INTERLACING INTRAFRAMES IN MULTIPLE-DESCRIPTION VIDEO CODING

Ermin Kozica, Dave Zachariah and W. Bastiaan Kleijn

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)
Stockholm, 10044 Sweden

ABSTRACT

We introduce a method to improve performance of multiple-description coding based on legacy video coders with pre- and postprocessing. The pre- and post-processing setup is general, making the method applicable to most legacy coders. For the case of two coders, a relative displacement of the intra-coding mode between the coders is shown to give improved robustness to packet loss. The optimal displacement of the intra-coding mode is found analytically, using a distortion minimization formulation where two independent Gilbert channels are assumed. The analytical results are confirmed by simulations. Tests with an H.263 coder show significant improvement in YPSNR over equivalent systems with no relative displacement of the intra-coding operation.

Index Terms— Video coding, multiple description, legacy coder, packet loss, robustness

1. INTRODUCTION

Video coding standards from about a decade ago, e.g., H.263 and MPEG-2, were developed for the purpose of efficient video compression. Their applications have mainly included communication over circuit-switched networks, e.g., videoconferencing and HDTV, but also storing video material for later retrieval, e.g., the DVD. Newer standards, e.g., MPEG-4 and H.264, have a performance that is significantly improved over their predecessors and are very efficient in achieving low bit-rates for given video quality. We will denote these coders, whose main purpose is efficient compression, legacy coders.

Legacy coders are, however, seldom suited for transmission over unreliable networks in general and packet-switched networks in particular. The reason for this is the non-zero probability of packet loss, in combination with the dependency amongst packets. A single packet loss often affects the video signal for several seconds. Adaptation of the coders to the new environment is crucial, since failure to overcome transmission errors generally results in annoying erroneous video being displayed or, in the worst case, total communication breakdown.

To overcome the challenges of live video communication over packet-switched networks, a pre- and post processing setup, in which the input video is divided into multiple descriptions that are coded in separate legacy coding units, may be deployed. The received descriptions at the receiver side are merged to produce the final result. Such an approach to the problem has its advantages and drawbacks. The minimal development time and the complete reuse of legacy systems minimize the development costs. The drawback on the other hand is, as with most solutions on a high level of implementation, that the performance obtained is not optimal.

The pre-processing parts of previously proposed systems differ in how the multiple descriptions are obtained. The methods can be divided into two classes: temporal or spatial subsampling. In temporal subsampling, the signal is subsampled in time. For instance, Apostolopoulos [1] assigns odd frames to description one and even frames to description two, respectively. A somewhat more general procedure is used in [2], where any combination of the frames is allowed. In spatial subsampling, on the other hand, the signal is subsampled in space. Different spatial techniques are described in [3], [4] and [5]. Some of these techniques include oversampling to increase the redundancy in the signal. Finally, Lotfallah [6] deploys a strategy that allows for switching between temporal and spatial subsampling, with the notion that temporal subsampling is more suitable for low-motion portions of the video, while spatial subsampling is more suitable for high-motion portions.

At the receiver side, the proposed systems decode received descriptions in respective decoders. Access to all descriptions renders the best quality after the post-processing step, which is the inverse of the pre-processing technique deployed. When packet losses occur, some descriptions are not available. These descriptions will be erroneous, due to error propagation, until they have been updated by an intra-coded frame. Let us define a corrupted description, as a description that either is not available or is erroneous. When there are corrupted descriptions, error concealment has to be performed. For temporal subsampling, some frames are not available for display, resulting in a jerky viewing experience. For spatial subsampling, the available descriptions are used to estimate the lost description, resulting in a blurry image. These effects are naturally not desirable. The worst case scenario is when all descriptions are corrupted at the same time. The quality of the displayed video drops quickly and the errors propagate through time, resulting in severe visual artifacts.

In this paper, we investigate one feature of the proposed systems that has not been addressed so-far: the independence of the legacy coders. More specifically, we consider the relative displacement of the intraframes in each respective coder. By altering the relative displacement of the intraframes we can improve the quality of the displayed video, without increasing rate, complexity or jeopardizing possible standard compliance of the legacy coders. Finally, we analytically derive the optimal displacement of the intraframes for the two description case, with respect to given channel probabilities.

The paper is organized as follows. Section 2 describes the general framework of a multiple-description video coder, deploying pre- and post-processing with multiple legacy coders. Interlacing intraframes is described in Section 3. Simulation results are presented in Section 4. Section 5 concludes the paper.

2. SYSTEM FRAMEWORK

The system framework for the pre- and post-processing system with multiple legacy coders has been designed such that minimal information about the legacy coders used is needed. There only needs to exist a way to communicate the input video to the encoders and to extract decoded video from the decoders. Note that these strict constraints on the system framework minimize the need for interaction with the legacy coders and are, hence, very conservative. Previously proposed coding systems that can be fit into the framework wholly or with some modification can be found in [1-9].

Consider the framework depicted in Figure 1, where variations of the symbol x denote representations of a particular frame. The input video signal x_0 is divided in block a into (for the sake of simplicity) two descriptions x_i, $i \in \{1, 2\}$. Each description is coded by a legacy encoder f_i, that uses intra-coded frames (intraframes or I-frames) and inter-coded frames (interframes or P-frames). The coded signals are sent over the network on the two channels c_i. At the receiver side, the coded descriptions are decoded in decoders g_i, producing descriptions $\hat{x}_i$. These descriptions are merged in block b, to form a single approximation $\hat{x}_0$ of the original input.

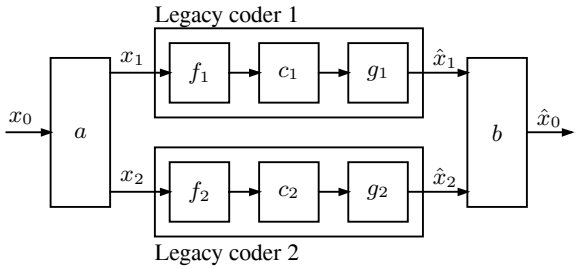

Fig. 1. Framework for the pre- and post-processing system with two legacy coders.

Depending on the realizations of the channels c_i, either of the two decoded descriptions $\hat{x}_i$, or both, may be corrupted in a particular frame. A corrupted description differs from its equivalence in the encoder f_i, which leads to prediction mismatch and error propagation. Hence, a corrupted description is corrupted until it has been updated with an error-free I-frame.

When either of the descriptions $\hat{x}_i$ is corrupted, or if both descriptions are corrupted, it is not possible to perform the inverse operation of the pre-processing block a. Hence, error concealment has to be applied in block b. To keep the need for access to the legacy coders minimal, the error concealment scheme that exists in the legacy coders is not tampered with. In case only one description is corrupted, the built-in error concealment of the decoders is not trusted. Thus, the merging block b does not take the corrupted description into account, but estimates the final result $\hat{x}_0$, from the received description. This is done until the corrupted description is updated with an error free I-frame. As mentioned earlier, both frames can be corrupted simultaneously. In that case, the description that was corrupted most recently is used for estimation of the final result.

Let us define D_0 as the expected distortion when none of the descriptions are corrupted and D_1 and D_2 as the expected distortion when $\hat{x}_2$ and $\hat{x}_1$ are corrupted, respectively. Further, let us denote the expected distortion when both descriptions are corrupted with D_T. The relation between expected distortions is for practical purposes

$$D_0 < D_1, D_2 \ll D_T. \tag{1}$$

3. INTERLACING INTRAFRAMES

In the previous section, the system framework was defined with the conservative assumption that the communication with the encoders is limited to the input video and that the communication with the decoders is limited to the output video. This results in total independency of the legacy coders, which also has been discussed in previous publications. However, there is the possibility of loosening the assumptions without affecting the integrity of the legacy coders. This can be done by assuming that the encoding mode (I or P) of the coders can be controlled, which is the case for most legacy coders. To the best of our knowledge, this opportunity for making the system robust to packet losses has not been discussed earlier in the literature.

For the purpose of illustration, we assume that the network consists of two independent Gilbert channels c_i, see Figure 2. The Gilbert channel model is convenient because it facilitates numerical optimization of the proposed pre-and post processing method. Further, it is a frequently used model for network packet-loss behavior [10]. Each channel has two states, state zero (0) and state one (1). When a channel is in state 0 the transmission is successful and when it is in state 1 the transmission fails. The probability that channel c_i is in state 1, given that the previous transmission was successful, is p_i. The probability that channel c_i is in state 0, given that the previous transmission failed, is q_i. To simplify notation below, let us define the steady state probability of state 0 as

$$\pi_i = \frac{q_i}{p_i + q_i} \tag{2}$$

and the probability of remaining in state 0 as

$$r_i = 1 - p_i. \tag{3}$$

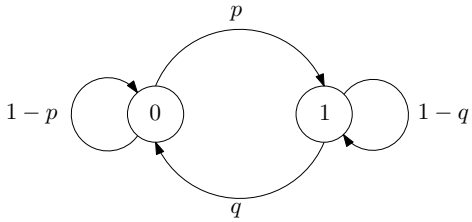

Fig. 2. The Gilbert two-state model of a bursty packet loss channel.

The freedom to choose the encoding mode of the legacy coders, in combination with the presented channel model, allows for optimization of the overall expected distortion of the final result $\hat{x}_0$. Assume that the group of pictures (GOP) length, i.e., the interval between two I-frames, for the coders is equal and set to K. With the traditional coding setup, the same frame will be coded in intra mode in the two coders, i.e., frames $\{\ldots, -K, 0, K, \ldots\}$. Altering the relative placement of the I-frame in one coder with respect to the other, yields the possibility for improvement of the overall expected distortion. Let us denote this displacement Δ, yielding intra-coding of frames $\{\ldots, -K, 0, K, \ldots\}$ in coder 1 and frames $\{\ldots, -K+\Delta, \Delta, K+\Delta, \ldots\}$ in coder 2. The expected distortion

at the receiver end then equals

$$\begin{aligned}\mathrm{E}[D] \quad = \quad & \frac{1}{2\kappa+1}\sum_{k=-\kappa}^{\kappa}\Bigg(D_0\pi_1 r_1^{\bmod\frac{k}{K}}\pi_2 r_2^{\bmod\frac{k-\Delta}{K}} \\ & + D_1\pi_1 r_1^{\bmod\frac{k}{K}}\Big(1-\pi_2 r_2^{\bmod\frac{k-\Delta}{K}}\Big) \\ & + D_2\pi_2 r_2^{\bmod\frac{k-\Delta}{K}}\Big(1-\pi_1 r_1^{\bmod\frac{k}{K}}\Big) \\ & + D_T\Big(1-\pi_1 r_1^{\bmod\frac{k}{K}}\Big)\Big(1-\pi_2 r_2^{\bmod\frac{k-\Delta}{K}}\Big)\Bigg), \quad (4)\end{aligned}$$

which is a distortion weighted summation of the probabilities of the possible outcomes, where $\bmod\frac{a}{b}$ denotes modulo b division of a, k denotes the frame index and κ is an arbitrary constant such that $K \ll \kappa$.

The optimal displacement Δ can be analytically calculated for the two-channel case. Since the nature of the modulo operation is cyclic, with a period of K, the summation can be taken over any interval of length K. By letting the displacement Δ be continuous, $\Delta \in [0, K)$, and the summation be approximated by an integration, the expression of the expected distortion is made continuous. The expression can be differentiated and the result set to zero, to determine where the function has its extremum. The displacement that gives the extremum is given by

$$\Delta_e = \frac{\ln\Big((r_1^K-1)r_2^K R\Big) - \ln\Big((r_2^K-1)(1-R)\Big)}{\ln(r_1)+\ln(r_2)}, \quad (5)$$

where $R \equiv \frac{\ln(r_2)}{\ln(r_1)+\ln(r_2)}$.

Two candidates for the optimal displacement are given by rounding the extremum displacement Δ_e up and down, respectively. Since the optimization is performed over an interval, two additional candidates are given by the interval boundaries. Hence, the set of possible optimal displacements is given by

$$\Delta \in \{0, \lfloor\Delta_e\rfloor, \lceil\Delta_e\rceil, K-1\}, \quad (6)$$

where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the round down and round up operations, respectively. Evaluation of the given set in Equation (4), gives the optimal displacement of intraframes for the case of two descriptions. The effect of the optimal displacement Δ on the performance of the coding system is investigated in the following section.

4. SIMULATION RESULTS

This section motivates and describes the setup of the performed simulations. Further, results that quantify the performance of the proposed technique are presented.

The optimization of the relative displacement Δ is done to minimize the expected distortion at the receiver side, given the channel probabilities. Hence, it is interesting to investigate how the channel probabilities affect the expected distortion, as well as the optimal displacement. For this purpose, the overall probability of packet loss was fixed to 10%. The channel probabilities q_i were fixed to 0.5, while the p_i were altered. Assuming that the expected distortions are given by the mean square error values

$$\{D_0, D_1, D_2, D_T\} = \{65, 205, 205, 1300\},$$

corresponding to PSNR values of $\{30, 25, 25, 17\}$ dB, the optimal displacements are illustrated in Figure 3, where the GOP length

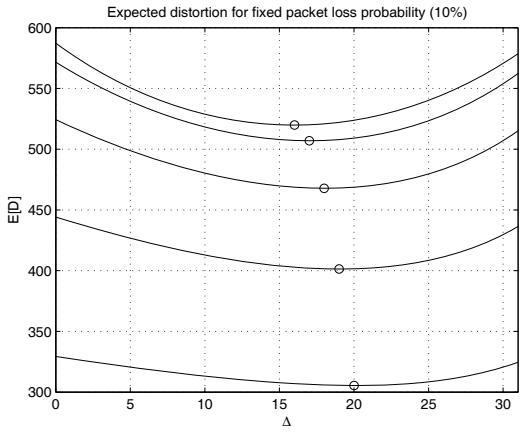

Fig. 3. Overall packet loss probability equals 10%. The packet loss probability of channel 1 is 2%, 4%, 6%, 8% and 10%, producing curves from bottom to top in the figure, respectively. The optimal displacement is illustrated with a circle for the different channel setups.

K was 32 frames. Equal loss probabilities for both channels, top curve, gives an optimal displacement of half the GOP length, i.e., $\Delta = K/2$. Improving the statistics of channel 1, while keeping the overall probability of packet loss fixed, the average distortion decreases and the optimal displacement increases. The increase of the optimal displacement is logical, channel 1 can be trusted for a greater number of frames, while channel 2 is used to improve performance where channel 1 is less reliable. The result is symmetric in the channel probabilities, i.e., the optimal displacement decreases if channel 2 is more reliable than channel 1. This is, however, not plotted for readability. Further, it is visible that the method holds greatest advantage, about 0.5 dB in PSNR, to traditional placement of the I-frames, $\Delta = 0$, when the channel probabilities are equal.

The probability that both descriptions are corrupted, i.e., the probability that a corrupted frame is displayed, is interesting, since these frames hold visual artifacts that the viewer detects instantaneously and finds the most annoying. Figure 4 illustrates this probability, for all frames in a GOP of length 32, for the traditional I-frame displacement, $\Delta = 0$, and the optimal I-frame displacement, $\Delta = 16$. The average probability is also plotted. The effect of the intraframe displacement on the probability of two corrupted descriptions is obvious. The probability is decreased by every new I-frame, while the minimum probability is slightly higher. As is shown in Figure 4, the average probability of both descriptions being corrupted is less when the I-frame displacement is optimal. In fact, the optimal I-frame displacement minimizes the average probability of simultaneously corrupted descriptions. This can be seen from the expression for the cost function in Equation (4), where each possible outcome is weighted by its expected distortion. Since the optimal displacement is not dependent of the expected distortion of the outcomes, the optimal displacement minimizes not only the expected distortion, but also the probability of corrupted display.

The YPSNR measure was used to objectively evaluate the performance of the proposed technique on the standard video sequence *foreman* in CIF format. For this purpose, a standard H.263 coder implementation was used. Two descriptions were generated with subsampling, by assigning odd pixels from odd rows to description 1

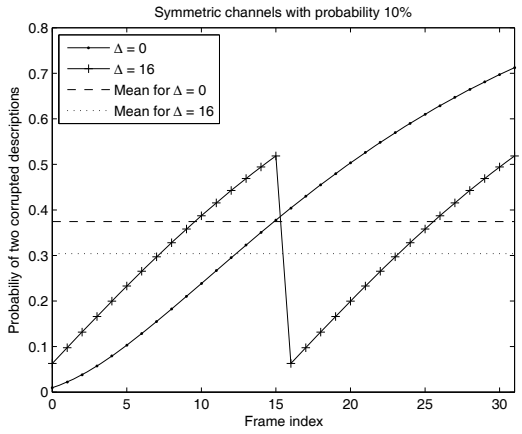

Fig. 4. Overall packet loss probability equals 10% and the channels have equal statistics, $p_i = 0.055$ and $q_i = 0.5$.

and even pixels from even rows to description 2. Zero-padding was not used. Each description was coded by an H.263 coder operating on the QCIF format at 200 kbps with a GOP length of 32 frames. The overall frame rate of the coder was chosen to 15 frames per second. When packetizing, the descriptions were kept in separate packets of the maximum packet size 1500 bytes. The channel realizations were based on traces from network routers. The probability of channel erasure was varied to assess its effect on the performance. Note that the assumption of two independent channels was dropped in the simulations, since deployment of such a system in practice is not straightforward. Figure 5 illustrates how the average YPSNR changes for different packet loss probabilities, for the traditional and optimal displacement of the I-frames. It is clear that the optimal placement of the I-frames outperforms the traditional by up to 1 dB.

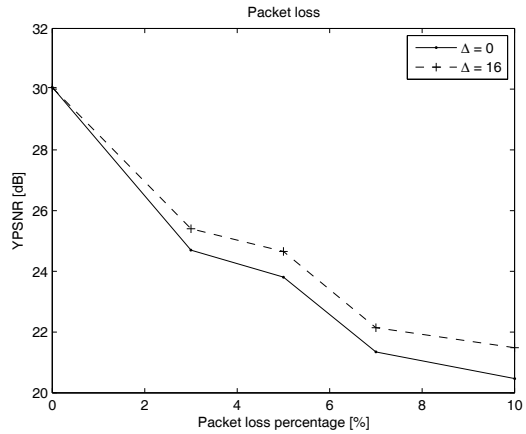

Fig. 5. The average YPSNR for the *foreman* sequence in CIF format. For each packet loss probability, 100 simulations were run with the traditional and the optimal I-frame displacement.

5. CONCLUSIONS

Interlacing intraframes in multiple-description video coding is a technique for optimal displacement of the intra-coding modes of multiple coders. Displacement of the intra-coding modes is, besides input video to encoders and output video from decoders, one central degree of freedom in pre- and post-processing systems that deploy multiple-description coding that has not been discussed in the literature so-far. To assess its effect on performance, statistically and in a true video coding system, we designed a system framework where minimal interaction with the legacy coders is allowed. Prohibition of, e.g., access to the internal state of the legacy coders, makes the framework applicable to a wide range of legacy coders. This does, however, not exclude the possibility of obtaining even better results in systems where such solutions are possible. Simulation results show that exploiting the possibility of interlacing intraframes indeed does improve overall performance.

6. REFERENCES

[1] John G. Apostolopoulos, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," *Visual Communications and Image Processing*, October 2001.

[2] Jin Young Lee and Hayder Radha, "Interleaved source coding (ISC) for predictive video coded frames over the internet," in *IEEE International Conference on Communications*, May 2005, vol. 2, pp. 1224–1228.

[3] Michael Gallant, Shahram Shirani, and Faouzi Kossentini, "Standard-compliant multiple description video coding," in *IEEE International Conference on Image Processing*, October 2001, pp. 946–949.

[4] D. Wang, N. Canagarajah, D. Redmill, and D. Bull, "Multiple description video coding based on zero padding," in *International Symposium on Circuits and Systems*, May 2004, vol. 2, pp. II–205–208.

[5] Nicola Franchi, Marco Fumagalli, Rosa Lancini, and Stefano Tubaro, "Multiple description video coding for scalable and robust transmission over IP," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 3, pp. 321–334, March 2005.

[6] O.A. Lotfallah and S. Panchanathan, "Adaptive multiple description coding for Internet video," in *International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. V–732–735.

[7] J. G. Apostolopoulos and M. D. Trott, "Path diversity for enhanced media streaming," *IEEE Communications Magazine*, vol. 42, no. 8, pp. 80–87, August 2004.

[8] Guanjun Zhang and Robert L. Stevenson, "Efficient error recovery for multiple description video coding," in *International Conference on Image Processing*, 2004, pp. 829– 832.

[9] G. Olmo and T. Tillo, "Directional multiple description scheme for still images," in *International Conference on Electronics, Circuits and Systems*, December 2003, vol. 2, pp. 886–889.

[10] Xunqi Yu, James W. Modestino, and Xusheng Tian, "The accuracy of Gilbert models in predicting packet-loss statistics for a single-multiplexer network model," in *Proc. IEEE INFOCOM*, March 2005, pp. 2602–2612.