Research Article

Spatial Multiple Description Coding for Scalable Video Streams

Roya Choupani,¹ Stephan Wong,¹ and Mehmet Tolun²

¹ Computer Engineering, EEMCS, P.O. Box 5031, 2600 GA Delft, The Netherlands

² Elektrik-Elektronik Mühendisliği Bölümü, Mühendislik Fakültesi, Aksaray Üniversitesi, 68100 Aksaray, Turkey

Correspondence should be addressed to Roya Choupani; roya@cankaya.edu.tr

Received 14 April 2014; Revised 21 July 2014; Accepted 9 August 2014; Published 25 August 2014

Academic Editor: Ekram Khan

Copyright © 2014 Roya Choupani et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The need for adapting video stream delivery over heterogeneous and unreliable networks requires self-adaptive and error resilient coding. Network bandwidth fluctuations can be handled by means of a video coding scheme which adapts to the channel conditions. However, packet losses which are frequent in wireless networks can cause a mismatch during the reconstruction in the receiver end and result in an accumulation of errors which deteriorates the quality of the delivered video. A combination of multiple description coding in pixel domain and scalable video coding schemes which addresses both video adaptation and robustness to data loss is proposed in this paper. The proposed scheme combines error concealment with spatial video scalability. In order to improve the fidelity of the reconstructed to the original frames in presence of packet loss, a multilayer polyphase spatial decomposition algorithm is proposed. Classical multiple description methods interpolate the missing data which results in smoothing and artifact at object boundaries. The proposed algorithm addresses the quality degradation due to low-pass filtering effect of interpolation methods. We also comparatively analyze the trade-off between robustness to channel errors and coding efficiency.

1. Introduction

Several error concealment methods have been proposed to deal with data loss in unreliable networks among which the most important methods are forward error correction [1], intra/intercoding mode selection [2], layered coding [3], and multiple description coding (MDC) [4]. MDC methods are developed for increasing the reliability of data transmission over unreliable networks. In MDC methods, video is decomposed into descriptions which are transmitted over a probably independent network channel [4]. This decomposition can be performed before applying any transform to the video data or after application of the transform and hence to the transform coefficients. The decomposition of data can be done in spatial resolution by assigning pixels to different descriptions [5–7], in temporal resolution by assigning frames to different descriptions [8], and in signal-to-noise ratio (SNR) by transmitting less accurate pixel values in each description [9]. This decomposition should be optimized by minimizing the reconstruction error when one or some of the descriptions are lost and also by minimizing the redundancy in the descriptions. The extreme case in the MDC methods

is duplicating data and transmitting identical data at every description. In this case the reconstruction error in presence of a description loss or corruption is eliminated, and receiving only one description provides the total video data. However, the duplication of data reduces the coding efficiency. Hence, a trade-off should be sought between coding efficiency and error resilience of the video. Generally, descriptions have the same importance and data rates and each description can be decoded independently from other descriptions, even though this is not a necessary requirement. Independency of descriptions if provided means that the loss of some of these descriptions does not affect the decoding of the rest [10]. The accuracy of the decoded video depends on the number of received descriptions [11]. Figure 1 depicts the basic framework for a multiple description encoder/decoder with two descriptions.

In case of a failure in one of the channels, the output signal is reconstructed from the other description. Besides, the reduced video quality in terms of lower spatial or temporal resolutions, or lower bit per pixel quality when only some of the descriptions are delivered, can be utilized to add scalability property to the video. In case of spatial decomposition

FIGURE 1: Multiple descriptions coding block-diagram.

of video into descriptions, polyphase downsampling of the frame data [12–14] and quincunx subsampling [15] are used. Figure 2(a) depicts polyphase subsampling with four subsets. Each subset is transmitted in a description and in case of a data loss, the lost data is estimated by interpolation over its adjacent neighbors. This technique relies entirely on the correlation between adjacent pixels present in the video frames. Figure 2(b) depicts the division of the frame pixels into two subsets by quincunx subsampling as described in [15].

In [13] authors combine spatial and temporal decomposition of video into multiple descriptions. Each block of 8×8 is decomposed into four polyphase groups of 4×4 where groups 0 and 4 are inserted into description D_1 and groups 2 and 3 into description D_2 . Motion compensation is carried out before decomposition of the blocks and hence the same motion vectors are shared by both descriptions. This leads to retrieving the motion vectors whenever a description is lost. Meanwhile they decompose the video temporally by transmitting even and odd frames in different streams. The missing block is reconstructed by interpolating the block at previous and next frames.

In video coding a transform is used to create uncorrelated data. The correlation present in video data indicates a statistical dependency present between the pixel values which is considered as a redundancy that can be exploited for more effective coding [16]. This correlation is removed by applying transforms such as discrete cosine transform (DCT). MDC for error concealment can be applied to transform coefficients as well [17]. Decomposing the coefficient set into two or more descriptions has the problem of estimating the missing data from the received descriptions as the coefficients are no longer correlated after applying the transform. An attempt to create a correlation between coefficients was made in [18]. In their work, the authors defined two subsets from the coefficients by putting odd and even coefficients in different subsets. Assuming that σ_1^2 and σ_2^2 are the variances of the subsets S_1 and S_2 , respectively, the descriptions are created as

$$\begin{aligned} \gamma_1 &= 2^{-1/2} (S_1 + S_2), \\ \gamma_2 &= 2^{-1/2} (S_1 - S_2) \end{aligned} \quad (1)$$

with correlation coefficient $(\sigma_1^2 - \sigma_2^2)/(\sigma_1^2 + \sigma_2^2)$ known to the receiver end. Thus, when one description is lost it can be more effectively estimated from the received description than if the original subsets are used as descriptions.

2. Error Concealment by Interpolation

Many different interpolation algorithms, such as Near Neighbor Replication (NNR), Bilinear Interpolation [13], and Bicubic Interpolation, have been used in literature [5, 19]. However, interpolating the missing data in pixel domain when one of the descriptions is lost does not always provide satisfactory results from a subjective perspective. Even though the reconstructed video quality is high with respect to objective metrics such as PSNR, subjective evaluations may indicate degraded quality in video in some cases. This is the result of overall quality assessment that is performed in PSNR; however, the subjective assessments take into account the regional and structural features of the objects present in the video. This characteristic is more visible at the boundaries of objects because the interpolation performs like low-pass filters. Figures 3 and 4 depict a sample frame and the result of its reconstruction when one of the descriptions is lost and the corresponding pixels are interpolated by finding the average values of the adjacent pixels. As it can be seen from Figure 4, the pixels belonging to bright thin objects are replaced with darker pixel values after interpolation and cause artifacts. Edge preserving interpolation methods have been proposed as a solution to the low-pass filtering impact of interpolation. In [5], authors propose a nonlinear method called edge sensing to interpolate the missing data while preserving edge pixels. In this method, the horizontal ΔH and vertical gradients ΔV are computed for each missing pixel using its adjacent pixels. If one of the gradient values is greater than a predefined threshold, it is assumed that the pixel is on an edge and only the adjacent pixels along the edge direction are used for interpolation. If none of the gradient values is larger than the threshold, the average of four adjacent pixels is used for interpolation. Although their proposed method improves the performance of linear interpolators, the performance degrades in cases such as very thin (one pixel thick) objects and edges which are not along vertical and horizontal directions.

3. Proposed Method

Our proposed method is a multilayer MDC video coding method which decomposes video spatially into four descriptions. The descriptions indicated with labels D_1 to D_4 represent four spatial subsets of the pixels in a frame as depicted in Figure 2(a) corresponding to subsets S_i for $i = 1, \dots, 4$ of the initial set S . The decomposition defines a partition where no overlap exists between the subsets, and the partitions sum up to the initial set as defined below:

$$\begin{aligned} S_i \cap S_j &= \emptyset \quad \text{for } i, j = 1, \dots, 4, \quad i \neq j, \\ \bigcup_{i=1}^4 S_i &= S. \end{aligned} \quad (2)$$

FIGURE 2: Multiple descriptions (a) using polyphase downsampling and (b) using quincunx downsampling.

FIGURE 3: Sample frame with a thin object of bright color.

FIGURE 4: Reconstructed frame when one spatial description is missing.

Although the spatially proximate pixels are correlated, decomposing frames into disjoint descriptions can diminish correlation when the frame contains thin and small objects with high contrast. This reduced correlation deteriorates frame quality when reconstruction is done in presence of packet loss. Since in motion compensated temporal filtering (MCTF) a frame is reconstructed from its reference frame(s),

the reduced quality after reconstruction can accumulate to drift error. To reduce the impact of reconstruction with missing descriptions, we include a downsampled block as a common base layer in all descriptions. Hence, each description is built using the common base-layer and an enhancement layer which gives the difference between the base layer and one of the subsets depicted in Figure 2(a). Our

FIGURE 5: Block diagram of the proposed method.

proposed method decomposes a macroblock of 16×16 pixels into 4 blocks of 8×8 pixels which are used for creating the base and the enhancement layers. Our motivation is based on our observation that the current spatial MDC methods for video assume a missing description can be interpolated from the remaining descriptions delivered intact. This assumption is not valid when the video contains objects of a high contrast with its background and a sharp boundary. Figure 3 depicts an example where the missing description is interpolated using the delivered descriptions. The dark points on bright areas of the pole (shown after zooming-in in Figure 4) are an example of this effect. Our proposed solution for this problem is described below.

The main idea in our proposed method is that when the descriptions are completely disjoint, interpolating the missing data (missing description) is carried out by utilizing the correlation between the pixels. However, the spatial decomposition of the frames can diminish this correlation resulting in lower fidelity of the reconstructed frame which in turn can cause drift error. In order to include the missing pixel values in the interpolation process and hence increase the spatial correlation between the pixels we introduce a base layer included in all descriptions. The base layer averages the values of the four descriptions in frequency domain. After decomposing a macroblock into four blocks, we motion-compensate each block, apply DCT transform and quantization, and compute the base layer which is included in all descriptions and the enhancement layers which carry the difference with the base layer. The base layer is obtained by finding the average of the quantized DCT coefficients of the blocks obtained by decomposing the macroblocks. Since each macroblock is decomposed into four 8×8 blocks, the base layer is also an 8×8 block where each element is the average of the coefficients at the corresponding positions of the four blocks of quantized DCT coefficient. Figure 5 depicts the block diagram of the proposed method where a thick arrow represents four outputs, BL refers to the base layer, and EL indicates the enhancement layer. The mathematical definition of the base and the enhancement layers is given in (3). The enhancement layer for each description is defined as the difference between the quantized DCT transform coefficients of a block and the quantized DCT coefficients of its base layer:

$$BL = \frac{1}{4} \sum_{i=1}^4 Q(DCT(Polyphase_i)), \quad (3)$$

$$EL_i = Q(DCT(Polyphase_i)) - BL,$$

where $Polyphase_i$ refers to the i th part of a block after its polyphase decomposition. The coefficients of the base layer and the enhancement layer are run-length and entropy encoded before transmission although they have not been shown in (3). In most cases the difference between the base layer DCT coefficients and the DCT coefficients of the block is very small. Hence, the enhancement layer does not add to the total bit-per-pixel rate of the descriptions. In some cases however, where the pixels values of the description are highly different from the average of the descriptions or the base layer, the enhancement layer will affect the bit-per-pixel rate. Reconstructing a block in presence of loss of one of the descriptions is carried out as follows.

Since the base layer is the average of the quantized DCT coefficients of all descriptions, the quantized DCT coefficients of the missing description can be found by subtracting the delivered descriptions from the base layer. The enhancement layer of the missing description is the difference of the coefficients obtained in this way and the base layer. This procedure shows that when a description is lost, the video is reconstructed using the remaining descriptions without any distortion. In case of data loss in more than one description, the missing descriptions of the block are interpolated by adding the average of the delivered enhancement layers to the base layer. When only one description is delivered, the proposed method is equivalent to using the delivered description in place of all missing descriptions. Equation (4) shows the interpolation in presence of more than one description loss:

$$Desc_i = BL + EL_i \quad \text{for } i = 1, \dots, n,$$

$$Desc_x = (n + 1) \times BL - \sum_{i=1}^n (Desc_i), \quad (4)$$

$$EL_x = Desc_x - BL,$$

where n is the number of delivered descriptions and EL_x and $Desc_x$ are the interpolated enhancement layer and the interpolated quantized DCT coefficients to be used for all missing descriptions, respectively.

Some important features of the proposed method are as follows.

- (i) In case of data loss in only one description, the proposed method can reconstruct the frame without any error. However, data loss in more than one description requires interpolation which is carried out as shown in (4).

- (ii) Although the proposed method introduces a redundant base layer, its performance in terms of bit-per-pixel when the video does not include high frequency content in object boundaries approaches the traditional polyphase MDC coding. This is due to the fact that the difference between the information transmitted in each description and the base layer (average of four descriptions) is small and hence the enhancement layers are very small.
- (iii) The proposed method provides the possibility of spatial and SNR scalability of video through decomposing each block spatially and encoding the data as base and enhancement layers. Spatial scalability is achieved by delivering only one description which will not result in any drift error. SNR scalability is achieved through delivering base layer only while this capability causes quality degradation due to drift error.

In [7] the authors propose a method which decomposes the video into multiple descriptions by redundantly transmitting a downsampled or low frequency version of the frame in all descriptions. Although the method proposed in this work is similar to the method described in [7], the algorithm for defining the enhancement layers and hence the interpolating and reconstructing video in presence of packet loss are different. The authors of [7] use DWT to create a low resolution common base layer and transmit the high frequency coefficients of each subband as enhancement layer at each description. In our proposed method the enhancement layer is the difference between the common base layer and the coefficients of the block being transmitted by that description. This lets us fully reconstruct the frame when a single description is lost.

4. Experimental Results

In the following paragraphs we introduce the experiments we have conducted to verify the performance of the proposed method.

4.1. Test Setup. The proposed method is experimentally verified using some video sequences. We have selected the video sequences in a way that they contain both low frequency smooth frames and high frequency contents. Table 1 lists the test videos and their respective properties.

The encodings are based on MPEG standard with the assumptions that the blocks of a frame have the same reference frame, and the GOP length is fixed to 16 with frame types of IBBBBBBBBBBBBBB. After polyphase decomposition of the macroblocks into blocks of 8×8 , each block is motion-compensated separately and hence has its own motion vectors. Downsampling ratio of chroma values is 4:4:4.

The set of experiments we have considered are as follows.

- (i) The proposed method defines a base layer which is repeated in all descriptions. The first experiment

TABLE 1: Video sequences used for performance evaluation.

Name	Rows $\times$ columns	Frame rate
Foreman	352×288	30
Stefan	768×576	30
Container	352×288	30
Deadline	352×288	30

FIGURE 6: PSNR values at different bit-rates using polyphase and the proposed method (Foreman sequence).

verifies the impact of this redundancy on the bit-per-pixel value of each test video. Since the changes in the bit-per-pixel value is dependent on the frequency content of each frame and in order to illustrate the changes more clearly, we have compared the bit-per-pixel values framewise in each video sequence.

- (ii) An important feature of our proposed method is its lossless delivery of the video when only one description is lost. In the second set of experiments, we compare the performance of our proposed method with interpolation methods.
- (iii) Our third set of experiments consider two- or three-description loss cases. We demonstrate the performance of proposed method vis-a-vis interpolation methods experimentally.

Figure 6 depicts the results of performance comparison between the proposed method and the polyphase decomposition of video when all descriptions are delivered intact.

The better performance of the polyphase method is due to the redundancy caused by repeated base layer in our method. The redundancy and reduction in PSNR value for any given bit rate are the price we pay for the better robustness against packet losses. As it is clear from Figure 6 the proposed method performs better (close to polyphase method) in low bit rates where the high frequency content of video is eliminated.

In our second set of experiments, we assume one description is lost in the entire video sequence. The description is computed using the proposed method and interpolated using

FIGURE 7: PSNR values at different bit-rates using interpolation and the proposed method (Foreman sequence) when one description is lost.

FIGURE 8: PSNR values at different bit-rates using interpolation and the proposed method (Stefan sequence) when one description is lost.

FIGURE 9: PSNR vis-a-vis bit-rates when two descriptions are lost (Foreman sequence).

averaging the delivered descriptions, bilinear interpolation, and edge sensing. As depicted in Figures 7 and 8, the proposed method outperforms interpolation methods. The performance differences in low bit rates however are very close. Besides, in videos with higher frequency content, the proposed method shows better performance (Figure 8).

Our final experiment evaluates the robustness of the proposed method in presence of more than one description loss. The experiment includes the case of two-description loss only. This is because of the reason that three-description loss reduces to replacing the video frames with the information from the delivered description only, which means no interpolation is carried out. The descriptions lost in the video sequence are randomly selected but remain fixed during the transmission. This assumption is compatible with the transmission error in a channel which may last for a few seconds causing loss of a description in consecutive frames. Figure 9 depicts the comparative results for the third experiment which indicates its superiority over interpolation methods.

The results of the experiments indicate that the proposed method outperforms the traditional interpolation methods in video coding in presence of frame losses. The proposed method includes the average of four descriptions in each one of them. This means that when two descriptions are lost, using the average of four descriptions and the enhancement layer of the delivered descriptions, we can retrieve the average of the enhancement layer of the lost descriptions. This property is the main reason for the better performance of the proposed method when more than one description is lost. The method proposed in [13] is compared with our proposed method. We consider two-description loss in our method but only one-description loss in the method proposed in [13] because our method reconstructs the block with no distortion when only one description is lost. The maximum reduction in our proposed method is 4.1 dB in PSNR while the method proposed in [13] can reach 8 dB PSNR quality loss. Figure 10 depicts the reduction in PSNR value of the frames in all test sequences. We have assumed two descriptions are lost in each GOP from a random position.

An important feature of our proposed method which needs clarification is that, in higher bit rates, the amount of high frequency content sent in each description increases. This increase results in larger enhancement layers which degrade the performance of the proposed method. However, having very different DCT coefficients in different descriptions (such as positive coefficients in one description and negative coefficients in the other) can only happen if the pixel blocks are highly different. Considering that the pixel blocks (8×8) used at each description are obtained by downsampling the same macroblock (16×16) through polyphase, in practice the enhancement layers are small. The large differences may happen when the macroblock is taken from the boundaries of objects with sharp contrast or very thin objects which is the main concern of our method. However, since these areas are proportionally small compared to the whole frame, the overall performance does not change dramatically.

As a subjective comparison, part of a frame from Stefan sequence has been reconstructed assuming two descriptions

FIGURE 10: PSNR reduction when two descriptions are lost in each GOP.

FIGURE 11: Original data (left), reconstruction using proposed method (middle), and reconstruction using bilinear interpolation (right).

are lost. Figure 11 depicts the original data (Y component), reconstruction using the proposed method, and reconstruction using bilinear interpolation.

5. Conclusions

A new method for spatially decomposing video into multiple descriptions is proposed. The proposed method addresses the quality degradation due to the low-pass filtering effect of interpolation whenever a description is lost. The proposed method is capable of recovering the video in a lossless form when one description is lost. This characteristic is coming with the cost of extra redundancy added to each description. In case of two-description losses, the proposed method outperforms interpolation methods. The performance difference between the proposed method and the interpolation methods increases with increase in bit-per-pixel value. This is an indication that the proposed method is more suitable for transmission of high quality video in presence of communication errors. Besides, the availability of a base and an enhancement layer in each description provides possibility of spatial and SNR scalability of video which make the method applicable for networks with bandwidth fluctuations. An extension of the method can be decomposing the video into more than 4 descriptions and combining the interpolation methods with the proposed method in estimating the enhancement layer data when more than one description is lost.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

- [1] A. Nafaa, T. Taleb, and L. Murphy, "Forward error correction strategies for media streaming over wireless networks," *IEEE Communications Magazine*, vol. 46, no. 1, pp. 72–79, 2008.
- [2] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, 2000.
- [3] C.-M. Fu, W.-L. Hwang, and C.-L. Huang, "Efficient post-compression error-resilient 3D-scalable video transmission for packet erasure channels," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 2, pp. 305–308, March 2005.
- [4] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–70, 2005.
- [5] R. Bernardini, M. Durigon, R. Rinaldo, L. Celetto, and A. Vitali, "Polyphase spatial subsampling multiple description coding of video streams with H264," in *Proceedings of the International Conference on Image Processing (ICIP '04)*, vol. 5, pp. 3213–3216, October 2004.
- [6] J. Jia and H. Kim, "Polyphase downsampling based multiple description coding applied to H.264 video coding," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 6, pp. 1601–1606, 2006.
- [7] T. shanableh, S.-T. Hsiao, and F. Ishtiaq, "Methods and apparatus for encoding and decoding video," U.S. Patent Application no. 12/108,680, 2008.
- [8] S. Gao and H. Gharavi, "Multiple description video coding over multiple path routing networks," in *Proceedings of the International Conference on Digital Telecommunications (ICDT '06)*, pp. 42–47, 2006.
- [9] O. Campana and R. Contiero, "An H.264/AVC video coder based on multiple description scalar quantizer," in *Proceedings of the 40th Asilomar Conference on Signals, Systems and Computers (ACSSC '06)*, pp. 1049–1053, Pacific Grove, Calif, USA, October–November 2006.
- [10] R. Venkataramani, G. Kramer, and V. K. Goyal, "Multiple description coding with many channels," *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2106–2114, 2003.
- [11] V. K. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, 2001.
- [12] N. Franchi, M. Fumagalli, G. Gatti, and R. Lancini, "A novel error-resilience scheme for a 3-D multiple description video coder," in *Proceedings of the Picture Coding Symposium (PCS '04)*, pp. 373–376, December 2004.
- [13] W.-J. Tsai and J.-Y. Chen, "Joint temporal and spatial error concealment for multiple description video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 12, pp. 1822–1833, 2010.
- [14] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

- [15] C.-S. Kim and S.-U. Lee, "Multiple description coding of motion fields for robust video transmission," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 9, pp. 999–1010, 2001.
- [16] V. K. Goyal and V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, 2001.
- [17] S. Cen and P. C. Cosman, "Decision trees for error concealment in video decoding," *IEEE Transactions on Multimedia*, vol. 5, no. 1, pp. 1–7, 2003.
- [18] Y. Wang, M. T. Orchard, and A. R. Reibman, "Multiple description image coding for noisy channels by pairing transform coefficients," in *Proceedings of the IEEE 1st Workshop on Multimedia Signal Processing*, pp. 419–424, Princeton, NJ, USA, June 1997.
- [19] N. Memon and X. Wu, "Recent developments in context-based predictive techniques for lossless image compression," *Computer Journal*, vol. 40, no. 2-3, pp. 127–136, 1997.