Multiple Description Scalable Coding For Video Transmission Over Unreliable Networks

Roya Choupani[†‡], Stephan Wong[†], and Mehmet R. Tolun[‡]

[†]Computer Engineering Department, TUDelft, Delft, The Netherlands
[‡] Computer Engineering Department, Çankaya University, Ankara Turkey
roya@dutepp0.et.tudelft.nl
J.S.S.M.Wong@tudelft.nl
tolun@cankaya.edu.tr

Abstract. Developing real time multimedia applications for best effort networks such as the Internet requires prohibitions against jitter delay and frame loss. This problem is further complicated in wireless networks as the rate of frame corruption or loss is higher in wireless networks while they generally have lower data rates compared to wired networks. On the other hand, variations of the bandwidth and the receiving device characteristics require data rate adaptation capability of the coding method. Multiple Description Coding (MDC) methods are used to solve the jitter delay and frame loss problems by making the transmitted data more error resilient, however, this results in reduced data rate because of the added overhead. MDC methods do not address the bandwidth variation and receiver characteristics differences. In this paper a new method based on integrating MDC and the scalable video coding extension of H.264 standard is proposed. Our method can handle both jitter delay and frame loss, and data rate adaptation problems. Our method utilizes motion compensating scheme and, therefore, is compatible with the current video coding standards such as MPEG-4 and H.264. Based on the simulated network conditions, our method shows promising results and we have achieved up to 36dB for average Y-PSNR.

Key words: Scalable Video Coding, Multiple Description Coding, Multimedia Transmission

1 Introduction

Communications networks, both wireless and wired, offer variable bandwidth channels for video transmission [1], [3]. Display devices have a variety of characteristics ranging from low resolution screens in small mobile terminals to high resolution projectors. The data transmitted for this diverse range of devices and bandwidths have different sizes and should be stored on media with different capacity. Moreover, an encoding which makes use of a single encoded data for all types of bandwidth channels and displaying devices capacities could be of a remarkable significance in multimedia applications. Scalable video coding (SVC) schemes are intended to be a solution for the Internet heterogeneity and receiver

display diversity problem by encoding the data at the highest quality but enabling the transmitter or receiver to utilize it partially depending on the desired quality or available bandwidth and displaying capacities. The main drawback of the available scalable video coding methods is that they are not suitable for non-reliable environments with a high rate of frame loss or corruption such as wireless networks. This problem stems from the fact that the methods are based on the motion compensated temporal filtering scheme and frames are coded as difference with a (generally prior) reference frame. In case that a reference frame is lost or corrupted, the whole chain of difference frames depending on it becomes unrecoverable. To increase the error resilience of the video coding methods, Multiple Description Coding (MDC) methods have been introduced [4], [5], [7]. These methods improve the error resilience of the video with the cost of adding redundancy to the code. In case that a frame is lost or corrupted, the redundancy is used to replace it with an estimated frame. Franchi, et al., proposed a method to send a video by utilizing independent multiple descriptions. Their method however, does not combine scalability features with multiple description coding and therefore only addresses frame loss or corruption and variations of bandwidth have not been dealt with [16]. The combination of scalable video coding methods and multiple description coding has attracted the interest of researchers recently [2], [3], [13]. The introduction of scalable extension of H.264 standard recently, which relaxes some of the restrictions of other video coding schemes such as using immediate prior frame as reference frame, provides a suitable framework for combining scalability of H.264 with error resistance of MDC schemes. This paper describes a new method which is a combination of the SVC extension of H.264 standard with MDC schemes in a way that no redundancy in the form of extra bits is introduced during the video coding. The remainder of this paper is organized as follows. Section 2 introduces the main multiple description coding methods. Section 3 explores the scalability features of H.264 standard which are used in our proposed method. Section 4 describes the details of our proposed method. In Section 5, we introduce the theoretical base of our performance evaluation method and provide the experimental results and finally, in Section 6, we draw the conclusions.

2 Multiple Description Coding

As a way of encoding and communicating visual information over lossy packet networks, multiple descriptions have attracted a lot of attention. A multiple description coder divides the video data into several bit-streams called descriptions which are then transmitted separately over the network. All descriptions are equally important and each description can be decoded independently from other descriptions which means that the loss of some of them does not affect the decoding of the rest. The accuracy of the decoded video depends on the number of received descriptions. Descriptions are defined by constructing P non-empty *sets* summing up to the original signal f. Each set in this definition corresponds to a description. The sets however, are not necessarily disjoint. A signal sample

may appear in more than one set to increase error resilience property of the video. Repeating a signal sample in multiple descriptions is also a way for assigning higher importance to some parts/signals of the video. The more a signal sample is repeated the more reliably it is transmitted over the network. The duplicate signal values increases the redundancy and hence the data size which results in reduced efficiency. Designing descriptions as partition does not necessarily mean that there is no redundancy in the data. In fact, designing the descriptions as a partition prevents extra bits to be added to the original data for error resilience but still the correlation between the spatially or temporally close data can be used for estimating the lost bits. The estimation process is commonly referred to as error concealment and relies on the the preserved correlation in constructing the descriptions. Fine Granular Scalability (FGS)-based MDC schemes partition the video into one base layer and one or several enhancement layers [8]. The base layer can be decoded independently from enhancement layers but it provides only the minimum spatial, temporal, or signal to noise ratio quality. The enhancement layers are not independently decodable. An enhancement layer improves the decoded video obtained from the base layer. MDC schemes based on FGS puts base layer together with one of the enhancement layers at each description. This helps to partially recover the video when data from one or some of the descriptions are lost or corrupt. Repeating base layer bits in each descriptor is the overhead added for a better error resilience. In Forward Error Correction (FEC)-based MDC methods, it is assumed that the video is originally defined in a multi-resolution manner [6], [9]. This means if we have M levels of quality, each one is adding to the fidelity of the video to the original one. This concept is very similar to the multi-layer video coding method used by FGS scheme. The main difference, however, is that there exist a mandatory order in applying the enhancements. In other words, it is sensitive to the position of the losses in the bitstream, e.g., a loss early in the bitstream can render the rest of the bitstream useless to the decoder. FEC-based MDCs aim to develop the desired feature that the delivered quality become dependent only on the fraction of packets delivered reliably. One method to achieve this is Reed Solomon block codes. Mohr, et.al., [15] used Unequal Loss Protection (ULP) to protects video data against packet loss. ULP is a system that combines a progressive source coder with a cascade of Reed Solomon codes to generate an encoding that is progressive in the number of descriptions received, regardless of their identity or order of arrival. The main disadvantage of the FEC-based methods is the overhead added by the insertion of error correction codes. Discrete Wavelet Transform (DWT)-based video coding methods are liable for applying multiple description coding. In the most basic method, wavelet coefficients are partitioned into maximally separated sets, and packetized so that simple error concealment methods can produce good estimates of the lost data [2], [10], [11]. More efficient methods utilize Motion Compensated Temporal Filtering (MCTF) which is aimed at removing the temporal redundancies of video sequences.

If a video signal f is defined over a domain D, then the domain can be expressed as a collection of sub-domains {S1;..;Sn} where the union of these sub-domains

is a cover of D. Besides, a corrupt sample can be replaced by an estimated value using the correlation between the neighboring signal samples. Therefore, the sub-domains should be designed in a way that the correlation between the samples is preserved. Domain-based multiple description schemes are based on partitioning the signal domain. Each partition, which is a subsampled version of the signal, defines a description. Chang [8] utilizes the even-odd splitting of the coded speech samples. For images, Tillo, et.al., [11] propose splitting the image into four subsampled versions prior to JPEG encoding. There, domain partitioning is performed first, followed by discrete cosine transform, quantization and entropy coding. The main challenge in domain-based multiple description methods is designing sub-domains so that the minimum distance between values inside a domain (inter-domain distance) is maximized while preserving the auto-correlation of the signal.

3 Scalable Video Coding Extension of H.264

As a solution to the unpredictability of traffic loads, and the varying delays on the client side problem, encoding the video data is carried out in a rate scalable form which enables adaptation to the receiver or network capacities. This adaptation can be in the number of frames per second (temporal scalability), frame resolution (spatial scalability), and number of bits allocated to each pixel value (signal to noise ratio scalability). In this section, we briefly review the scalability support features of H.264 standard which are used in our proposed method. The scalability support features of H.264 standard were introduced based on an evaluation of the proposals carried out by MPEG and the ITU-T groups. Scalable video coding (SVC) features were added as an amendment to H.264/MPEG4-AVC standard [14].

3.1 Temporal Scalability

Temporal scalability is achieved by dropping some of the frames in a video to reach the desired (lower) frame rate. As the motion compensated coding used in video coding standards encodes the difference of the blocks of a frame with its reference frame (the frame coming immediately before it), dropping frames for temporal scalability can cause some frames to become unrecoverable. H.264 standard relaxes the restriction of choosing the previous frame as the reference frame for current frame. This makes it possible to design hierarchical prediction structures to avoid reference frame loss problem when adjusting the frame rate.

3.2 Spatial Scalability

In supporting spatial scalable coding, H.264 utilizes the conventional approach of multilayer coding, however, additional inter-layer prediction mechanisms are incorporated. In inter-layer prediction the information in one layer is used in the other layers. The layer that is employed for inter-layer prediction is called

reference layer, and its layer identifier number is sent in the slice header of the enhancement layer slices [12]. Inter-layer coding mode is applied when the macroblock in the base layer is inter-coded. To simplify encoding and decoding macro-blocks in this mode, a new block type named base mode block was introduced. This block does not include any motion vector or reference frame index number and only the residual data is transmitted in the block. The motion vector and reference frame index information are copied from those of the corresponding block in the reference layer.

4 Our Proposed Method

Our proposed method involves using the scalability features of the H.264 standard. To make the video resilient against frame loss or corruption error we define multiple descriptions. However, to achieve a high performance which is comparable to single stream codes, we do not include any error correction code in the descriptions. The error concealment in our proposed method is based on the autocorrelation of the pixel values which is a decreasing function of spatial proximity. Generally, the differences among the pixels values about a given point are expected to be low. Based on this idea we have considered four descriptions D_1 to D_4 representing four spatial sub-sets of the pixels in a frame as depicted in Figure 4. Each description correspond to a subset S_i for $i = 1..4$. The subsets define a partition as no overlap exists in the subsets and they sum up to the initial set.

$$S_i \bigcap S_j = \emptyset \quad for \quad i = 1, .., 4 \quad and \quad i \neq j$$

$$\bigcup_{i=1}^{4} S_i = D$$

Each description is divided into macro-blocks, motion compensated, and coded independently. The decoder extracts frames and combines them as depicted in Figure 4. When a description is lost or is corrupted, the remaining three de-

1	2	1	2	1
3	4	3	4	3
1	2	1	2	1
3	4	3	4	3
1	2	1	2	1

Fig. 1. Organization of the pixels in the descriptions

scriptions provide nine pixel values around each pixel of the lost description for interpolation during error concealment. Figure 4 depicts the pixel values utilized for interpolating a pixel value from a lost description. For interpolation, we

1	2	1	2	1
3	4	3	4	3
1	2	1	2	1
3	4	3	4	3
1	2	1	2	1

Fig. 2. Pixels used (blue) for interpolating the value of a missing pixel (red)

are using a weighted interpolation where the weights are normalized by the Euclidean distance of each pixel from the center as given below. We have assumed

$$\frac{1}{6.828} \times \begin{array}{|c|c|c|} \hline \frac{\sqrt{2}}{2} & 1 & \frac{\sqrt{2}}{2} \\ \hline 1 & 0 & 1 \\ \hline \frac{\sqrt{2}}{2} & 1 & \frac{\sqrt{2}}{2} \\ \hline \end{array}$$

the residue values and motion vectors and other meta-data in a macroblock is transmitted as a data transmission unit and hence are not available when the data packet is lost. The succeeding frames which utilize the estimated frame as their reference frame, will suffer from the difference between the reconstructed frame and the original one. The error generated in this way is propagated till the end of the GOP. However, if no other frame from the same GOP is lost, the error is not accumulated. The multilayer hierarchical frame structure of H.264 reduces the impact of frame loss to at most $log_2 n$ succeeding frames where n is the number of frames in a GOP. Our proposed method has the following features.

- Multiple description coding is combined with video scalable coding methods with no redundant bits added.
- Each description is independent from the rest and the base-enhancement relationship does not exist between them. This feature comes without the extra cost of forward error correction bits added to the descriptions. Any lost or corrupted description can be concealed regardless of its position or order with respect to the other descriptions.
- The proposed method is compatible with the definition of the multi-layer spatial scalability of H.264 standard. This compatibility is due to the possibility of having the same resolution in two different layers in H.264 and using inter-coding at each layer independently. We have not set the motion

prediction flag and let each description to have its own motion vector. This is because of the independent coding of each description. Setting the motion prediction flag can speed up encoder but it reduces the coding efficiency slightly as the most similar regions are not always happen at the same place in different descriptions.
– The proposed method is expandable to more number of descriptions if the error rate of the network is high, a higher level of fidelity with the original video is required, or higher levels of scalability are desired.

5 Experimental Results

For evaluating the performance of our proposed method, we have considered measuring Peak Signal to Noise Ratio of the Y component of the macroblocks (Y-PSNR). Equations 1 and 2 describe Y-PSNR used in our implementation mathematically.

$$PSNR = 20 \log_{10} \frac{Max_I}{\sqrt{MSE}} \tag{1}$$

$$MSE = \frac{1}{3mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} ||I(i,j) - I'(i,j)||^2 \tag{2}$$

where Max_I indicates the largest possible pixel value, I is the original frame and I is the decoded frame at the receiver side. Y-PSNR is applied to all frames of video segments listed in Table 5 by comparing the corresponding frames of the original video segment and after using our multiple description coding method. We have considered the case where one of the descriptions is lost and interpolated. We have randomly selected the erroneous description. We put 32 frames in each GOP and a diadic hierarchical temporal structure has been used for motion compensated coding. We have furthermore imposed the same reference

Table 1. Average Y-PSNR values when loss is in only one frame of each GOP.

Sequence Name	Resolution	Frame rate	Average Y-PSNR (db)
Foreman	352×288	30	36.345
Stefan & Martin	768×576	30	33.110
City	704×576	60	34.712

frame for all macroblocks of a frame for simplicity although H.264 supports utilizing different reference frame for macroblocks of a frame. In additionally, we have restricted the number of descriptions lost to one for each GOP. This means at most one forth of a frame is estimated during error concealment step. The location of the lost description in the GOP is selected randomly and the Y-PSNR is obtained for the average of each video segment. The average Y-PSNR values are reported in Table 5. The second set of evaluation tests considers the average

Y-PSNR value change for each video segment with respect to the number of frames affected by the lost description. Still however, we are assuming only one description is lost each time and the GOP length is 32. Figure 5 depicts the result of multiple frame reconstruction for three video segments. Despite having multiple frames affected by the loss or corruption problems, the results indicates that the ratio of peak signal to noise ratio is relatively high. As a benchmark

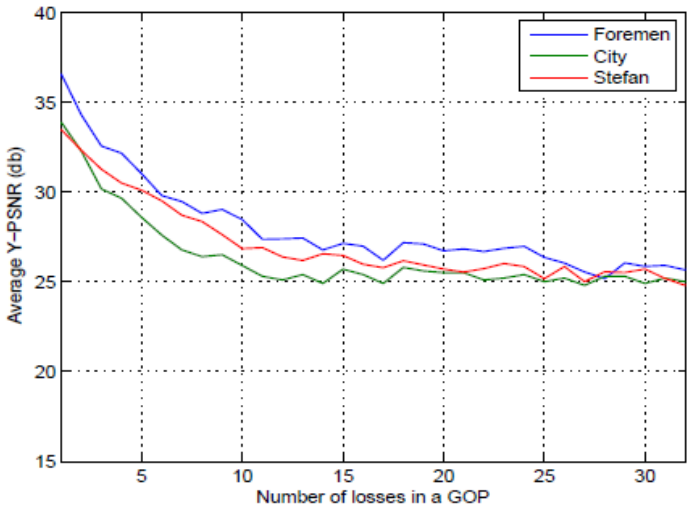

Fig. 3. Multiple Description Schemes with a) 9 Descriptions, b) 16 Descriptions

to evaluate the efficiency of our algorithm, we have compared average Y-PSNR value of Foreman and City video segments with single layer video coding. Figure 4 and 5 depict the comparison results.

6 Conclusion

A new method for handling the data loss during the transmission of video streams has been proposed. Our proposed method is based on multiple description coding however, coding efficiency is not sacrificed as no extra bit data redundancy is introduced for increasing resilience of the video. The proposed method has the capability of being used as a scalable coding method and any data loss or corruption is reflected as reduction in the quality of the video slightly. Except for the case when all descriptions are lost, the video streams do not experience jitter at play back. The compatibility of the proposed method with H.264 standard simplifies the implementation process. Our proposed method is based on spatial scalability features of H.264 however, a reasonable extension of the work is inclusion of SNR scalability.

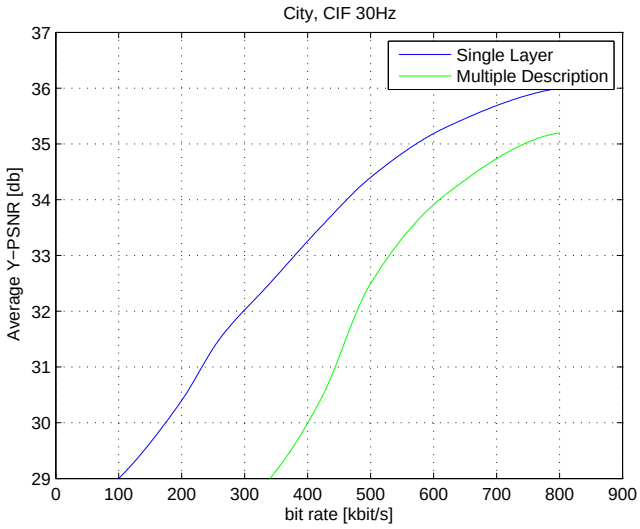

Fig. 4. Coding efficiency comparison between single layer and our proposed method using City video segment

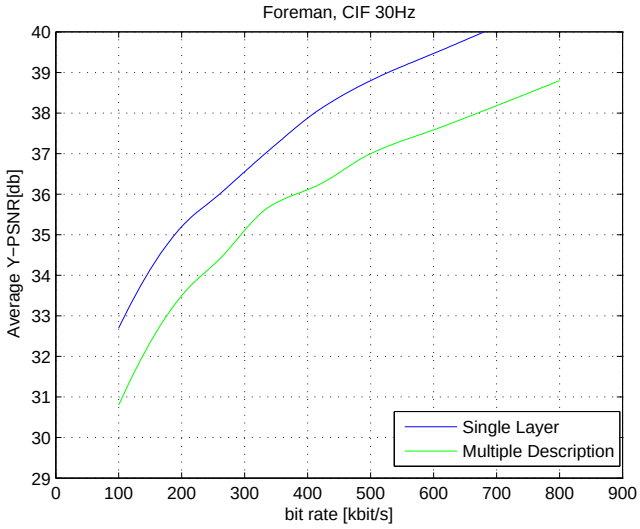

Fig. 5. Coding efficiency comparison between single layer and our proposed method using Foreman video segment

References

1. G. Conklin, G. Greenbaum, K. Lillevold, A. Lippman, and Y. Reznik, "Video Coding for Streaming Media Delivery on the Internet", IEEE Transaction on Circuits

and Systems for Video Technology, March 2001.
2. Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Fully-scalable Wavelet Video Coding using in-band Motion-compensated Temporal Filtering", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 417–420, 2003.
3. J. Ohm, "Advances in Scalable Video Coding", Proceedings of the IEEE, vol. 93, no. 1, Jan. 2005.
4. V.K. Goyal, "Multiple Description Coding: Compression Meets the Network", Signal Processing Magazine, IEEE Publication, vol. 18, issue 5 pp. 74–93, Sep. 2001.
5. Y. Wang, A. R. Reibman, L. Shunan, "Multiple Description Coding for Video Delivery", Proceedings of IEEE, vol. 93, No. 1, Jan. 2005.
6. R Puri, K Ramchandran, "Multiple Description Source Coding using Forward Error Correction Codes", Signals, Systems, and Computers, vol. 1, pp. 342-346, 1999.
7. R. Venkataramani, G. Kramer, V.K. Goyal, "Multiple Description Coding with many Channels", IEEE Transaction on Information Theory, vol. 49, issue: 9, pp. 2106–2114, Sept. 2003.
8. S. K. Chang, L. Sang," Multiple Description Coding of Motion Fields for Robust Video Transmission:, IEEE Transaction on Circuits and Systems for Video Technology, vol. 11, issue 9, pp 999–1010, Sep. 2001.
9. Y. Wang S. Lin, "Error-resilient Video Coding using Multiple Description Motion Compensation", IEEE Transaction on Circuits and Systems for Video Technology, vol. 12, issue 6, pp. 438–452, Jun. 2002.
10. Y. Xuguang, K. Ramchandran, "Optimal Subband Filter Banks for Multiple Description Coding", IEEE Transaction on Information Theory, vol. 46, issue 7, pp. 2477–2490, Nov. 2000.
11. T. Tillo, G. Olmo, "A Novel Multiple Description Coding Scheme Compatible with the JPEG2000 Decoder", IEEE Signal Processing Letters, vol. 11, issue 11, pp. 908–911, Nov. 2004 .
12. T. Wiegand, G.J. Sullivan, G. Bjontegaard, A. Luthra, "Overview of the H.264/AVC Video Coding Standard", IEEE Transaction on Circuits and Systems for Video Technology, vol. 13, issue 7, July 2003.
13. H. Schwarz, D. Marpe, T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard", IEEE Transaction on Circuits and Systems for Video, 2007
14. C. Hewage, H. Karim, S. Worrall, S. Dogan, A. Kondoz, "Comparison of Stereo Video Coding Support in MPEG-4 MAC, H.264/AVC and H.264/SVC" Proceeding of the 4^{th} Visual Information Engineering Conference, London, July, 2007.
15. A.E. Mohr, E.A. Riskin, R.E. Ladner, "Unequal Loss Protection: Graceful Degradation of Image Quality over Packet Erasure Channels through Forward Error Correction", IEEE Journal of Selected Areas in Communications, vol. 18, issue 6, pp. 819–828, Jun. 2000.
16. N. Franchi, M. Fumagalli, R. Lancini, S. Tubaro, "A Space Domain Approach for Multiple Description Video Coding", ICIP 2003, pp. 253–256, vol.2, 2003.