Layered Video Coding Offset Distortion Traces for Trace-Based Evaluation of Video Quality after Network Transport

Patrick Seeling, Martin Reisslein, and Frank H.P. Fitzek

Abstract

Currently available video traces for scalable encoded video with more than one layer are a convenient representation of the encoded video for the evaluation of networking mechanisms. The video distortion (RMSE) or quality (PSNR) for individual video frames in these traces, however, only allow for the calculation of the video quality of correctly received video frames; for lossy network transport, only a rough approximation can be made. With the availability of *scalable offset distortion* traces, which we introduce and evaluate in this paper, networking researchers are enabled to accurately calculate the video quality of scalable encoded video as it is perceived by the receiving client after lossy network transport.

I. Introduction

In the future Internet, multimedia applications and services are expected to account for a large portion of the overall traffic. Among the different forms of multimedia, video data presumably account for a major fraction of multimedia data transported over networks. Video is typically encoded before transport over networks to save on the required bandwidth. For networking research in the area of video transmission, the encoded video can be represented by (i) the encoded bit stream, (ii) video traces, or (iii) a model. While size, copyright issues, and requirements on equipment and experience for video encoding are typical problems associated with research based on the actual encoded data, video traces and models are more convenient in their utilization in networking research. Accurate and parsimonious video traffic models, however, are still an ongoing research issue. Video traces provide an appealing approach for conducting research on the transmission of video. Video traces are typically in simple text format and carry only the video frame sizes and the video frame qualities. In contrast to encoded video data, video traces do not carry the actual video information and are therefore exchangeable among researchers without copyright issues. Another benefit of video traces is that no special equipment is needed; video traces can be employed in network simulators, widely used in networking research. Video traces have evolved from simple frame size traces to traces that contain video qualities and more information [1]. Networking research has taken advantage of the availability of these traces, see e.g., [2]–[10].

For networking research, the frame loss probability, which is defined as the long run fraction of frames that miss their playout deadline at the receiver, can be easily determined. To determine the video quality, however, subjective tests or objective metrics have to be applied to video bit streams. The mean opinion score (MOS) [11] for the evaluation of the video quality requires several test subjects for a single transmitted video which is impractical for utilization in networking research. The objective video quality is typically measured in terms of the root mean square error (RMSE) and the peak signal to noise ratio (PSNR), which is computed from the RMSE. (Throughout this paper we refer to the RMSE as *distortion* and to the PSNR as *quality*.)

The determination of the video quality perceived by the recipient(s) without any losses can be conveniently accommodated with conventional video traces [1]. Most video transport mechanisms, however, incorporate bandwidth limitations or loss probabilities which result in frame losses. For single layer (non-scalable) video, offset distortion traces have recently become available to allow for accommodation of video frame losses [12] without requiring explicit experiments with the encoded video [13]–[15] or general approximations. Scalable video encoding mechanisms and corresponding video streaming mechanisms, however, use multiple layers that add to the video quality. This popular type of encoded video cannot be accommodated in the currently available video traces.

Conventional scalable video coding encodes the source video into hierarchically organized layers: a base layer and one or more enhancement layers. A basic video quality can be achieved if the base layer can be successfully decoded. Adding one or more enhancement layers to the decoding process increases the quality of the decoded video. We consider temporal and spatial scalability encodings with one enhancement layer here; data partitioning-based and signal-to-noise ratio based scalability or multiple enhancement layers can be accommodated in a similar fashion. Temporal scalable encodings encode the base and enhancement layers by interleaving them. The base layer of temporal scalable encoded video thus provides a lower frame rate at the resolution of the encoding. Adding the enhancement layer increases the frame rate. Spatial scalable video encoding provides a low resolution version of the video when only the base layer is received. Upsampling is used to display the received

P. Seeling is with the Dept. of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, patrick.seeling@asu.edu

M. Reisslein is with the Dept. of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, reisslein@asu.edu

F. H.P. Fitzek is with the Dept. of Communication Technology, Aalborg University, Aalborg, Denmark, DK-9220, ff@kom.aau.dk

base layer in full (or enhancement layer) resolution. If the enhancement layer is received in addition to the base layer, the full resolution video can be displayed. In addition to the loss of frames, which is typically concealed by re-display of the last decoded frame, other combinations of received layers add to the complexity of determining the video quality as perceived by the client for scalable video.

In this paper, we introduce and evaluate scalable offset distortion traces. These traces, when combined with currently available video traces, enable networking researchers to meaningfully assess the perceived video quality for scalable video using only video traces. The scalable offset distortion traces contain the qualities of upsampled and re-displayed video frames. The impact of different video transmission outcomes (for base and enhancement layer(s) of the scalable video) on the video stream quality can be accurately determined using these scalable video traces.

II. Video Frame Quality

The objective video quality is typically calculated as peak signal to noise ratio (PSNR) between the unencoded original video data and the encoded and subsequently decoded video data. The PSNR is calculated using the root mean squared error (RMSE) between the pixels of the unencoded and the encoded and subsequently decoded video frame. Each individual pixel is represented by an 8-bit value for the luminance (Y) component, and a sub-sampled version of the image is used to store the two 8-bit values for the chrominance components hue (U) and intensity (V). Typically only the luminance component is taken into consideration for the calculation of the RMSE and PSNR, as the human eye is most sensitive to this component [16]. Let q denote the quantization scale (which relates inversely to quality) for an arbitrary video encoding and let N denote the total number of video frames in the video stream. We denote an individual pixel's luminance value in the nth original video frame at position (x, y) as $F^q_n(x, y)$ and its encoded and subsequently decoded counterpart by $f^q_n(x, y)$. Let X and Y denote the resolution in pixels of the source video. We calculate the video frame distortion as RMSE for all the luminance differences of an individual frame n encoded with the quantization scale q as

$$RMSE^q_n = \sqrt{\frac{1}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}[F^q_n(x,y) - f^q_n(x,y)]^2}. \tag{1}$$

The video frame quality as PSNR can be calculated from the RMSE as

$$Q^q_n = 20\log_{10}\frac{255}{RMSE^q_n}. \tag{2}$$

With the N frames in a given video stream, we calculate the average video quality or video stream quality as

$$\overline{Q}^q = \frac{1}{N}\cdot\sum_{n=1}^{N} Q^q_n \tag{3}$$

and the variability of the video frame qualities measured as standard deviation as

$$\sigma^q = \sqrt{\frac{1}{(N-1)}\sum_{n=1}^{N}(Q^q_n - \overline{Q}^q)^2}. \tag{4}$$

We calculate the corresponding distortion metrics in analogous manner. The video stream quality is generally maximized if the quality of individual frames is maximized and the variability of the quality among the frames of a video stream is minimized [17]. For scalable encodings with one base and one or more enhancement layers the quality can thus vary not only due to the encoding process, but also with the availability of the individual layers at the encoder.

III. Assessing Impact of Lost Frames with Video Bit Stream or through Approximation

In this section, we describe how the video quality for temporal and spatial scalable video with one base and one enhancement layer can be evaluated through experiments with the actual video bit stream or through approximation.

A. Temporal Scalable Video

We consider a basic temporal scalability scheme with an *IBBPBBPBB…* GoP pattern. For such a GoP pattern, the B frames constitute the enhancement layer, as no other frame relies on them. In the example illustrated in Figure 1, the base layer consists of I and P frames and the reception of the base layer gives 1/3 of the original frame rate at the decoder. The enhancement layer B frames are encoded with respect to the preceding I or P frame and the succeeding I or P frame in the base layer. As illustrated, the loss of a base layer (reference) frame results in the loss of the referencing frames in the enhancement layer. Simultaneously, the loss of a frame in the base layer spreads to the following frames in the base layer until a new I frame is received and the reference thus updated. In the illustrated example, the loss of the P frame at position 7 causes the referencing B frames 5 and 6 in the enhancement layer to be lost. Similarly, the not illustrated following frames in the base and enhancement layers would be lost as well until a new reference frame can be sent. The decoder in this example now re-displays frame 4 in place of frames $5, 6, 7$, and the remaining frames in this GoP.

In case that the actual bit stream was available, the video distortion or quality for the video frames $5, 6, 7, \ldots$ could be calculated from comparison of the original (unencoded) video frames $5, 6, 7, \ldots$ with the encoded and subsequently decoded video frame 4. Without the availability of the original video, only a rough approximation for the not displayed video frames, e.g., $Q = 20$dB can be made.

B. Spatial Scalable Video

For spatial scalable video encodings, we assume that the enhancement layer has been encoded only with respect to the base layer. In addition, the enhancement layer could also be encoded using motion estimation and compensation techniques among the enhancement layer frames to increase the compression efficiency further at the expense of more inter-frame dependencies which can be accommodated in our offset distortion traces in analogous manner. For more than single enhancement layers, the same mechanisms described here apply analogously for the additional layers. For the receiver of the video with a CIF-sized display, which we consider to fix ideas, several different cases of receiving the base and enhancement layer can occur. Following the traditional layered video streaming approach, the available data at the receiver on a video frame basis can be (i) the base and enhancement layer, resulting in the full size display of the current video frame, (ii) the base layer only, resulting in the display of the upsampled (and possibly additionally filtered) base layer video frame, and (iii) neither base or enhancement layer, resulting in re-display of the last successfully decoded frame, which can be either a full-size enhancement layer video frame or an upsampled base layer video frame. We illustrate these different possibilities with an example in Figure 2. In this example, the enhancement layer data for frames 1–4 and the base layer data for frames 1–5 is available at the decoder. The decoder thus displays frames 1–4 in full enhancement layer resolution, frame 5 as upsampled base layer frame, and re-displays the upsampled base layer frame 5 for frames 6 and 7.

In case that the actual video bit stream was available, the video distortion measured as RMSE or video quality measured as PSNR would be calculated comparing the full resolution frames 1–4 of the original unencoded video (i.e., the video in enhancement layer resolution) with the encoded and decoded enhancement layer frames 1–4 and for frames 5–7 with the upsampled low resolution base layer frame 5. Without access to the actual video data, only a very rough approximation can be made. Using *offset distortion traces* [12] for the enhancement layer would allow to calculate the distortion or quality caused by re-display of an enhancement layer resolution frame (i.e., in case that the same number of base and enhancement layer frames are available at the decoder – in the example in Figure 2 this would occur if the enhancement layer frame 5 would be available at the decoder as well). However, if only the upsampled version of the last frame (before the loss), i.e., frame 5 in the illustrated example, is available, then the offset distortion traces of [12] will not allow for determining the video quality after the loss.

IV. Offset Distortion Traces for Scalable Video Coding

In this section, we introduce the offset distortion for scalable encoded video for temporal and spatial scalable video.

A. Temporal Scalable Video

For temporal scalability, the distortions caused by the re-display of the enhancement layer frames can be determined in the same manner as for video encoded into a single layer. In particular, our earlier research results showed that for open-loop encoded video, the temporal enhancement layer can be obtained by extraction of the B frames from a single layer encoding [1].

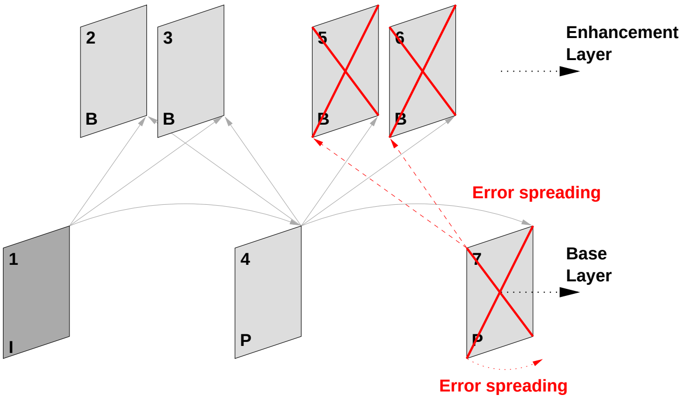

Fig. 1. Temporal scalable video with inter-frame dependencies and different error spreading possibilities.

For encoding with rate control, however, the base layer bit allocation is different from the allocation in case of a single layer encoding. For this particular case, the single layer traces cannot be used. Instead, the combined base and enhancement layers have to be considered in order to generate an offset distortion trace with all frames available for the calculation. With the thus calculated offset distortion trace, the qualities of the missing frames can be calculated analogously to the single layer case, see [12].

B. Spatial Scalable Video

Following the three loss scenarios outlined in Section III, networking researchers require several different (offset) video traces to accurately determine the impact of lost video frames at the decoder. For the case of two-layered spatial scalable video, which we consider here as an example, four different video traces are needed to accommodate the different scenarios possible at the receiver's decoder. Let d denote the offset in frames between the last successfully decoded video frame and the frame n under consideration. The needed distortion and quality values in the case of spatial scalable two-layered video are:

1) The *traditional* video frame distortion or quality comparing the unencoded original enhancement layer resolution video frame with the encoded and subsequently decoded frame, denoted as $RMSE_{EL}^{n,q}(0)$ and $Q_{EL}^{n,q}(0)$.
2) The *upsampling* distortion or quality for a received base layer frame which compares the unencoded original enhancement layer resolution video frame with the encoded, decoded, and subsequently upsampled base layer frame, denoted as $RMSE_{UP}^{n,q}(0)$ and $Q_{UP}^{n,q}(0)$.
3) The *offset* distortion or quality for the enhancement layer frames, denoted as $RMSE_{EL}^{n,q}(d)$ and $Q_{EL}^{n,q}(d)$, $d \geq 1$.
4) The *scalable offset* distortion or quality which is obtained by upsampling and re-displaying the base layer video frame, denoted as $RMSE_{UP}^{n,q}(d)$ and $Q_{UP}^{n,q}(d)$, $d \geq 1$.

We now take a closer look at the calculation of the individual video distortion or quality value calculations.

The traditional video frame distortion or quality is calculated using Equations (1) and (2) and can be obtained from current video traces. The upsampling distortion caused by upsampling the base layer frame in low resolution to the enhancement layer frame high resolution can also be obtained from the current layered video traces available at [18]. The old traces give the video frame quality in terms of the PSNR (from which the RMSE values can be calculated according to 2) for the upsampled base layer frames. The enhancement layer trace contains the quality improvement in PSNR when both the base and the enhancement layer are available at the decoder. The offset distortion or quality values are obtained from offset distortion traces, which have recently been developed for inclusion into current video traces, see [12]. All these traces, however, only enable to determine the (upsampled) base layer qualities and the enhancement layer qualities. For cases where the base layer was only partially received, such as in Figure 2, these traces do not allow the calculation of the video distortion or quality for re-displaying the base layer frame at the enhancement layer resolution. To determine these distortions and qualities correctly, new *scalable offset distortion* traces are required.

The calculation of the RMSE for the re-display of the upsampled base layer frame n instead of an enhancement layer frame at position $n + d$ is given as function of the frame offset d similar to Equation (1) as

$$RMSE_{UP}^{n,q}(d) = \sqrt{\frac{1}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}[F_{EL}^{n+d,q}(x,y) - f_{UP}^{n,q}(x,y)]^2} \quad (5)$$

where $f_{UP}^{n,q}$ denotes the encoded (at quantization scale q) and upsampled base layer frame n and $F_{EL}^{n+d,q}$ denotes the original unencoded enhancement layer frame at offset d. The corresponding video frame quality can be calculated similar to Equation (2)

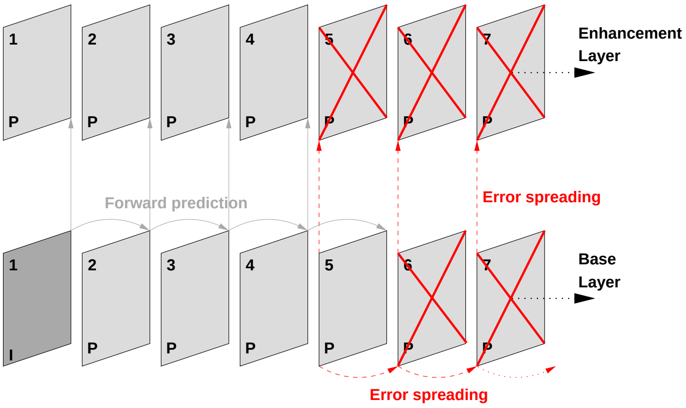

Fig. 2. Spatial scalable video with inter-frame dependencies and different error spreading possibilities.

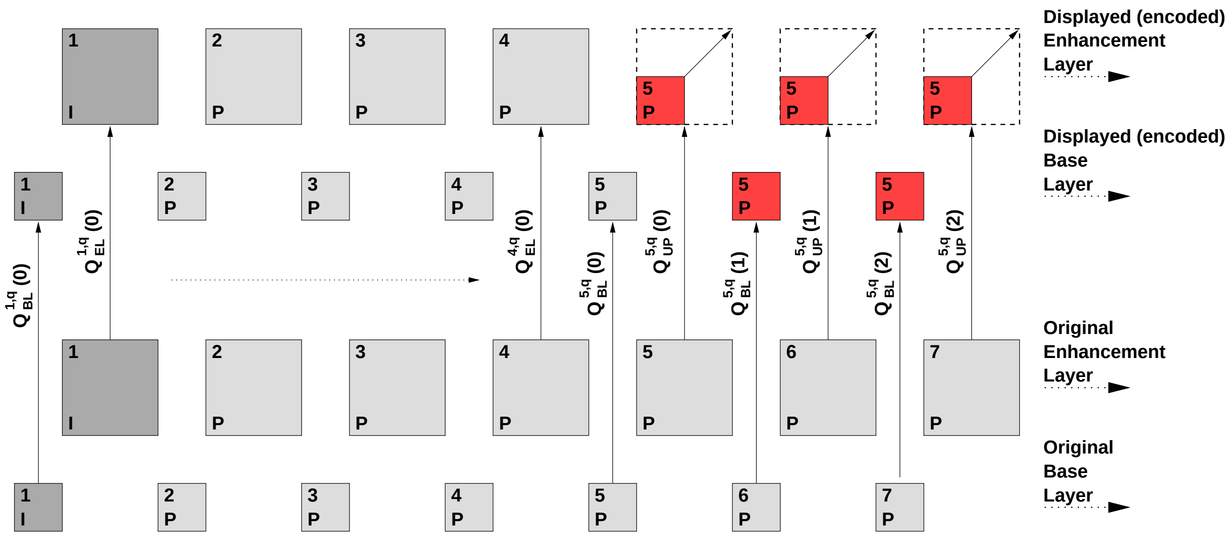

Fig. 3. Spatial scalable video with 2 layers after erroneous transmission with corresponding video frame quality values $Q^{n,q}$.

as

$$Q^{n,q}_{UP}(d) = 20 \log_{10} \frac{255}{RMSE^{n,q}_{UP}(d)}. \tag{6}$$

These values can be stored in a plain text file where for each encoded base layer frame n a row indexed with n contains the scalable offset distortion information $RMSE^{n+d,q}_{UP}$ in column d. As in general, transmission errors can be healed after a certain number of frames, accordingly we calculate the scalable offset distortion up to a maximum of $d = d_{\max}$ frames. We continue the example illustrated in Figure 2 showing the different quality values needed in Figure 3. As illustrated, the quality (and similarly the distortion) values for frames 1–4 in the enhancement layer resolution can be obtained from the current video traces. For the enhancement layer frame 5, only the base layer data is available at the decoder. Thus the decoder upsamples (and possibly filters) the base layer frame prior display at the client. This distortion value can be obtained from the readily available video traces at [18] as $RMSE^{n,q}_{UP}(0)$. For frames 6 and 7, no base layer information is available. The decoder thus re-displays the last frame in memory, which in this case is the upsampled base layer frame 5. The distortions caused by upsampling and re-displaying $RMSE^{5,q}_{UP}(1)$ and $RMSE^{5,q}_{UP}(2)$ are not part of any available video trace and thus limit the evaluation of the video quality to rough approximations. Using the *scalable offset distortion* traces, which we introduce and evaluate in this paper, this information becomes available to networking researchers, who in turn can calculate the distortion and quality for layered video in this manner.

We illustrate the *offset distortion* for the base layer resolution and the enhancement layer resolution as well as the *scalable offset distortion* for frame 100 from the *News* video sequence in Figure 4. We observe that the enhancement layer resolution offset distortion values closely follow the base layer resolution values. In addition, we observe that the scalable offset distortion values exhibit a similar characteristic, albeit on a higher level of distortion. Closer examination reveals that the difference in distortion, however, is not monotonous, but declining as the offset in frames increases. We examine the differences in between

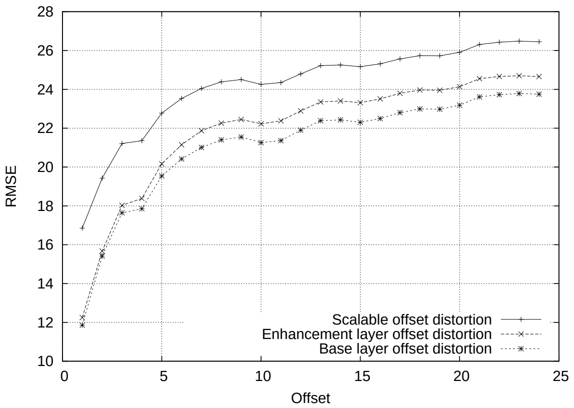

Fig. 4. Offset distortion values for frame 100 from the *News* sequence encoded with quantization scale $q = 9$.

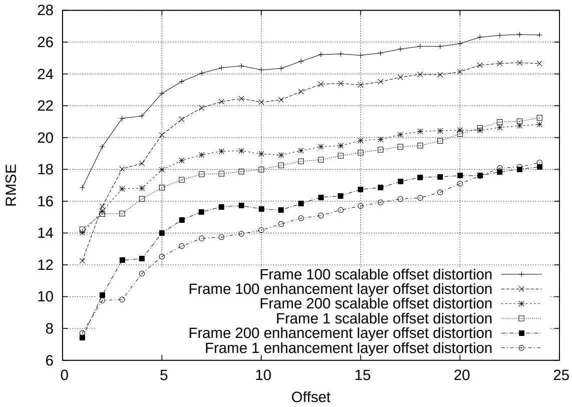

Fig. 5. Scalable offset distortion and enhancement layer resolution offset distortion values for frames 1, 100, and 200 from the *News* sequence encoded with quantization scale $q = 9$.

different frames for the enhancement layer resolution offset distortion and the scalable offset distortion in Figure 5. We observe that the scalable offset distortions for all frames follow the characteristic shapes of the enhancement layer resolution offset distortion. We observe additionally that the differences in between the two offset distortions vary by frame and offset, which in turn shows that assuming fixed distortion (or quality) values does not capture these behaviors at all.

V. Influence on Simulation Results

In this section, we evaluate the general impact of the scalable offset distortion on simulation results using the video sequence *News* encoded with the official Microsoft MPEG-4 reference encoder [19]. We use a QCIF-sized base layer and a CIF-sized enhancement layer. We employ the GoP pattern *IPPP*... using a GoP length of 24 frames and a frame rate of 24 frames per seconds (i.e., we start a new GoP every second). The enhancement layer is encoded only with respect to the base layer (i.e., we do not enable motion estimation and compensation between enhancement layer frames) in our evaluation, but we note that different encoding schemes can be regarded in a similar manner. Without loss of generality, we evaluate lossless transmission of the encoded video over a bandwidth-limited link to illustrate the difference between a rough approximation using a low value of $Q = 20$db or $RMSE = 25.5$ for the video quality versus the actual value determined by the scalable offset distortion trace. We consider the standard layered video streaming approach, where the base layer is transmitted before the enhancement layer in combination with the typically used RTP/UDP/IP protocol encapsulation.

We illustrate the influence on the video distortion for the enhancement layer (EL) resolution in Figure 6 for the *News* sequence encoded with a quantization scale parameter of $q = 9$. We observe that the enhancement layer distortion is significantly higher for approximation and calculation of the distortion values over a wide range of evaluated bandwidths. The reason for this behavior is that the encoded base layer requires 73.2 kbps on average, whereas the enhancement layer requires 1038.9 kbps on average. In addition to this effect, the intra coded I frames in the base layer and the larger sized enhancement layer frames require packetization into multiple IP packets, which adds an additional protocol overhead.

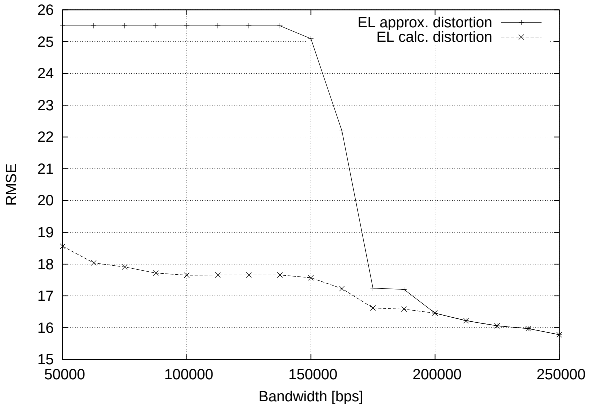

Fig. 6. Mean distortion from approximation and calculation for spatial scalable video streaming of the *News* video sequence.

Only if most of the base layer can be transmitted successfully, the approximated distortion for the enhancement layer approaches the level of the calculated distortion, as we can determine the distortion for the upsampling from the current traces.

Importantly, we note that the approximation of the distortion with a fixed value results in a too high estimate of the distortion. A simple approximation with fixed values is thus not desirable.

For the variability of the video distortion, we illustrate the standard deviation for the two layers in Figure 7. We observe that the approximation of the distortion values decreases the variability significantly compared to the actual variability. The approximation does not capture the behavior of the calculated variability, instead the variability is approximated as too low for very low bit rates and as too high as the bit rate increases.

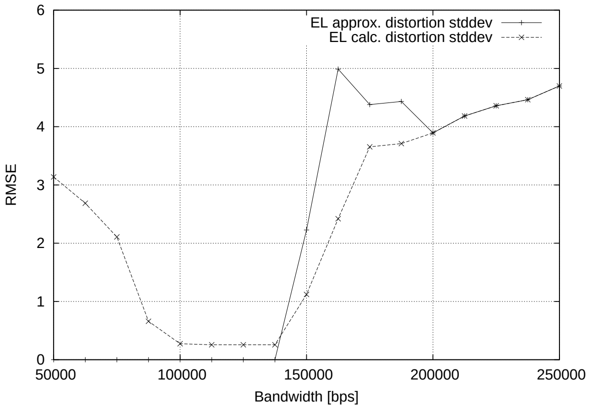

Fig. 7. Standard deviation of the distortion from approximation and calculation for spatial scalable video streaming of the *News* video sequence.

Overall, we find that using an approximation value instead of calculated values increases the introduced error in video quality estimation on a trace basis quite significantly. We conclude that utilizing the scalable video distortion and quality traces we are currently including into our video trace library [18] results in an accurate estimation of the video quality after (lossy) network transport.

VI. Conclusion

In this paper, we reviewed currently available video traces and their limited application in determining the video quality after lossy network transport for scalable video encodings. Scalable offset distortion traces, which we introduced and evaluated in this paper, in combination with the other currently available video traces allow networking researchers to calculate the video distortion or quality for the different combinations of base and enhancement layer frames available to the decoder. We exemplarily illustrated the impact of using the scalable offset distortion traces on simulation results where we found significant differences between using these traces and using an approximative value. Scalable offset distortion traces thus enable networking researchers to conduct trace-based evaluation of networking mechanisms and to accurately determine the perceived video quality at the receiver.

Acknowledgement

Supported in part by the National Science Foundation under Grant No. ANI-0136774.

References

[1] P. Seeling, M. Reisslein, and B. Kulapala, "Network performance evaluation with frame size and quality traces of single-layer and two-layer video: A tutorial," *IEEE Communications Surveys and Tutorials*, vol. 6, no. 3, pp. 58–78, Third Quarter 2004.

[2] M. Dai and D. Loguinov, "Analysis and modeling of MPEG-4 and H.264 multi-layer video traffic," in *Proceedings of IEEE INFOCOM*, Miami, FL, Mar. 2005.

[3] W.-C. Feng, *Buffering Techniques for Delivery of Compressed Video in Video–on–Demand Systems*. Kluwer Academic Publisher, 1997.

[4] P. Koutsakis and M. Paterakis, "Call-admission-control and traffic-policing mechanisms for the transmission of videoconference traffic from MPEG-4 and H.263 video coders in wireless ATM networks," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1525–1530, Sept. 2004.

[5] M. Krunz and S. Tripathi, "Exploiting the temporal structure of MPEG video for the reduction of bandwidth requirements," in *Proc. of IEEE Infocom*, vol. 1, no. 1, Kobe, Japan, Apr. 1997, pp. 67–74.

[6] M. Krunz, R. Sass, and H. Hughes, "Statistical characteristics and multiplexing of MPEG streams," in *Proceedings of IEEE INFOCOM*, Boston, MA, Apr. 1995, pp. 455–462.

[7] J. Liebeherr and D. Wrege, "Traffic characterization algorithms for VBR video in multimedia networks," *Multimedia Systems*, vol. 6, no. 4, pp. 271–283, July 1998.

[8] J. W. Roberts, "Internet traffic, QoS, and pricing," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1389–1399, Sept. 2004.

[9] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modelling in ATM systems," University of Wuerzburg, Institute of Computer Science, Tech. Rep. 101, Feb. 1995.

[10] U. Sarkar, S. Ramakrishnan, and D. Sarkar, "Study of long-duration MPEG-trace segmentation methods for developing frame-size-based traffic models," *Computer Networks*, vol. 44, no. 2, pp. 177–188, Feb. 2004.

[11] ITU-T Recommendation P.800.1, "Mean opinion score (MOS) terminology," Mar. 2003.

[12] P. Seeling, M. Reisslein, and F. H.-P. Fitzek, "Offset distortion traces for trace-based evaluation of video quality after network transport," in *In Proc. International Conference on Computer Communication and Networks (ICCCN)*, Oct. 2005.

[13] W. Luo and M. ElZarki, "Analysis of error concealment schemes for MPEG–2 video transmission over ATM based networks," in *Proceedings of SPIE Visual Communications and Image Processing 1995*, Taiwan, May 1995, pp. 102–108.

[14] ——, "MPEG2Tool: A toolkit for the study of MPEG–2 video transmission over ATM based networks," Department of Electrical Engineering, University of Pennsylvania, Tech. Rep., 1996.

[15] J. Shin, J. W. Kim, and C.-C. J. Kuo, "Quality–of–service mapping mechanism for packet video in differentiated services network," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 219–231, June 2001.

[16] S. Winkler, "Vision models and quality metrics for image processing applications," Ph.D. dissertation, EPFL, Switzerland, 2000.

[17] T. Lakshman, A. Ortega, and A. Reibman, "VBR video: Tradeoffs and potentials," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 952–973, May 1998.

[18] "Video traces for network performance evaluation," Traces available at: http://trace.eas.asu.edu.

[19] Microsoft, "Mpeg-4 visual codec version 2.5.0," Aug. 2004.