Multiple Description H.264 Video Coding with Redundant Pictures

Ivana Radulovic
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
ivana.radulovic@epfl.ch

Ye-Kui Wang, Stephan Wenger, Antti Hallapuro, Miska M. Hannuksela
Nokia Research Center
Tampere, Finland
{ye-kui.wang, stephan.wenger, antti.hallapuro, miska.hannuksela}@nokia.com

Pascal Frossard
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
pascal.frossard@epfl.ch

ABSTRACT

Multiple description coding (MDC) offers a competitive solution for video transmission over lossy packet networks, with a graceful degradation of the reproduced quality as the loss rate increases. This paper illustrates how redundant pictures, an error resilience tool included in H.264/AVC, can be employed in conjunction with multiple state video coding scheme, previously proposed in the literature. The proposed MDC solution is shown to provide superior performance to state-of-the-art techniques, in terms of improved average luma peak-signal-to-noise-ratio (PSNR), fewer temporal fluctuations in the picture quality, and improved robustness to bad estimation of the loss probability in the network.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications applications - Computer conferencing, teleconferencing, and videoconferencing

General Terms

Design, Measurement, Performance.

Keywords

Multiple description video coding, redundant pictures, error resilience

1. INTRODUCTION

Even though there is continuous growth in available bandwidth in access networks, processor speed and memory, real-time video communication applications operating over the Internet and wireless networks still face packet losses as well as variable and unpredictable throughput. Retransmission is commonly used to overcome packet losses but is not acceptable for low-delay applications, such as conversational or interactive applications (e.g., video phone and conferencing). Another common strategy for tackling transmission errors is the use of forward error correction (FEC) codes. However, in order to decorrelate typical errors of wireless channels, FEC codes should be calculated over a relatively long period of media data, thus making them impractical to be used in conversational applications. It is therefore a common solution to rely on the source coding level error resilience methods in conversational video communication applications.

One suggested solution providing error resilience in the source coding level is multiple description coding (MDC). The key objective of MDC is to represent a signal in more than one description in such a way that a high quality is achieved when all descriptions are reconstructed in combination, and the quality degrades gracefully when one or more descriptions are not available for reconstruction. The most common MDC model refers to two descriptions with rates R_1 and R_2 respectively, that are sent over two lossy channels. Receiving only the description i (with $i = \{1, 2\}$) results in the *side* distortion D_i, while receiving both descriptions induces the *central* distortion D_{12}. In this work we consider the so-called balanced case, where rates $R_1 = R_2$, and when the side distortions are approximately equal.

Numerous interesting MDC video algorithms have emerged over the years [1]. Among these, the schemes based on information splitting have so far been mostly explored in MD video coding, mainly due to their simplicity and compatibility with existing video coding standards. They can operate either in the temporal [2], spatial [3] or the frequency domain [4], where each partition corresponds to a different description. This class of algorithms takes advantage of redundancy inherently present in a source. For example, there is usually a high degree of correlation between the neighboring pixels/lines/columns in a frame or between neighboring frames. Therefore, lost descriptions can be recovered based on the correctly received ones. From all the solutions in this category, the most popular one is certainly the so-called multiple state video coding (MSVC), which has been proposed by Apostolopoulos in [2]. The idea is similar to the one of video redundancy coding [5], except that the sync-frames are not used. As depicted in Figure 1, the input video is split into sequences of odd and even frames, each being coded as an independent description, and with its own prediction process and state. In MSVC, even if one description is completely lost, the other one can be independently decoded, and the reconstructed video can be rendered

MV'07, September 28, 2007, Augsburg, Bavaria, Germany.

at half of the frame rate. It has also been suggested to recover lost frame(s) in a damaged representation by utilizing temporally adjacent frames(s) in another description, and use these recovered frames for future prediction. However, when recovering reference pictures, a significant drift is introduced, which often leads to annoying artifacts.

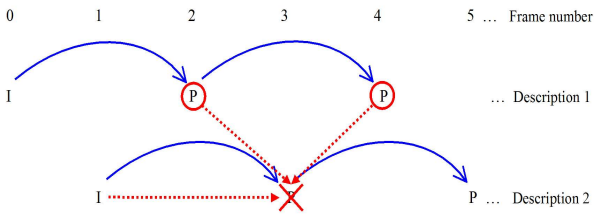

Figure 1: Multiple state video coding scheme, proposed in [2].

To reduce the decoding complexity of [2] and to improve the recovery in case of losses, Zhang et al. [6] propose to use multi-hypothesis motion-compensated prediction (MHMCP). Another solution, also based on odd/even frame separation, was proposed by Wang et al. in [7], where a second-order motion compensation prediction of a current frame based on two previous reference frames is built to increase error resiliency. In [8], a slice group based MDC scheme has been proposed. Each frame is divided into two slices by a checkerboard pattern, and each description contains a finely quantized slice, and the coarse of another slide. A slightly different polyphase down-sampling technique has been proposed in [9] and the results reported an improvement to a simple checkerboard slice pattern. In [10], Wang et al. use GOB alternation and low-quality macroblock update for MD video coding.

Using a coarsely quantized version of a stream to replace the possibly lost good quality parts has also been addressed in MD audio coding. For example, Hardman et al. [11] propose a robust audio tool (RAT) scheme, where redundant information for any given packet of speech is piggy-backed onto a later packet. Jiang and Ortega [12] suggest generating one description from quantized even speech samples in fine resolution, and the difference between even and odd samples in coarse resolution, while the situation would be the other way round in another description. Finally, in distributed video coding, the authors in [13] propose generating a supplementary coarsely quantized bitstream using Wyner-Ziv encoding, which is combined with the error-prone main stream to produce an improved video quality.

In this paper, we build on the popular MSVC scheme [2], and we propose to use *redundant pictures* (RP) in order to attenuate the error drift in case of loss. We refer to this scheme as Multiple-State Video Coding with Redundant Pictures (MSVC-RP). The redundancy is adapted to the expected loss rate, and controlled by acting on the quantization parameters used to code the redundant pictures. The addition of redundant pictures permits to considerably reduce the error drift, thereby to increase the error resilience and the quality at the receiver. When comparing MSVC-RP with the unmodified MSVC, as well as with two single-description coding schemes based on intra refresh, we are able to show significant improvements in terms of average luma PSNR, amount of temporal fluctuations of the quality of the reconstructed video, and robustness to unknown network conditions.

The paper is organized as follows. Section II provides more detailed description of our technology. Section III compares the proposed MSVC-RP method against three other error resilient coding methods. Finally, Section IV summarizes the paper.

2. MDC VIDEO CODING WITH REDUNDANT PICTURES

2.1 Proposed scheme

Redundant pictures (RP) are one error resilient tool included in H.264/AVC. According to the standard, each picture may be associated with one or more RPs, which a decoder can reconstruct in case a primary picture (or parts thereof) is missing. H.264/AVC does not define how to generate redundant pictures as long as a decoded redundant picture is visually similar in appearance to the corresponding decoded primary picture. One prominent scenario is producing a redundant picture from the same source picture as the primary one, but more coarsely quantized. Therefore, if a primary picture is encoded with the quantization parameter Q_p, the same parameter for the redundant picture can take any value between Q_p and the maximal possible value (51 in H.264/AVC).

Figure 2 illustrates the MSVC-RP scheme proposed in this paper. The input video sequence is split into sequences of odd and even source pictures, as in [2]. When encoding, each primary picture in the even/odd description is predicted only from other pictures of the same description, typically the previous picture. In addition, RPs are included in the bitstream of each description. These RPs carry the information from the alternate description and, in the time domain, they are placed such that they can substitute a possibly lost primary picture. In the current implementation, redundant pictures are coded as P pictures and each primary frame has its redundant version. The streams built in such a way are independent, and therefore a reconstruction at a full rate is possible even with one of them only.

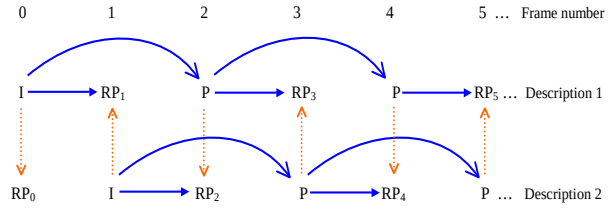

Figure 2: The proposed scheme for MDC video.

The descriptions are sent possibly over two different lossy links, or over one link in an interleaved fashion. In the current implementation, if primary pictures are received error-free, the corresponding RPs are treated as pure redundancy and thus discarded. Although the information from the redundant picture could be exploited to further enhance the quality of the corresponding primary picture, we refrain from this solution, thus trying to keep the decoding process as simple as possible. If a primary picture (or parts thereof) has been lost, the timewise corresponding redundant picture

is reconstructed and used to replace its missing parts. Typically, replacing the lost parts of a primary picture with the same content, but more coarsely quantized, creates much smaller artifacts than if the missing parts were concealed with the information from the neighboring macroblocks from the same and/or subsequent frames. Finally, if both primary and redundant parts of a picture are lost, the missing information is copied from the closest available previous frame from either description. Bi-directional motion-compensated error concealment, as suggested in [2], is not considered in this paper, because many of the receiving devices are not expected to have the processing power required for retroactive and bi-directional error concealment. After the necessary discarding/replacement/concealment, the two descriptions are subsequently interleaved to produce the final reconstruction.

Two sources of redundancy are introduced in MSVC-RP. First, there is the redundancy due to the fact that the frames in each description are now spaced temporarily further apart, as in MSVC. In addition, encoding of redundant pictures requires additional bitrate. Figure 3 shows the introduced redundancy for the Foreman QCIF sequence, compared to both single description coding and MSVC. In this case, we fixed $Q_p = 20$ (which corresponds to the total single description rate of 222.6 kbits/s), and we vary Q_r from Q_p to 50. We can see that, compared to MSVC [2], the amount of redundancy varies from 2.85%, when $Q_r = 50$, up to 80.34%, when $Q_p = Q_r = 20$. Compared to the single description case, the redundancy varies from 37.46% to 140.5%. As we will see later, the redundancy introduced thereby helps to significantly lessen the drift in case of repair.

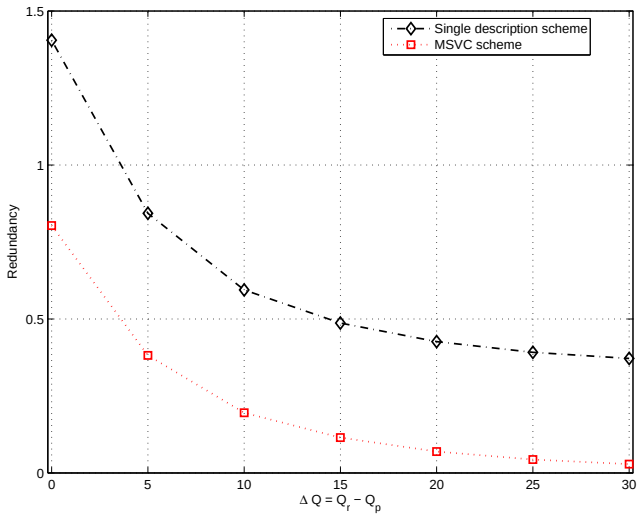

Figure 3: Redundancy introduced in our scheme, compared to MSVC [2] and the single description case(Foreman QCIF, encoded at 7.5 fps and with $Q_p = 20$).

2.2 Selecting Q_p and Q_r

The resoltuion of redundant pictures should be chosen by taking the network loss rate into account. If the loss rate is very low, the probability of a RP being used is low as well, and therefore RPs should be quantized coarsely. In that case the redundancy compared to MSVC will be as small as few percents. On the other hand, with increasing loss rates, better quality of RPs becomes more advisable, which at the fixed total rate, comes at a price of reducing the quality of primary pictures. However, in the average sense, this will be more beneficial, since now the probability that a RP will be used as a replacement is high.

To demonstrate this, we first consider the Foreman QCIF sequence encoded at 7.5 fps, with a target bit rate of 144 kbps. We use constant Q_p and Q_r parameters to encode the sequence and we manually check which combinations of QPs satisfy the aforementioned bitrate constraint. For this sequence, five different combinations of QPs permit to match the total bitrate of 144 kbps : $\{(Q_p = 25, Q_r = 42), (Qp = 26, Q_r = 34), (Q_p = 27, Q_r = 31), (Q_p = 28, Q_r = 29), (Q_p = 29, Q_r = 29)\}$. The five bitstreams have been tested for each loss rate and the ones that give the highest average luma PSNR have been selected. The discovered best QP combinations are shown in the left part of the Table I. As expected, the difference between QPs for primary and redundant pictures resulting to the highest average luma PSNR is significant at low loss rates, whereas at high loss rates, the optimal QP difference is small.

Next, we consider the Stefan CIF sequence, which has been encoded at 30 fps and 512 kbits/s. The following combinations of quantization parameters lead to the desired bitrate: $\{(Q_p = 41, Q_r = 49), (Q_p = 42, Q_r = 44), (Q_p = 43, Q_r = 43)\}$. The discovered optimal parameters for the four considered loss ratios are given in the right part of Table I. It can be seen that the optimal difference between quantization parameters between the primary and redundant pictures ranged from 8 to 2 for loss rates 3% and 20%, respectively.

p	Foreman QCIF		Stefan CIF	
	Q_p^{opt}	Q_r^{opt}	Q_p^{opt}	Q_r^{opt}
3%	25	42	41	49
5%	26	34	41	49
10%	28	29	42	44
20%	28	29	42	44

Table 1: Combinations of quantization parameters that give minimal average distortions, as a function of p.

3. PERFORMANCE EVALUATION

3.1 Testbed

Our testbed corresponds to the common error resilience testing conditions specified in JVT-P206, [14]. The NAL unit size is limited to 1400 bytes, and a single slice is packetized per NAL unit. An overhead for the RTP/UDP/IPv4 headers of 40 bytes is also taken into account. We use the four packet loss pattern files, included in ITU-T VCEG Q15-I-16 [15], which correspond to average packet losses of 3%, 5%, 10% and 20%. Only the first frames in all the video sequences are encoded as I pictures. Experimental results were produced for Foreman QCIF at 7.5 fps and 144 kbits/s, and Stefan CIF sequence at 30 fps and 512 kbits/s.

We compare MSVC-RP with three state-of-the-art schemes for the same overall bitrate :

- *MSVC* scheme [2]: we focus on a simple algorithm

that replaces a lost picture with the temporally closest correctly reconstructed picture from either description [16].

- *Adaptive intra refresh* scheme (AIR) scheme [6], which takes into account both the source and expected channel distortion (due to losses) when choosing an optimal mode for each macroblock. Therefore, it is likely to place intra macroblocks in more "active" areas.
- *Random intra refresh* (RIR) scheme, which selects randomly a number of macroblocks to be coded in the intra mode. The number of introduced intra macroblocks in both AIR and RIR schemes is driven by the network losses [17].

In our MSVC-RP implementation, parts or entirely lost pictures are replaced with their redundant versions, taken from the alternate description. If both primary and redundant picture are lost, we copy the temporally closest decoded picture from either description. For the other schemes, in case of partial frame losses, the missing pieces are copied from the corresponding places in the previous pictures. If an entire picture is lost, we copy the entire previous picture, as is done in our scheme.

3.2 Simulation results

Figure 4 first demonstrates how the average luma PSNR develops as a function of the loss rate, for the Foreman sequence. For MSVC-RP, besides the results utilizing optimized QPs as discussed above, we also plot the results for the case when $Q_p = Q_r = 29$ – a rather naive choice. We can observe that both MSVC-RP cases perform better than all the other schemes over the whole range of loss rates. At $p = 3\%$, the use of MSVR-RP with appropriate QPs results in improvements of 2.5 dB and 6.9 dB in terms of average luma PSNR over the AIR and RIR, respectively. At 20% of loss rate, the gain over the AIR scheme is 1.5 dB, while it can be as high as 10.6 dB, compared to the RIR .

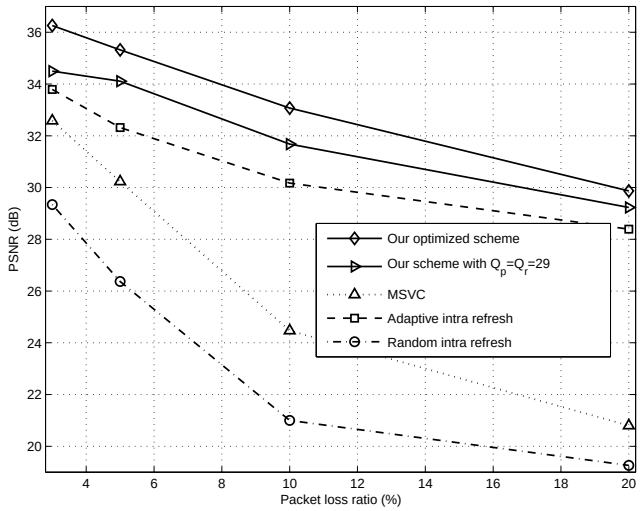

Figure 4: Average PSNR, for four schemes and four loss patterns (Foreman QCIF, 7.5 fps, 144 kbits/s).

To further evaluate the performance of MSVC-RP, we investigated how the luma PSNR changed over time. We applied identical loss pattern taken from the error pattern files to all four schemes. Figure 5 depicts the temporal behavior at 5% average loss rate. We can observe that the losses in the first pictures cause large oscillations in quality over time, for both MSVC and AIR. However, while the AIR scheme seems to recover gradually over time, this is not the case with MSVC, which recovers only after a scene change. RIR appears to be very vulnerable and its performance is inferior to the three other schemes, except in the last parts of Foreman where the scene is almost static. Finally, MSVC-RP shows a clear superior performance, not only in terms of an average PSNR, but also in the robustness to losses, leading to more stable quality.

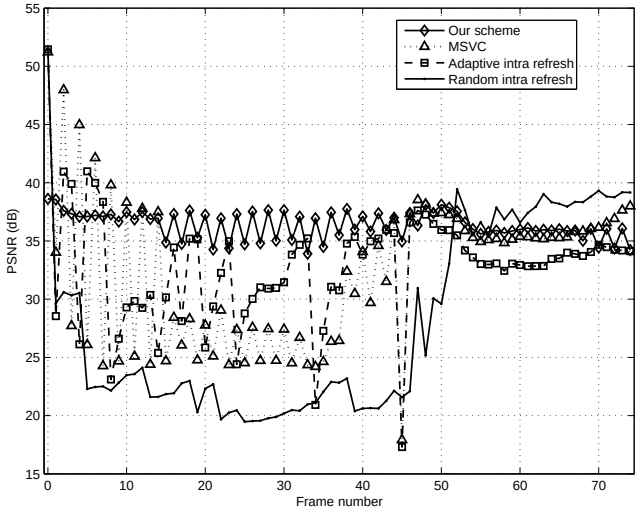

Figure 5: Reconstructed video quality, on a frame basis, when PLR = 5% (Foreman QCIF, 7.5 fps, 144 kbits/s).

Figure 6 illustrates the average distortion evolution versus the total bit rate, when the packet loss ratio is 5%. We show the comparison of MSVC-RP and AIR only, since the other schemes behaved significantly worse. We observe that MSVC-RP outperforms AIR in the entire considered range of bitrates and that the gain increases with the target bitrate. The improvement is 0.6 dB in terms of average luma PSNR for the rate 32 kbits/s, while it reaches 2.7 dB when the total rate is 192 kbits/s.

Finally, we compare our scheme and the AIR approach optimized for a given loss ratio p, but when the actual channel characteristics are different (as it may happen in practical scenarios when channel characteristics change). Figure 7 shows performance for the schemes optimized for $p = 5\%$, but when the actual loss ratio varies from 3% to 20%. For the sake of completeness we also plot the best performance of MSVC-RP and AIR at each loss ratio. The difference between the optimized and actual performance for both schemes are 0.39 dB and 0.14 dB respectively, when $p = 3\%$. Not surprisingly, the gap between the optimized and actual performance increases as the actual loss ratio moves away from 5%. At $p = 10\%$ these gaps for both schemes are 0.9 dB and 1.33 dB respectively, while at $p = 20\%$ the corresponding gaps are 1.32 and 2.78 dB. Therefore, we can conclude that MSVC-RP is more robust to unknown network conditions.

Similar results are observed for the Stefan CIF sequence.

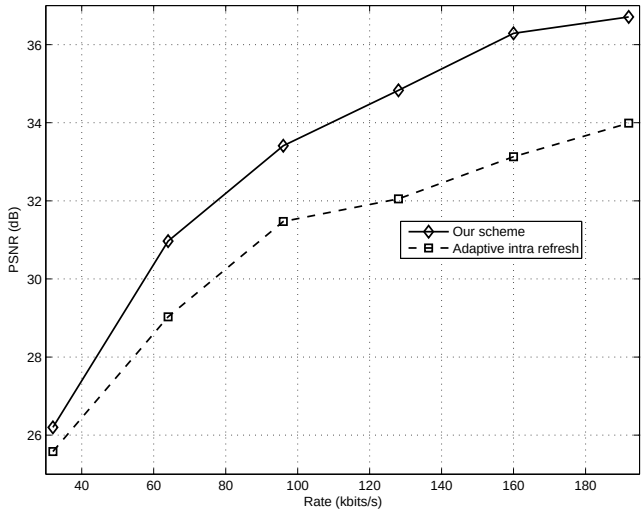

Figure 6: Average distortion, as a function of encoding rate, when PLR = 5% (Foreman QCIF, 7.5 fps).

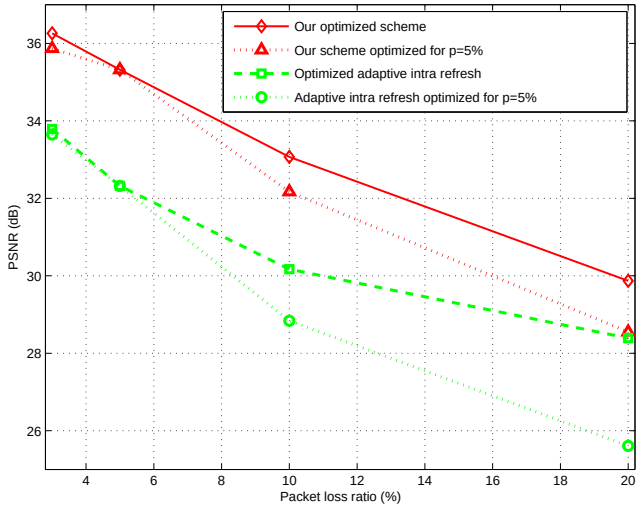

Figure 7: Average distortion when all the schemes are optimized for PLR = 5%, but the actual PLR is different (Foreman QCIF, 7.5 fps).

Figure 8 shows the average luma PSNRs as functions of the network loss ratio. For MSVC-RP, we considered the hand-tuned QP choice and $Q_p = Q_r = 43$ as a naive choice. MSVC-RP yielded better performance in average luma PSNR compared to the other tested schemes in three out of four packet loss rates. MSVC-RP with optimized QPs performs 1.2 dB better than AIR when $p = 3\%$, while AIR is slightly better at $p = 20\%$ (0.3 dB). At $p = 3\%$, our MSVC-RP scheme with optimized QPs offers a gain of 0.9 dB over MSVC-RP with $Q_p = Q_r = 43$. The improvement is 0.5 dB at $p = 20\%$. Gain over the MSVC ranges from 2.1 to 3.4 dB, while it varies from 3.8 dB to 5.4 dB for the RIR.

Figure 9 illustrates the behavior over time for the Stefan sequence under 10% loss. Significant artifacts were visible in all cases at the beginning of the sequence. Once more,

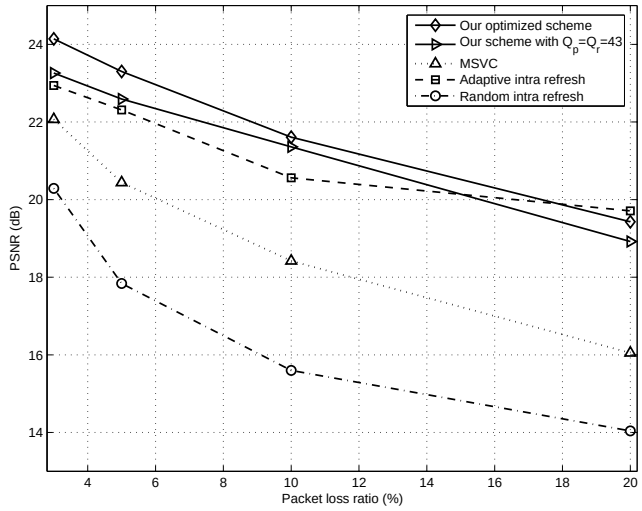

Figure 8: Average PSNR, for four schemes and four loss patterns (Stefan CIF, 30 fps, 512 kbits/s).

only MSVC-RP is able to maintain an acceptable quality level and also produced more stable reconstruction quality over time.

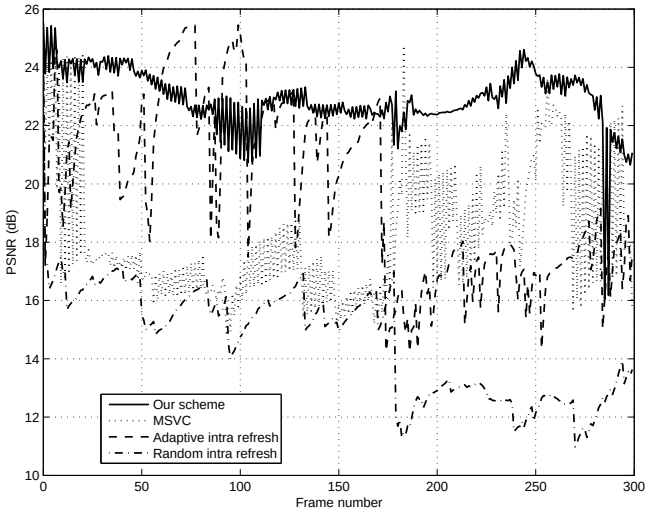

Figure 9: Reconstructed video quality, on a frame basis, when PLR = 10% (Stefan CIF, 30 fps, 512 kbits/s).

Finally, Figure 10 shows the average PSNRs for the MSVC-RP and AIR schemes optimized for $p = 3\%$, but when the actual loss rates are different than the expected ones. We can see that the gap between our optimized scheme and the one optimized for $p = 3\%$ stays very small in the whole range of considered losses, with the maximal value of only 0.2 dB (at $p = 20\%$). On the other hand, AIR seems to be much more vulnerable to unknown conditions, resulting in a gap of 4.1 dB between the optimized scheme and the one optimized for $p = 3\%$, when the actual loss rate is 20%.

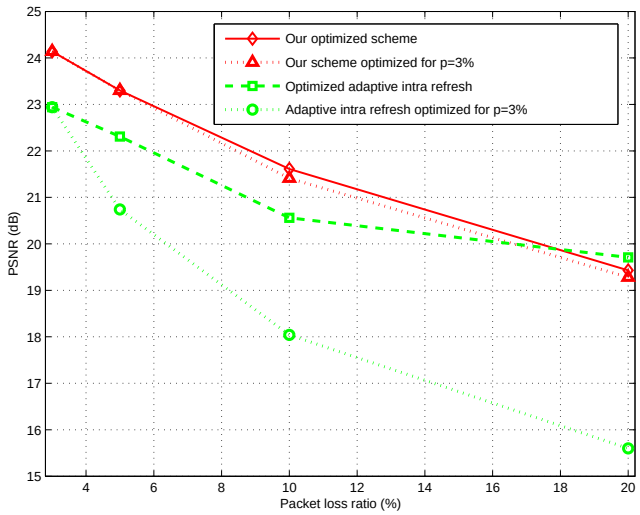

Figure 10: Average distortion when all the schemes are optimized for PLR = 3%, but the actual PLR is different (Stefan CIF, 30 fps).

4. CONCLUSIONS

In this work, we have shown how redundant pictures can be advantageously used as a resiliency tool in multiple description video coding. Comparisons with state-of-the-art error resilience techniques have shown a superior performance of the proposed MSVC-RP scheme in terms of average luma PSNR, stability of reconstructed picture quality over time, and robustness to unknown network conditions. In MSVC-RP, it appears that at lower loss rates relatively small number of bits should be spent on redundant pictures (by using coarser quantization), while at high loss rates the redundant pictures should be almost as finely quantized as the primary ones. However, a theoretical analysis of the rate-distortion properties and an adaptive rate selection algorithm for the different picture types in MSVC-RP remain as subjects of the future work.

5. REFERENCES

[1] Y. Wang, A.R. Reibman and S. Lin, "Multiple description coding for video delivery," *Proceeding of the IEEE*, vol. 93, no. 1, pp. 57–70, Jan. 2005.

[2] J. Apostolopoulos, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," *Proceedings of Visual Communications: Image Processing*, pp. 392–409, Jan. 2001.

[3] N.Franchi, M.Fumagalli, R. Lancini and S. Tubaro, "Multiple description video coding for scalable and robust transmission over IP," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 3, pp. 321–334, Mar. 2005.

[4] A.R. Reibman, H. Jafarkhani, M.T. Orchard and Y. Wang, "Multiple description video using rate-distortion splitting," in *Proceedings of IEEE International Conference on Image Processing*, Oct. 2001, pp. 978–981.

[5] S. Wenger, "Video Redundancy Coding in H.263+," in *Proceedings of Workshop on Audio-Visual Services for packet networks*, 1997.

[6] Y. Zhang, W. Gao, H. Sun, Q. Huang, Y. Lu, "Error resilience video coding in H.264 encoder with potential distortion tracking," in *Proceeding of International Conference on Image Processing*, vol. 1, Oct. 2004, pp. 163–166.

[7] Y. Wang and S. Lin, "Error-resilient video coding using multiple description motion compensation," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 438–452, June 2002.

[8] D. Wang, N. Canagarajah and D. Bull, "Slice group based multiple description video coding using motion vector estimation," in *Proceedings of IEEE International Conference on Image Processing*, vol. 5, Oct. 2004, pp. 3237 – 3240.

[9] J. Jia and H.-K. Kim , "Polyphase Downsampling Based Multiple Description Coding Applied to H.264 Video Coding," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89ŰA, no. 6, pp. 1601–1606, June 2006.

[10] Y. Wang and C. Wu, "Low complexity multiple description coding method for wireless video," in *Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA)* , vol. 2, Mar. 2005, pp. 95– 98.

[11] V. Hardman, M.A. Sasse, M. Handley and A. Watson, "Reliable audio for use over the Internet," in *Proceedings of INET*, 1995.

[12] J. Wenqing and A. Ortega, "Multiple description speech coding for robust communication over lossy packet networks," in *Proceedings of IEEE International Conference on Multimedia and Expo*, vol. 1, Aug. 2000, pp. 444–447.

[13] S. Rane, A. Aaron and B. Girod, "Systematic lossy forward error protection for error-resilient digital video broadcasting - a Wyner-Ziv coding approach," in *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 5, Oct. 2004, pp. 3101–3104.

[14] Y.-K. Wang, S. Wenger and M.M. Hannuksela, "Common conditions for SVC error resilience testing," *JVT document P206*, Aug. 2005.

[15] S. Wenger, "Proposed error patterns for Internet Experiments," *ITU-T VCEG document Q15-I-16*, Oct. 1999.

[16] B. Heng, J. Apostolopoulos, J.S. Lim, "End-to-end rate-distortion optimized MD mode selection for multiple description video coding," *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. Article ID 32592, 2006.

[17] P. Haskell, D. Messerschmitt, "Resynchronization of motion compensated video affected by ATM cell loss," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Mar. 1992, pp. 545–548.