Perceptual Issues of Streaming Video

Alan C. Bovik, Christos Bampis, Todd R. Goodall

Laboratory for Image and Video Engineering (LIVE)
The University of Texas at Austin, Austin TX

Abstract

The large-scale streaming of videos on demand, as exemplified by Netflix, Amazon, and YouTube, is a remarkable modern engineering achievement that embodies significant advances in such fields as video compression and communications, digital networks, high-speed computation, and display technologies. Yet even today, there remain significant challenges in providing the highest quality compressed digital video content to the consumer. We consider two of the main issues. The first is source inspection, whereby the intrinsic quality and possible impairments of source videos of interest are determined. The second is that of balancing the tradeoffs that can occur between video compression and rebuffering effects. We describe recent human studies that we have conducted on these problems and the types of automatic prediction models that we have been developing.

Author Keywords

Streaming video; source inspection; motion picture quality; picture quality prediction; video distortion; video rebuffering.

1. Introduction

The large-scale streaming of videos on demand, e.g., by Netflix, Amazon, and YouTube, is a remarkable modern engineering achievement embodying significant advances in video compression and communications, digital networks, high-speed computation, and display technologies. Yet even today, there remain significant challenges in providing the highest quality compressed digital video content to the consumer. Here we discuss two of the main issues.

The first is the problem of source inspection, whereby the intrinsic quality and possible impairments of source videos of interest are determined. This is of particular importance given the massive amount of legacy contents (e.g., older television programs and motion pictures) that are of low intrinsic quality and/or suffering from various artifacts arising from conversion from older formats to newer formats, such as upscaling, combing and aspect ratio conversion effects. We discuss methods that we have recently been developing to detect these effects automatically, which would be of great value for large-volume video streaming enterprises.

The second problem we discuss is that of balancing the tradeoffs that can occur between video compression and rebuffering effects. While it is understood that excessive compression leading to perceptual artifacts is undesirable, and that the degree of perceptual annoyance arising from it can be predicted reasonably accurately, the perceptual effect of rebuffering events (video stalls from buffer emptying) on the overall human Quality of Experience (QoE), and how they balance against choices made regarding compression, have been less well studied. Creating models that predict the degree of annoyance arising from combined compression-rebuffering scenarios could lead to effective automatic rate control protocols that minimize the likelihood of rebuffering events, while maintaining acceptable levels of compressed picture quality.

We also describe recent human studies that we have conducted on these problems and the types of automatic prediction models that we have been developing that we envision could lead to greatly enhanced solutions to both of these problems.

2. Video Quality Assessment Scenarios

The development of models that predict human judgments of motion picture or video quality is well-studied [1]-[3]. Algorithms that predict motion picture quality vary according to the relevant reference picture information, ranging from full-reference (FR) [3]-[6], to reduced-reference (RR) [7], to no-reference (NR) or blind [8]-[12]. NR algorithms, which do not use a reference video, instead rely on measuring the degree of statistical 'naturalness' that is lost by the introduction of distortion.

Full-reference (FR) video quality assessment (VQA) models and algorithms have gained the greatest currency in digital television and cinematic applications by their application to rate monitoring and control problems, whereby broadcast, cable, satellite, or Internet VOD video encoding is perceptually monitored by the highly perceptually-relevant VQA algorithms, to either predetermine rate (compression level) or on the fly, in real time. The Emmy-winning Structural Similarity model (SSIM) [4] is probably the most successful example, of this, as it is deployed globally by broadcast, cable and satellite television providers to monitor and control picture quality, thereby affecting the viewing experiences of millions of viewers on a daily basis. Given the high percentage of video content occupying global (wireline as well as wireless) bandwidths, SSIM has a tremendous effect on global bandwidth allocation and consumption. A more recent model, the MOVIE algorithm, is also deployed globally [6].

However, FR VQA models are less useful at other points along the streaming video pipeline. This is depicted in Fig. 1, which divides the pipeline into Source, Transmission, and Client.

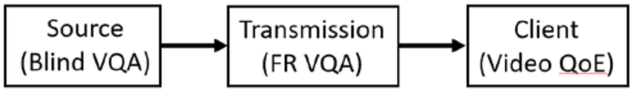

Figure 1. Stages of video quality analysis for streaming video.

The first (Source) stage involves the interesting and complex problem of video source inspection. A content provider, such as Netflix, acquires enormous amounts of content "sight unseen" from a large variety of sources. This content may be afflicted by any of a wide variety of distortions, such as compression artifacts, or by impairments arising from legacy processing, such as deinterlacing 'comb' effects, or by conversion errors arising from upscaling, aspect ratio conversion, 3-2 pulldown, and many other possibilities. Since no reference is generally available in any of these scenarios, then either human inspection (which is not feasible) or automatic blind VQA algorithms are required.

The second (Transmission) stage is already ably handled using existing full-reference VQA models, as discussed above.

The third (Client) stage presents the very difficult problem of handling new video impairments, variously called stalls or rebuffering events, that arise from of the interplay between available bandwidth at the receiver, the current video bitrate, and the status of the video buffer(s) in the client's reception device(s). Video stalls or freezes are severe impairments that greatly affect a viewer's visual Quality of Experience (QoE).

We discuss the first and third stages in the following sections.

3. Video Source Inspection

Videos obtained from diverse sources may include legacy content such as old movies or television programs that were created at lower resolution, digitized from analog form, compressed using an older codec, or subjected to interlacing or other degradations. All of these, when displayed on a modern high resolution screen, can affect the perceptual quality of the presentation to some degree. Some artifacts, such as those arising from upscaling videos from low resolutions to higher ones, may be less noticeable to a casual observer, while others such as "comb" distortions in deinterlaced videos may be quite obvious. However, distorted source videos generally share one characteristic: there isn't a high-quality reference video available to make quality comparisons against, hence NR VQA models are required.

An important development in the field of vision science that has evolved over the past few decades are highly regular, practical statistical models of photographic images and video. It is now well established that the bandpass responses of picture luminances and chrominances over space and time obey simple and reliable statistical laws. In an empirical study of natural photographs, Rudermann [13] observed that when pictures have their local mean luminances removed, which is a sort of bandpass process, followed by a divisive normalization by local energy (luminance standard deviation) invariably have resulting first-order empirical distributions (histograms) that are approximately Gaussian and significantly decorrelated. This result has been observed on a wide variety of bandpass / normalized image signals produced using steerable pyramids, wavelets, Gabor filters, and so on [14].

This is a significant perceptual relevance to this observation, since the visual signal is subjected to both bandpass processing [15] and divisive normalization of the bandpass signals at various stages of the visual pathway, including retina, lateral geniculate nucleus (LGN) and primary cortex. Cortical models of this type have been used to drive machine vision algorithms from early on [16], [17]. The vision system appears to have evolved to adapt the regular statistics of pictures [18] to produce highly efficient, sparse low-level image representations [19].

When images are visibly impaired, the impression of distortion is largely pre-attentive. Likewise, if a picture is impaired, this statistical regularity is destroyed, and when viewed, is strongly correlated with an annoying sense of distortion The statistical structure of distorted images is reflected in in the behavior of bandpass, normalized versions of them. This was first observed in [20] and used to create an FR IQA model called Visual Information Fidelity (VIF) which delivered top performance for several years, and which also explained, in part, the success of SSIM [21]. VIF has recently been introduced as the core FR IQA model for HDR video evaluation in the MPEG/ISO HDR Tools package [22]. The statistical regularity of photographs, and the loss of it by distortion, has also driven the development of top-performing NR IQA models [1], [8]-[12] which closely approach the performance of FR models [23], [24].

The simple and successful NSS-based NR IQA model BRISQUE [10] conducts quality prediction by computes mean-subtracted contrast normalized (MSCN) coefficients on images I :

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C}$$

where $C = 1$ and where $\mu(i, j)$ and $\sigma(i, j)$ are the local (gaussian-windowed) mean and deviation of luminance, respectively:

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_{k,l}(i,j)$$

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} (I_{k,l}(i,j) - \mu(i,j))^2}$$

This process strongly spatially decorrelates and gaussianizes high-quality picture data, but this is modified by distortion. Histograms of MSCN coefficients (and products of neighboring pairs of them) are fit to parametric generalized Gaussian densities, and the best-fitting parameters are used as quality-predictive features.

Distortions encountered in source inspection, such as upscaling artifacts, sometimes require a high degree of sensitivity to detect and assess. Towards this we improved the capabilities of BRISQUE by first pre-filtering images to be assessed by an orthogonal filter bank computed over 1 million natural image patches using principle component analysis (PCA). This set of 25 5x5 filters is depicted in Fig. 2.

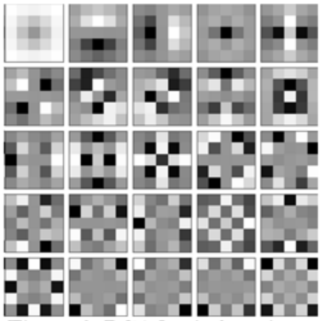

Figure 2. PCA basis functions.

Each of the 25 filtered images is subjected to MSCN processing, then the best parametric fits to the response histograms are found, yielding 125 parametric features [25].

As a simple visualization of the discrimination power of the PCA-BRISQUE features, Fig. 3 depicts boxplots of the sample variances computed on the MSCN maps and on the MSCN neighboring-product maps for all 25 PCA-filtered images, for different upscaling ratios (1x, 2x, and 3x). The plots indicate how these gross extracted energies become statistically well-separated when images are upscaled. Indeed, this simple energy measurement can be the basis of a very effective upscaling ratio predictor, although even better results are obtained by training on the 125 parametric NSS features. Indeed, even linear regression on the energies produces an effective predictor.

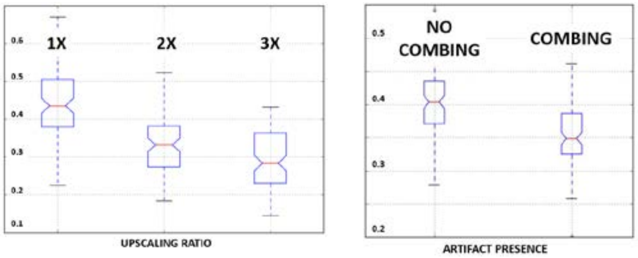

Figure 3. Boxplot of mean variance of MSCN maps: (left) against upscaling ratio and (right) against combing.

As we will show in the associated talk, the achieved prediction performance using the 125 NSS features is very high. Indeed, the accuracy obtained using a linear discriminant analyzer against multiple ratios of bilinear, bicubic, or Lanczos upscaling applied to frames obtained from a sizable Netflix database was very high (94%) as compared to the state-of-the-art model in [26], which reached only 67%. Even better result are obtained on combing artifacts, which tend to be more obvious.

4. Predicting Client QoE of Streaming Video

Given the volatile and increasingly crowded network conditions that feed videos to mobile devices, streaming content providers now deploy adaptive streaming strategies to mediate changing bandwidth conditions. At the client side, a viewer of streaming video may experience greatly reduced bandwidth, resulting in severe compression impairments, and/or video stalls/freezes (rebuffering events) from emptying of the client side buffer owing to a loss of adequate bandwidth. To study the perceptual effects of adaptive streaming on a client's subjective Quality of Experience (QoE), we designed and developed the new LIVE-NFLX QoE Video Database, which will soon be publicly released.

The new database uses a bandwidth usage equalization model, whereby it embodies various playout patterns that are allowed to use the same bandwidth and an equivalent buffer capacity. This allowed us to design directly comparable but highly diverse playout patterns, including constant encodes, adaptive rate drops, and mixtures of rate drops and rebuffering events.

We gathered almost 5000 human QoE subjective scores on 14 contents and 8 playout patterns. The 56 subjects viewed the impaired video contents on a mobile device. Longer video sequences (more appropriate for streaming applications) than are usual on VQA studies were used. The video content, either publicly available or provided by Netflix, is diverse and includes: action, drama, anime, and so on. We gathered both continuous and summary QoE ratings, which is enabling us to study both the temporal and overall (summary) effects of compression artifacts and rebuffering events on subject visual QoE, and how these types of impairments may be balanced against each other.

We found that rebuffering is always obvious and leads to significant, sharp drops in perceived QoE, while compression artifacts span a range of annoyances, depending on the content and compression level. Notably, there exists a threshold below which rebuffering is preferred over encoding a complex content (e.g. a sequence rich in motion) at a low bitrate.

An important observation was made with respect to recency effects: as expected, we found that summary QoE ratings were heavily biased by latest experiences. However, when a very negative QoE experience took place early on (such as a sequence of rebuffering events), it strongly impacted the reported endpoint QoE, i.e., subjects recalled them when making QoE judgments (a primacy effect). We also studied the temporal aspects of QoE, e.g., Fig. 4 shows the average subjective QoE reported during and following a sudden drop in bitrate. We have found that QoE recovery after a rate drop is a function of content complexity.

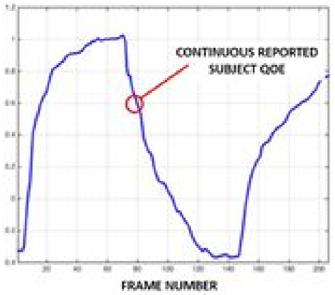

Figure 4. Plot of average reported continuous QoE of subjects viewing a video suffering a bit rate drop.

Analysis of the subjective results allowed us to define several QoE features: the video quality (distortion) during normal playback, the memory effects after video impairments occur, and the effects of video stalls (e.g. stall number, location, and density). We use these to train regressors to effectively predict QoE retrospectively. By using efficient VQA models such as SSIM [4] and ST-RRED [7], along with rebuffering and memory-related features, we have created retrospective QoE prediction engines that outperform the recent state-of-the-art. For example, Fig. 5 shows a scatterplot of predicted QoE scores using one of our models against human mean opinion (summary MOS) QoE scores on the new LIVE-NFLX QoE Video Database. Aside from one outlier, the relationship appears to be near-linear and well clustered and distributed. The linear correlation against MOS that was attained by this model using SSIM [4] as the VQA component was 0.75. By comparison, applying only SSIM on the database (where no stalls occurred) yielded a correlation of 0.63.

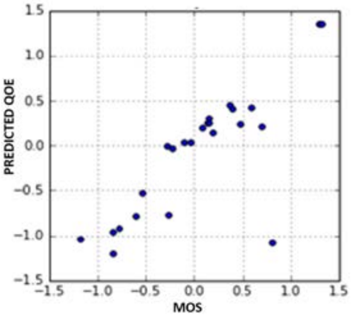

Figure 5. Plot of predicted QoE vs MOS.

The choice of VQA model used to predict the perceived effects of distortions caused by video compression and their ultimate effect on subject video QoE was important. As expected, PSNR delivered poor results (SROCC below 0.6) while SSIM [4] delivered good results. The best QoE prediction results were attained using the ST-RRED model [7], in line with other positive results obtained on this NSS-based algorithm.

5. Future Work

There are a number of exciting directions that we plan to pursue in the near future. In regards to source inspection, very often a content provider like Netflix may acquire content of a relatively poor quality but high intrinsic artistic merit, e.g., an old movie. The level of quality may reasonably be estimated using an NR VQA module. As with any content, decisions must be made with respect to bit rate, which further affects the quality of the video content, and which can be measured using an FR model. However, in this case the FR algorithm is relying on an imperfect reference, hence will yield perceptually less reliable results.

This suggests the development of a new, two stage framework of "Conditional NR-FR VQA,", whereby an NR source inspection is conducted on the reference I, which is then compressed to produce I_c. FR quality assessment of I_c is then conducted using I as a reference, with the result being conditioned on the predicted quality of I. This suggest a form of two-stage Bayesian prediction $Q_{FR}\{I, I_c \mid Q_{NR}[I] = q\}$ which would entail modeling of the conditional distribution, or equivalently, of the joint distribution of the perceptual qualities of I and I_c.

With regards to future work on streaming video QoE prediction, there are even more possibilities, of which perhaps the most obvious is to model and predict the continuous time variation of visual QoE [27]. In this way, QoE predictions could be used to adapt coding and buffer maintenance to the network conditions, and ultimately, to produce tools for controlling bandwidth usage. This could be approached as a probabilistic perceptual optimization, similar to SSIM-optimized image restoration [28] and compression / rate control [29], [30]. This idea of perceptually optimizing the video network, which is to say, most of the network, is a long-held desire that seems coming to fruition.

6. Acknowledgements

This work was supported by Netflix.

7. References

[1] A.C. Bovik, “Automatic prediction of perceptual image and video quality,” Proc. IEEE, **101**(**9**), 2008-2024 (2013).

[2] A.K. Moorthy and A.C. Bovik, “Visual quality assessment algorithms: What does the future hold?” Multimedia Tools and Applications, **51**(2), 675-696, 2011.

[3] K. Seshadrinathan and A.C. Bovik, “A structural similarity metric for video based on motion models,” IEEE Int’l Conf. on Acoustics, Speech and Signal Processing (2007).

[4] Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Transactions on Image Processing, **13**(**4**), 600-612 (2004).

[5] Z. Wang, L. Lu and A.C. Bovik, “Video quality assessment based on structural distortion measurement,” Signal Process. Image Commun., **19**(2), 212-132 (2004).

[6] K. Seshadrinathan and A.C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” IEEE Transactions on Image Processing, **19**(2), 335-350 (2010).

[7] R. Soundararajan and A.C. Bovik, “Video quality assessment by reduced reference spatio-temporal entropic differencing,” IEEE Trans Circ. Syst. Video Technol., **23**(4), 684-694 (2013).

[8] M. Saad and A.C. Bovik, “Blind prediction of natural video quality,” IEEE Transactions on Image Processing, **23**(**3**), 1352-1365 (2014).

[9] A.K. Moorthy and A.C. Bovik, “A two-step framework for constructing blind image quality indices,” IEEE Signal Process. Lett, **17**(5), 513-516 (2010).

[10] A. Mittal, A.K. Moorthy and A.C. Bovik, “No-reference image quality assessment in the spatial domain,” IEEE Trans. on Image Processing, **21**(12), 4695-4708, (2012).

[11] A. Mittal, G.S. Muralidhar, J. Ghosh and A.C. Bovik, “Blind image quality assessment without human training using latent quality factors,” IEEE Signal Processing Letters, **19**(2), 75-78 (2013).

[12] A. Mittal, R. Soundararajan and A.C. Bovik, “Making a ‘completely blind’ image quality analyzer,” IEEE Signal Processing Letters, **21**(**3**), 209-212 (2013).

[13] D.L. Ruderman, “The statistics of natural images,” Network: Comput. Neural Syst., **5**(4), 517–548 (1994).

[14] M.J. Wainwright and E.P. Simoncelli, “Scale mixtures of Gaussians and the statistics of natural images,” Adv. Neural Info Process Syst **12**, MIT Press, Cambridge MA (2000).

[15] M. Clark and A.C. Bovik, “Experiments in segmenting texton patterns using localized spatial filters,” Pattern Recognition, **22**(**6**), 707-717 (1989).

[16] A.C. Bovik, M. Clark and W.S. Geisler, “Multichannel texture analysis using localized spatial filters,” IEEE Trans. Pattern Anal. Machine Intell., 12(1), 55-73 (1990).

[17] I.K. Sethi and A.K. Jain, Artificial Neural Networks and Statistical Pattern Recognition, Elsevier (1991).

[18] D.J. Field, “Relations between the statistics of natural images and the response properties of cortical cells,” J. Opt. Soc. Am. A, **12**(4), 2379-2394 (1987).

[19] B.A. Olshausen and D.J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” Nature, **381**(6583), 607-609, 1996.

[20] H.R. Sheikh and A.C. Bovik, “Image information and visual quality,” IEEE Trans. Image Process., **15**(**2**), 430-444 (2006).

[21] K. Seshadrinathan and A.C. Bovik, “Unifying analysis of full reference image quality assessment,” IEEE Int’l Conf Image Process., San Diego, CA (2008).

[22] H.R. Tohidypour, M. Azimi, M.T. Pourazad and P. Nasiopoulos, “Software implementation of visual information fidelity (VIF) for HDRTools,” JCT-VC of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JCTVC-W0115, MPEG # m377950, San Diego, CA (2016).

[23] W. Xue, X. Mou, L. Zhang, A.C Bovik and X. Feng, “Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features,” IEEE Trans. Image Process., **23**(11), 4580-4862 (2014).

[24] L. Zhang, L. Zhang and A.C Bovik, “A feature-enriched completely blind image quality evaluator,” IEEE Trans. Image Process., **24**(8), 2579-2591 (2015).

[25] T.R. Goodall, I. Katsavounidis, Z. Li, A. Aaron and A.C. Bovik, “Blind picture upscaling ratio prediction,” IEEE Signal Process. Lett., **23**(12), 1801-1805 (2016).

[26] X. Feng, I. J. Cox and G. Doerr, “Normalized energy density-based forensic detection of resampled images,” IEEE Trans. Multimedia, **14**(3), 536-545 (2012).

[27] C. Chen, L.K. Choi, G. de Veciana, C. Caramanis, R.W. Heath and A.C. Bovik, “Modeling the time—varying subjective quality of HTTP video streams with rate adaptation,” **23**(5), 2206-2221, 2014.

[28] S.S. Channappayya, A.C. Bovik, C. Caramanis, and R.W. Heath, “Design of linear equalizers optimized for the structural similarity index,” IEEE Trans. Image Process., **17**(**6**), 857-872 (2008).

[29] S.S. Channappayya, A.C. Bovik, and R.W. Heath, “Rate bounds on SSIM index of quantized images,” IEEE Trans. Image Process., **17**(**9**), 1624-1639 (2008).

[30] T.S. Ou, UY.H. Huang and H.H Chen, “SSIM-based perceptual rate control for video coding,” IEEE Trans. Circ. Syst. Video Technol., **21**(5), 682-691 (2011).