John G. Apostolopoulos
and Amy R. Reibman

The Challenge of Estimating Video Quality in Video Communication Applications

Video communication over the Internet has gone from science fiction to reality. Not only is video streaming and two-way video conferencing becoming common, but, as recently reported by the Associated Press through ABC News [1], today even broadcast television programs such as eyewitness news often use live video from camera phones or Skype on a notebook or desktop computer to interview people and share information. A desired goal in these scenarios is to ensure the best video quality possible. This leads to the challenge of estimating video quality as a key step to achieving the above. This article's goal is to describe the need for video quality estimation and why it is challenging as well as highlight the different types of video quality estimators (VQEs) being developed. Signal processing is viewed as critical throughout this work.

APPLICATIONS

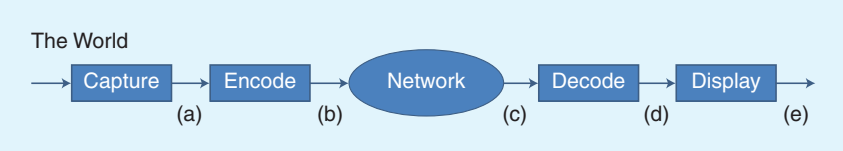
Video delivery over modern communication networks is becoming ubiquitous through its use with an ever larger range of different applications. Initial applications included television (TV) broadcasting where recent examples in the United States include services provided by AT&T U-Verse and Verizon FiOS, and streaming video-on-demand (VOD) by NetFlix, Amazon, and Hulu. Video surveillance systems are gaining in popularity for both commercial and consumer use (for example, monitoring one's home or children). Also, two-way software-based video communication usage, such as Skype and ooVoo, is rapidly increasing. Furthermore, eyewitness news programs and two-way video from cell phones or notebook computers help bring distant people and visual information to the newsroom and then to the rest of the world.

Each of the diverse range of applications mentioned above has its own needs in terms of video quality estimation, and even within a single application there are often multiple groups involved who have different needs. For example, in Internet Protocol TV (IPTV) or a VOD service, a single video might pass from content provider to service provider to network provider before reaching the end viewer. The content provider owns or is licensed to sell the content. The service provider sells IPTV service to the customer, while the network provider connects the service providers and its customers. The content provider wants to ensure that their video, which they created to have sufficient quality, is not further degraded when delivered to the final consumer. Service providers want to guarantee not only that the video they are given by the content provider has sufficient quality, but also that the network provider does not degrade it significantly. One of the challenges of TV broadcast is that network providers have no way of distinguishing whether a poor-quality video has been improperly acquired; it could also be a valuable yet low-quality eyewitness news video. Even among, say, network providers, there are a range of problems. In some cases, the goal is an accurate estimate of the video quality for a single video session [2]. In other cases, accuracy may be less important than a real-time, lightweight VQE that can be applied to provide online monitoring of large-scale network deployments [3]. The scenarios below therefore have different importance for each of the content, service, and network providers.

VIDEO DELIVERY OVER MODERN COMMUNICATION NETWORKS IS BECOMING UBIQUITOUS THROUGH ITS USE WITH AN EVER LARGER RANGE OF DIFFERENT APPLICATIONS.

In an application, the video quality depends on a variety of factors, as illustrated in the end-to-end Figure 1 and described in the following six sections. At each stage of the processing chain, video quality can be measured locally and system parameters can be adapted to obtain better quality. In addition, features affecting quality can also be extracted to assist in characterizing quality at a later stage of the processing chain. Video

(continued on page 156)



[FIG1] **Parts (a)–(e) show the components of an end-to-end video communication system and five locations where measurements can be performed. Some applications may impose constraints that limit the ability to measure at one or more of these locations.**

Digital Object Identifier 10.1109/MSP.2011.943553
Date of publication: 17 February 2012

1053-5888/12/$31.00©2012IEEE

 continued from page 160

quality estimation is used for both 1) online processing to monitor, detect problems, and improve the live session, and 2) offline processing for subsequent analysis and system improvement.

ACQUISITION QUALITY

For the viewer to experience high-quality video, the captured video should have minimal undesired attributes. If the captured video is afflicted by defocus (blur), motion shake from handheld cameras, noise due to low light, overexposure, improper framing of the content, or other undesirable effects, this may affect the ability of later stages to perform well. For example, images captured in low light are more difficult to compress than those captured in bright light.

COMPRESSION

To communicate video over current wired and wireless networks requires significant lossy compression—often the compressed video has a bit rate on the order of 100 times smaller than the original video. The compressed video may be afflicted by compression artifacts, such as blocking, ringing, or blurring. In addition, compressed video is much more vulnerable to transmission impairments such as packet loss, which can lead to visible artifacts in the decoded video. The severity of the artifacts depends on a number of factors including the type of network impairment and the details of how the compression and transport is performed. Error- or loss-resilient video coding refers to efforts to make the compressed video more resilient to errors or losses in transmission. This depends on the compression algorithm (e.g., MPEG-2 or H.264), the encoding parameter choices such as the group of pictures structure, the strategy used to map the bit stream into packets, and the error concealment performed at the decoder to hide losses.

TRANSPORT

Despite dramatic improvements in our communication networks that have made video delivery practical, these networks often still exhibit congestion and wireless interference. The resulting impairments may include packet loss or bit errors either in isolation or in bursts, delay and delay jitter, all of which vary over time due to network conditions.

DISPLAY

The display device, environmental viewing conditions, and viewer have a significant effect on the perceived quality. Low-contrast displays may produce poor quality in strong sunlight; motion blur may degrade quality for highly active scenes. Further, past experiences, current expectations, and the motivations of the viewer all affect perceived quality. For example, if this is the only way a soccer fan can watch World Cup Soccer, his or her pleasure in seeing any action may outweigh any otherwise annoying degradations.

THE DISPLAY DEVICE, ENVIRONMENTAL VIEWING CONDITIONS, AND VIEWER HAVE A SIGNIFICANT EFFECT ON THE PERCEIVED QUALITY.

VIDEO CONTENT

The video content itself also has a profound effect. Different types of video content can influence the perceived video quality at all points in the processing chain. For example, action sports such as football have quite different attributes such as motion as compared to typical head-and-shoulders video conferencing.

AUDIO CONTENT

While we focus our attention on video quality, the quality of the audio is paramount to the quality of the entire experience. For example, high-quality audio content may make a video "look" better, and bad audio or poor audio/video synchronization may make a high-quality video "look" bad.

OVERVIEW OF VIDEO QUALITY ESTIMATION TECHNIQUES

The importance of video quality estimation has led to the development of many diverse techniques for measuring quality in video communications. Different applications have different needs for quality monitoring. A video quality estimator need only be effective for its intended usage. First, it may only need to measure a subset of impairments. For example, if the VQE is measuring the impact of transcoding on quality, it only needs to measure the impact of compression. Alternatively, if the VQE is measuring the impact of network transmission, it only needs to quantify accurately the impact of packet losses. Second, it may only need to be accurate over a certain class of videos. For example, it may be targeted toward head-and-shoulders low-motion video, or high-quality nature photography. Third, it may only need to be accurate over a range of qualities. For example, it may be able to focus on low-rate videos delivered to mobile devices.

One of the biggest drivers of diversity in the techniques of video quality estimation, however, is the fact that the application of the VQE may constrain what is possible to be measured. For example, a video conferencing application may only be able to perform measurements at the end points (sender and receiver). A network provider may be able to measure in the network, but not at the end points. Figure 1 identifies five places where measurements can be performed.

We next highlight three different classes of approaches for video quality estimation [2]. They differ in terms of how much information is required, the level of processing complexity, and the accuracy they provide.

- *Full-reference VQEs* require access to both the decoded video and the original video (the reference) in the same physical location to estimate the quality. Because this may not be feasible in a networked application, they are most appropriate for in-the-lab testing. Also, these techniques typically assume that the original video has maximum quality. A thoughtful discussion of visual quality measures is given in [4]. By comparing the decoded video to the reference video, a

full-reference VQE is able to compute the distortion in the decoded video exactly. Its challenge is to determine which aspects of the distortion are visible and which do not affect quality. Thus, while full-reference VQEs are perhaps the simplest set of techniques conceptually, they have the advantage of being able to solve this challenge by easily incorporating extensive knowledge of the human visual system.

■ *No-reference pixel-based VQEs* attempt to estimate quality by decoding the video but without access to the original. Clearly, this is much more challenging than the full-reference case. No-reference pixel-based VQEs do not have the reference video and are therefore unable to exactly compute the distortion. Instead, they must infer the desired reference, distinguishing between desired signal and undesired distortion caused by compression and transmission artifacts. One approach is to model the signal, another is to model the distortion. A study of no-reference VQEs is given in [5], where the authors stress that these estimators only need to be as accurate as required by the applications they are being used for. A successful no-reference VQE is likely to limit its scope to the desired application and desired accuracy and to rely on accurate models of both the reference image and the distortion. To the extent possible, even no-reference VQEs should incorporate known principles of the human visual system, including the facts that image contrast is more meaningful perceptually than luminance, that errors may be hidden, or masked, by the desired signal, and that the eye is more sensitive to low-frequency information than high-frequency signals.

■ *No-reference bit stream-only VQEs* attempt to estimate quality without completely decoding the video to obtain decoded pixels. At one extreme, they may only count missing packets using packet header information. A VQE that measures only the number of missing packets [i.e., the network-layer packet loss rate (PLR)]has the lowest complexity, but poor performance. It would be unable to accurately compare two videos without packet loss, and it would not be able to capture the manner in which video quality with nonzero packet loss depends strongly on how the compression and transport is performed. For example, if the video is not coded with loss-resilience, the prediction between frames may cause the error from a single packet loss to propagate and create visible artifacts across a large number of consecutive video frames, as shown in Figure 2. On the other hand, with loss-resilient coding the same packet loss may only afflict a single frame, or after concealment the error may be negligible. Thus while such an extreme solution may be feasible inside the network, it is unlikely to be highly accurate.

However, a no-reference bit stream-only VQE might also obtain additional useful information by partially parsing the bit stream to examine compression parameters. For instance, the coarseness of quantization or the spatial and temporal location of missing packets may be extracted to enable a more accurate estimate of the video quality. These techniques have the advantage of having the lowest complexity; however they are also more restrictive in terms of the accuracy of the information they can provide.

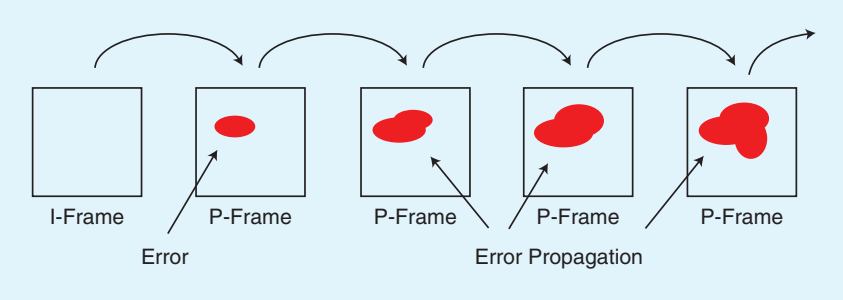
Complementary information can be obtained from the no-reference pixel-based and bit stream-only VQEs. From the bit stream, we can extract exactly what parameters were used during compression to create the bit stream, and we can identify exactly where in the video a packet loss has occurred. For example, extracting the quantizer step-size used during compression enables a bit stream-only VQE to determine if a region of video is likely to accurately represent the uncompressed video.

On the other hand, a no-reference pixel-based VQE can know exactly how the decoder concealed any missing information from lost packets and can understand exactly what is displayed to the viewer; it has no need to rely on assumptions about error concealment. Such a VQE can also examine the statistical characteristics of pixels to assess the perceptual attributes in the video, including colorfulness, blockiness, blurriness, temporal jerkiness, and noisiness.

Not surprisingly, therefore, more accurate quality assessment is possible by designing a VQE that combines all available information: bit stream, decoded pixels, and even original unencoded pixels if available.

ADDITIONAL CONSIDERATIONS IN QUALITY ESTIMATION

Another important consideration about measuring video quality is ascertaining exactly what the application requires in terms of quality. How is quality defined for the specific application? While average performance may be an adequate measure for some applications, in other



[FIG2] Video coders apply prediction between frames to achieve compression: the first frame is intracoded (I-frame) and the subsequent frames are predictively coded (P-frames) where the arrows show the prediction between frames. This approach leads to significant compression but also makes the video vulnerable to packet loss. This figure illustrates how a single lost packet leads to an error in decoded frame two and also to significant subsequent error propagation both in time and space.

cases the application requires immediate real-time knowledge of quality on a specific channel. Does the application require knowledge of actual "quality," or is it sufficient to merely understand how often a visible degradation occurs?

In general, the performance of objective quality estimators are evaluated using actual subjective estimates of quality obtained from actual viewers. To obtain this ground truth, however, requires careful subjective testing methodology. One of the most critical aspects is asking viewers the right questions. If the application requires knowledge of the actual "quality" of the video, then the subjective ground truth should be gathered by asking viewers about quality. If, on the other hand, the application merely needs to understand how often a visible degradation occurs, a different set of questions is necessary.

One particular challenge is how to measure quality perceived by humans of video subjected to rare events like packet losses. Not only is it difficult for viewers to evaluate their reaction to rare events, but also many such losses are not detected by many viewers. In these cases, the losses are masked by the video itself. Further, the viewers may be focused on the content and may not notice a small, transient artifact. Therefore, it may be an effective approach to break the problem up. In the first step, one quantifies when a packet loss creates a visible artifact; in the second step, the resulting quality of the video is computed based on the frequency of visible artifacts.

It is tempting to categorize the requirements of a specific video based on its type. For example, it is commonly suggested that video containing sports content requires both high bandwidth and low packet loss rate. This assumption is true whenever the video content contains periods of high-motion activity. However, there are many periods of time within a sports transmission during which instantaneous bandwidth is low and any packet losses will likely be invisible. This is because within a sports show, there are often extended periods of idleness: while waiting for a pitcher to communicate with the catcher, or while showing a close-up of a player or a referee. Hence, the immediate temporal content is often more important than the class of video when establishing requirements for both bit rate and packet loss rates. As a result, a segment from a TV soap opera containing a slow-motion pan may actually have stricter packet-loss requirements than many segments from a sportscast.

Stereo three-dimensional (3-D) video is quickly gaining in popularity, and with it the need for appropriate video quality estimation increases. For quality estimation of stereo 3-D, it is not sufficient to assess only the two individual videos that comprise the 3-D stereo, because the relationship between the two videos governs our perception. Viewing two high-quality videos of a stereo pair may lead to significant eye strain, fatigue, and headaches if the pair is not well matched. Correct depth perception requires careful attention to stereo capture and display. During capture, the interaxial distance (baseline) between the cameras, convergence, and zoom of a stereo camera must be carefully coordinated. Complications arise when the stereo video is to be viewed on displays of different sizes by viewers with different interocular distances (spacing between the eyes) at different distances from the display. Interested readers are encouraged to examine [6] and references therein to better understand the issues and best practices for capture and display of high-quality 3-D video.

ONE PARTICULAR CHALLENGE IS HOW TO MEASURE QUALITY PERCEIVED BY HUMANS OF VIDEO SUBJECTED TO RARE EVENTS LIKE PACKET LOSSES.

CONCLUDING REMARKS

As discussed at the beginning of this article, different applications have different needs for quality estimation as well as constraints on estimating quality. Selecting or designing an appropriate VQE depends on various factors including available information (e.g., original video, encrypted or nonencrypted content), computational capability, where measurement is performed (e.g., in-network on encrypted data or at end node after decryption), scalability requirement, and level of quality needed to measure and required accuracy.

The need for video quality estimation will increase in the future with the expected growth of video. Important areas deserving further work include:

- developing effective, low-complexity no-reference video quality estimators designed for the needs of specific applications
- lightweight VQEs that are amenable to large-scale network deployments and monitoring a large number of streams
- effective VQEs for encrypted content
- effective VQEs for the upcoming roll-out of stereo 3-D and other forms of "3-D video."

Signal processing clearly has a prominent role to play in this future.

AUTHORS

John G. Apostolopoulos (john_apostolopoulos@hp.com) is director of the Mobile and Immersive Experience Lab within HP Labs.

Amy R. Reibman (amy@research.att.com) is a lead member of technical staff at AT&T Labs–Research.

REFERENCES

[1] D. Bauder. (2010, Sept. 2). Networks using Skype in news reports, Skype becoming popular tool for network news telecasts. ABC News, New York. [Online]. Available: http://abcnews.go.com/Entertainment/wireStory?id=11541154

[2] A. R. Reibman, V. Vaishampayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Trans. Multimedia*, vol. 6, pp. 327–334, Apr. 2004.

[3] S. Tao, J. Apostolopoulos, and R. Guérin, "Real-time monitoring of video quality in IP networks," *IEEE/ACM Trans. Networking*, vol. 16, pp. 1052–1065, Oct. 2008.

[4] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[5] S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process. Image Commun.*, vol. 25, pp. 469–481, Aug. 2010.

[6] A. M. Tekalp, A. Smolic, A. Vetro, and L. Onural, Eds.,"Special issue on 3-D media and displays," *Proc. IEEE*, vol. 99, pp. 529–741, Apr. 2011.

[SP]