816, 2004.

[9] A. P. Hekstra et al.: "PVQM – a perceptual video quality measure." *Signal Processing: Image Communication* **17**(10):781–798, 2002.

[10] S. Kanumuri, P. C. Cosman, A. R. Reibman, V. A. Vaishampayan: "Modeling packet-loss visibility in MPEG-2 video." *IEEE Trans. Multimedia* **8**(2):341–355, 2006.

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

[12] S. Lee, M. S. Pattichis, A. C. Bovik: "Foveated video quality assessment." *IEEE Trans. Multimedia* **4**(1):129–132, 2002.

[13] J. Lubin, D. Fibush: "Sarnoff JND vision model." T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.

[14] F. X. J. Lukas, Z. L. Budrikis: "Picture quality prediction based on a visual model." *IEEE Trans. Comm.* **30**(7):1679–1692, 1982.

[15] J. L. Mannos, D. J. Sakrison: "The effects of a visual fidelity criterion on the encoding of images." *IEEE Trans. Inform. Theory* **20**(4):525–536, 1974.

[16] W. Osberger, A. M. Rohaly: "Automatic detection of regions of interest in complex video sequences." in *Proc. SPIE*, vol. 4299, pp. 361–372, San Jose, CA, 2001.

[17] R. Pastrana-Vidal, C. Colomes, J. Gicquel, H. Cherifi: "Caractérisation perceptuelle des interactions audiovisuelles: Revue." in *Proc. CORESA Workshop*, Lyon, France, 2003.

[18] A. E. Savakis, S. P. Etz, A. C. Loui: "Evaluation of image appeal in consumer photography." in *Proc. SPIE*, vol. 3959, pp. 111–120, San Jose, CA, 2000.

[19] R. Steinmetz: "Human perception of jitter and media synchronization." *IEEE J. Selected Areas in Comm.* **14**(1):61–72, 1996.

[20] C. J. van den Branden Lambrecht, O. Verscheure: "Perceptual quality measure using a spatio-temporal model of the human visual system." in *Proc. SPIE*, vol. 2668, pp. 450–461, San Jose, CA, 1996.

[21] O. Verscheure, P. Frossard, M. Hamdi: "User-oriented QoS analysis in MPEG-2 delivery." *Real-Time Imaging* **5**(5):305–314, 1999.

[22] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." 2000, available at http://www.vqeg.org/.

[23] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment – Phase II." 2003, available at http://www.vqeg.org/.

[24] Z. Wang, L. Lu, A. C. Bovik: "Video quality assessment based on structural distortion measurement." *Signal Processing: Image Communication* **19**(2):121–132, 2004.

[25] S. Winkler: "Issues in vision modeling for perceptual video quality assessment." *Signal Processing* **78**(2):231–252, 1999.

[26] S. Winkler: "A perceptual distortion metric for digital color video." in *Proc. SPIE*, vol. 3644, pp. 175–184, San Jose, CA, 1999.

[27] S. Winkler: "Visual fidelity and perceived quality: Towards comprehensive metrics." in *Proc. SPIE*, vol. 4299, pp. 114–125, San Jose, CA, 2001.

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

[29] S. Winkler: "Perceptual video quality metrics – a review." in *Digital Video Image Quality and Perceptual Coding*, eds. H. R. Wu, K. R. Rao, chap. 5, CRC Press, 2005.

[30] S. Winkler, R. Campos: "Video quality evaluation for Internet streaming applications." in *Proc. SPIE*, vol. 5007, pp. 104–115, Santa Clara, CA, 2003.

[31] S. Winkler, F. Dufaux: "Video quality evaluation for mobile applications." in *Proc. SPIE*, vol. 5150, pp. 593–603, Lugano, Switzerland, 2003.

[32] S. Winkler, C. Faller: "Audiovisual quality evaluation of low-bitrate video." in *Proc. SPIE*, vol. 5666, pp. 139–148, San Jose, CA, 2005.

[33] S. Winkler, C. Faller: "Perceived audiovisual quality of low-bitrate multimedia content." *IEEE Trans. Multimedia* **8**(5):973–980, 2006.

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