Multiple Description Wavelet Coding for Scalable Video Transmission

†‡Roya Choupany roya@cankaya.edu.tr
†Stephan Wong J.S.S.M.Wong@tudelft.nl
‡Mehmet Tolun tolun@cankaya.edu.tr
†Delft University of Technology, Delft, theNetherlands
‡Çankaya University, Ankara Turkey

***Abstract*—Scalable video coding (SVC) and multiple description coding (MDC) are the two different adaptation schemes for video transmission over heterogenous and best-effort networks such as the Internet. We propose a new approach to encode video for unreliable networks with rate adaptation capability. Our proposed method groups 3D discrete wavelet transform coefficients in different descriptions and applies a modified embedded zero tree data for rate adaptation. The proposed method reduces the impact of the drift error by organizing the frames in a hierarchical structure.**

***Index Terms*—Scalable Video Coding, Multiple Description Coding, Multimedia Transmission**

I. Introduction

Heterogeneity in current-day networks (especially in the Internet), the unpredictability of traffic loads, and the varying delays on the client side, make it impossible to correctly determine a specific bit rate for a video stream [1]. Consequently, the encoder should either consider the lowest possible bit rate that guarantees delivery without delay or choose an encoding scheme which can adapt with the fluctuations in the bit rate range. This means that it should be possible to partially decode the video stream at the incoming bit rate and at the video quality associated with that bit rate. A solution to this problem is encoding the video data in a rate scalable scheme for enabling adaptation to the receiver rendering device or network data rate capacities. Increasing the video quality gradually is the common characteristic of all Scalable Video Coding (SVC) schemes. The quality increase is accomplished though the gradual increased availability of the data units that were encoded in a granular manner. It is clear that a gradual increase in the frame size, bit rate, or frame rate is achieved through adapting the granularity of a stream to the bit rate capability of the network. A fine granularity scalability scheme defines the video content in a multi-layer format where the existence of at least one layer (the base layer) containing the most basic data is required. The remaining layers (enhancement layers) add to the quality of the video [5]. A higher quality for a video is attained by increasing the number of layers decoded at the receiver side. The most important drawback of the multi-layer coding schemes is a problem called drift error. Drift error arises when the encoder uses the enhancement-layer information in the motion-prediction loop and the enhancement-layer information is not received by the decoder. Drift error can degrade the quality of the video dramatically. A similar problem may occur if parts of the transmitted data are lost due to network communication failure. Forward Error Correction (FEC) codes or Multiple Description Coding (MDC) usage are the most common solutions to the data loss problem. FEC methods are in contrast with the goal of SVC as they increase the data size by incorporating extra bits in the original data for error detection. In this paper we are proposing a method which combines scalable video coding techniques with multiple description coding for reducing the quality degradation in presence of packet loss or video scale-down. We also address the drift problem by introducing a hierarchical structure for reducing the length of the frame prediction chain in a group of pictures. This work is an extension to our paper published in the 7^{th} International Conference on Digital Content, Multimedia Technology and its Applications [**?**]. The remaining parts of this paper have been organized as follows: Section II summarizes the previous work on reducing drift error effect in SVC. We also discuss the MDC methods used for reducing the packet loss error effect in this section. Section III gives the details of our proposed method. Section IV provides the experimental results of our method. We draw conclusions in Section V.

II. Related Work

The drift error problem which is common to all multi-layer scalability schemes has been addressed by several researchers. H.264 standard tries to control the drift error by introducing a new concept called key frames [2]. Key frames are not necessarily intra-coded frames. By

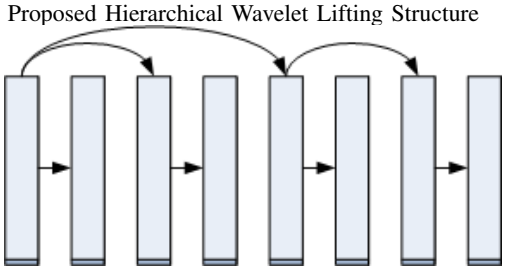
introducing a hierarchical reference frame organization, H.264 allows all enhancement layer frames to utilize the references with the highest available quality for motion estimation, which enables a high coding efficiency for these key frames [9]. H.264 standard does not eliminate the drift error completely but minimizes its effect and limits it to the frames between two consecutive key frames [6]. The tradeoff between enhancement layer coding efficiency and drift error can be adjusted by the choice of the number of frames between two consecutive key frames or the number of hierarchy stages. The discrete wavelet transform (DWT) has been used for scalable encoding of still image and video multimedia as described in JPEG2000 [**?**]. Spatial oriented trees such as Embedded Zero Tree Wavelets (EZW) and Spatial Partitioning in Hierarchical Trees (SPIHT) are used for organizing wavelet coefficients in their importance order for scalability [**?**], [**?**]. 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 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 [10]. 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. 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 put 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. 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., [19] 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 [7]. 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 [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 $S_1; ..; S_n$ 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.

III. PROPOSED METHOD

Our proposed method involves using the scalability features of discrete wavelet transforms. The frames of a GOP pass through a Haar lifting stage. The splitting and prediction steps of the lifting process are repeated in several stages in a hierarchical structure. The drift error

is reduced considerably by applying this hierarchical structure. The general view of the hierarchical structure and the applied lifting method are depicted in Figure 1. The frames in a group of pictures (GOP) are organized

Fig. 1. Proposed Hierarchical Wavelet Lifting Structure



in pairs and the second frame in each pair is predicted from the first frame. The first frames of the pairs from the first level are grouped in the next level of the hierarchy in pairs and the same prediction and wavelet encoding steps are applied to them. This means the first frames of the pairs which serve as the reference frames for the second frames at the same pair, are processed at a higher level where they are finally positioned as the second frame of a pair. This procedure is repeated at the following levels of the tree hierarchy. If only the lowest layer of the hierarchy is considered, the drift error is limited to one frame as the second frame at each pair is predicted and obtained using the first frame of the same pair. However, any accuracy change in the second layer affects the first frames of each pair in the lowest layer and therefore the error is accumulated. The worst case situation is when error is introduced in the topmost layer of the hierarchy which affects the whole tree. However, in this case the number of frames in a series of frames referencing each other is limited to the tree height and, therefore, the GOP size and hence tree height should be determined in a tradeoff with the max tolerable drift error. This structure reduces the drift error in a logarithmic manner. The proposed structure falls in the group of non-delay methods where no frame need to be buffered till the arrival of the following frame(s) for decoding. This makes the decoder implementation simple, with minimum memory requirement. Three levels of 2D wavelet transform is applied to the frames of each GOP after Haar lifting and organizing in hierarchical structure. The wavelet coefficients of each description are quantized and zig-zag scanned before transmission to put them in information importance order. This makes it possible for a receiver or a network node to truncate parts of data in case of low bandwidth or processing power. Modified embedded zero trees (EZW) are used to arrange the coefficients. In our modified zero tree, the coefficients are grouped in four descriptions and the zero tree is used for storing the coefficients in each description. The differences between the original zero tree and the proposed modified zero tree are two folds.

1) The zig-zag scanning is performed according to the order depicted in Figure 2.
2) Each branch of the tree corresponds to one the descriptions and contains the low frequency part of the coefficients.

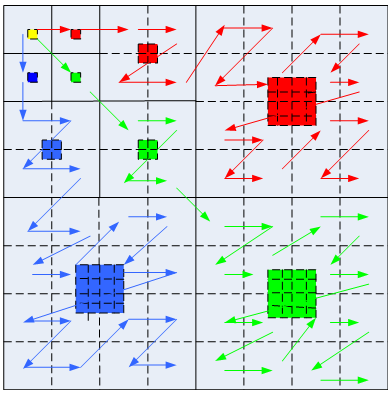


Fig. 2. Zig Zag Scanning of DWT Coefficients at each Description

The following subsections discuss the type of scalabilities provided by the proposed method.

A. *Bit rate scalability*

Bit rate scalability in the available scalable video coding standards such as MPEG-4 or H.264 is defined as multi-layer coding of the frames where the number of layers determines the granularity of the video. The multi-layer bit rate scalability suffers from the problem that using enhancement layer data in motion compensation loop may result in drift error. On the other hand motion compensation using base layer only can reduce compression efficiency. For example, the fine granularity quality scalable (FGS) coding in MPEG-4 was chosen in a way that drift is completely omitted by using base layer frames as reference frames in motion compensation and, therefore, any loss or modification of a quality refinement packet does not have any impact on the motion compensation loop. The number of layers in these standards should also be limited as carefully designed since the multi-layer concept for quality scalable coding becomes less efficient, when the relative rate difference between successive layers gets smaller. As our proposed method is based on discrete wavelet transform, multi-layer restrictions are not effective here. The number of bits used for representing wavelet transform coefficients of the motion compensated residues is reduced for a lower bit rate transmission over a low bandwidth channel. This goal can also be achieved by organizing the

wavelet coefficients in a spatial tree structure such as EZW or SPIHT in which cases, the transmitted bitstream is truncated considering the bandwidth capacity of the channel.

B. Spatial scalability

Scalability in frame resolution was introduced with the MPEG-2 standard. A multi-layer structure was considered for implementing the scalability where the base layer contained the lowest resolution and a higher resolution frame was reconstructed by upsampling the base layer data and adding the refinements from upper layers. The main problem with this method is integrating it with the motion compensation loop. Using high resolution frames for motion estimation can reduce the compression rate when only a low resolution frame is reconstructed at the receiver side. Motion estimation in low resolution frames on the other hand causes drift error problem as the difference frames obtained from low resolution base frames do not include any information about the eliminated rows and columns. One way to reduce the drift error is by limiting the length of a group of pictures in a series of motion compensated frames without reducing the compression efficiency. Our proposed method performs this by organizing the frames in a hierarchy and therefore is suitable for these types of scalability applications.

C. Temporal scalability

Temporal scalability in the traditional video coding methods is achieved through placing some of the frames in the base layer and the rest in the enhancement layer(s). An example is dividing the frames of a GOP by putting even numbered frames in base layer and odd numbered ones in the enhancement layer. A drift error problem will appear if the motion compensation involves the frames of the enhancement layer, if the receiver is capable of reconstructing the video in a lower frame rate. In our method a 50% temporal scalability is achieved by dropping the second frames of each frame pair at the lowest level of the hierarchy. This scalability is accomplished without any drop in the compression efficiency or drift error problem. A higher rate of scalability is possible by eliminating the second frames at the next level. The number of levels in the tree hierarchy is not an upper or a lower limit on the temporal scalability. This fact is described by considering the characteristic of the proposed hierarchy where the frames of a series are not chained together in a linear structure. This means that any possible rate of temporal scalability is achievable by eliminating only some of the second frames of the frame pairs. The temporal scalability achievable at each level can be expressed using Equation 1.

$$SR_n = \sum_{i=1}^{n} \frac{1}{2^i} \tag{1}$$

where SR_n refers to scalability rate at level n.

IV. Experimental Results

We implemented the proposed method using MATLAB. Some implementation considerations are as follows:

- The number of frames per GOP in our experiments we fixed at 32,
- The Biorthogonal 4.4 DWT kernel is used (bior4.4),
- For EZW coding of the wavelet coefficient, at each description we are replacing the coefficients belonging to other descriptions with zero,
- Finally we are encoding the EZW codes using Huffman encoding.

Replacing wavelet coefficients belonging to neighboring descriptions helps us to use the standard zig-zag scanning of the EZW method. The zeros added in this way are replaced by a zero tree symbol and do not have any significant impact on the bit per pixel rate of the method. Our evaluation is based on changing the initial threshold value of EZW coding and computing the fidelity of the frame using PSTN criteria. The computed PSTN values for different thresholding levels are drawn with respect to the bit per pixel obtained. We compute bit per pixel using the total length of the obtained code divided by the number of pixels in the frame. To verify the performance of the proposed method we have considered three video sequences with the specifications given at Table I. We have tried to include videos with high and low frequency contents for a better evaluation.

Hierarchical coding of the frames with 32 frame per group of picture is applied to each video sequence. The hierarchical structure is wavelet transformed and split into three description. The coefficients in each description is then coded using embedded zero tree wavelet (EZW) method. The test cases are devised in a way that both error resilience and scalability of the

TABLE I
VIDEO SEQUENCES USED FOR PERFORMANCE EVALUATION.

Name	Rows × Columns	Frame rate
Foreman	352 × 288	30
Stefan	768 × 576	30
Suzie	144 × 176	30

proposed method can be evaluated. We consider four cases for evaluation of the method as listed below:

1) The effect of the loss of a description on the reconstruction of the video,
2) The quality degradation due to loss of a description in a frame and the drift error effect on reconstructing the remaining frames of the GOP,
3) The drift error effect due to down scaling the video,
4) The drift error effect due to down scaling the video in presence of a description loss.

For the case of video reconstruction in presence of a description loss we reconstruct the video using two out of three descriptions. The reconstructed frames are compared with the original frames using PSNR values. The computation is carried out by putting aside one of the descriptions each time and the average of the resulted PSNR values is computed. In all experiments, the lost coefficients are replaced by zero when we perform inverse DWT. Table II shows the PSNR values for different cases of description losses at each of the test video sequences. The notation $PSNR_{ij}$ indicates that descriptions i and j have been received. The first experiment assumes all

TABLE II
PSNR VALUES FOR DIFFERENT CASES OF DESCRIPTION LOSSES

Video Sequence	$PSNR_{23}$	$PSNR_{13}$	$PSNR_{12}$	$PSNR_{Avg}$
Foreman	36.62	37.13	36.78	36.84
Stefan	35.88	35.73	36.03	35.88
Suzie	38.49	38.76	38.56	38.60

packets of one of the descriptions are lost. Hence The reconstruction is carried out by using the remaining data. This is an example of a burst error case. However, there exists the possibility of a single packet loss. Here we are assuming that a packet carries the data of one description in a frame. Even a single packet loss as expressed above can cause degradation in the quality of the reconstructed video. The quality degradation is the result of a drift error effect due to a change in one frame. The hierarchical structure used aims at minimizing the drift error effect. However, repeating part of the data (base layer) at each descriptions has also the impact of reducing the quality reduction effect. Our second experiment measure the quality degradation by randomly choosing one frame from each GOP, putting aside one description of it, and reconstructing the video. The PSNR values of the reconstructed frames occurring after the frame with a missing descriptions are computed and averaged for all GOPs in each video sequence. Table III shows the computed averages for each sequence separately.

TABLE III
PSNR VALUES IN PRESENCE OF A FRAME LOSS IN A SINGLE DESCRIPTION

Video Sequence	$PSNR_{23}$	$PSNR_{13}$	$PSNR_{12}$	$PSNR_{Avg}$
Foreman	40.35	40.93	40.68	40.65
Stefan	38.97	38.73	39.10	38.93
Suzie	41.49	40.87	41.53	41.30

The scalability capability of the proposed method is also verified in our experiments. The required bit rate determines the threshold for the number of EZW symbols we transmit and use for reconstructing the video. A slightly smaller PSNR values we obtain compared to the case when the video is coded as a single description [8] which is a result of added redundancy of repeated base layer. The drop in PSNR value between the multiple description and single description coding is also a function of the number of DWT levels used because increasing the number of DWT levels decreases the size of the base layer and hence a smaller redundancy is imposed. However, reconstruction error in presence of a description loss increases if the base layer is small. Figure IV depicts the results of the reconstructing video as PSNR with respect to bit rate. Here we have assumed all descriptions are received without error. In our last

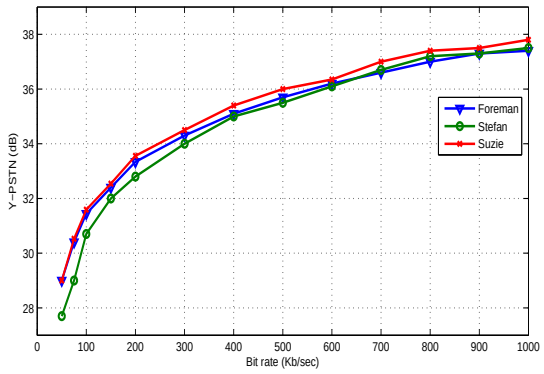


Fig. 3. Video reconstruction with respect to bit-rate using all descriptions

experiment we combine the scalability with description loss. The experiment is conducted by changing the threshold value of EZW encoder to obtain different bit rates. Then we reconstruct the video using only two out of three descriptions. We compute the PSNR for the reconstructed video and consider the average of the PSNR values of reconstructed video with one of the descriptions omitted. Figure IV depicts the obtained result from three video sequences. It should be noted that the bit rate in the last experiment is obtained from

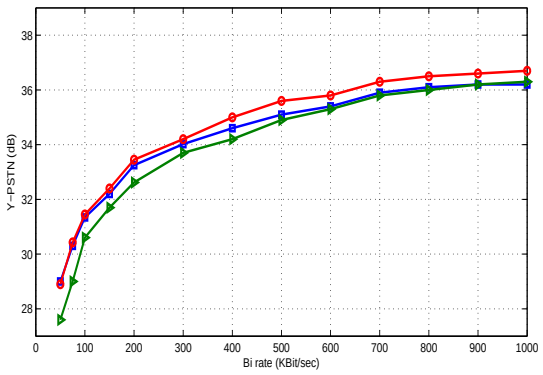


Fig. 4. Video reconstruction with respect to bit-rate in presence of a description loss

the total number of bits contained in two delivered descriptions.

V. Discussion

The combination of hierarchical coding and multiple description has the advantage of reducing the impact of partial data loss while providing the possibility of receiving video in lower bit rate by the receiver. On the other hand any change in the transmitted data, either due to data loss or omitting part of data intentionally for scalability, affects all frames coming afterwards. A combination of a scalable coding method, an error resilience transmission method, and a drift error effect reducing method is proposed. The proposed method is suitable for video transmission over heterogenous and best-effort networks such as the Internet. The proposed method provides the bit rate scalability by reducing the quality of the video whenever the transmission line suffers form a narrow bandwidth problem. A possible extension of the method is providing spatial scalability as many portable wireless devices come with low resolution screens.

References

[1] G. Conklin, G. Greenbaum, K. Lillevold, A. Lippman, and Y. Reznik, "Video Coding for Streaming Media Delivery on the Internet", IEEE Trans. Circuits and Systems for Video Technology, March 2001.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of Scalable Video Coding Extension of the H264/AVC Standard", IEEE Transaction on Circuits and Systems for Video Technology, Vol. 17, No. 9, September 2007.

[3] D. Marpe, T. Wiegand, and G. J. Sullivan, The H.264/MPEG4 advanced video coding standard and its applications, IEEE Commun. Mag., vol. 44, no. 8, pp. 134144, Aug. 2006.

[4] G. J. Sullivan, H. Yu, S. Sekiguchi, H. Sun, T. Wedi, S. Wittmann, Y.-L. Lee, A. Segall, and T. Suzuki, New standardized extensions of MPEG4-AVC/H.264 for professional-quality video applications, presented at the ICIP, San Antonio, TX, Sep. 2007

[5] H. Jiang, "Experiment on post-clip FGS enhancement", ISO/IEC JTC1/SC29/WG11, MPEG00/M5826, March 2000.

[6] A. Segall, CE 8: SVC-to-AVC Bit-Stream Rewriting for Coarse Grain Scalability, Joint Video Team, Doc. JVT-V035, Jan. 2007.

[7] M. Wien, H. Schwarz, and T. Oelbaum, Performance analysis of SVC, IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 9, pp. 11941203, Sep. 2007

[8] R. Choupani, S.J.M. Wong, M. R. Tolun, "A Drift-Reduced Hierarchical Wavelet Coding Scheme for Scalable Video Transmissions," First International Conference on Advances in Multimedia (MMEDIA), pp.68-73, 2009

[9] H. Kirchhoffer, H. Schwarz, and T. Wiegand, CE1: Simplified FGS, Joint Video Team, Doc. JVT-W090, Apr. 2007.

[10] R. Choupani, J.S.M. Wong, and M. R. Tolun, "Multiple Description Scalable Coding for Video Transmission over Unreliable Networks", Embedded Computer Systems: Architectures, Modeling, and Simulation, 9th International Workshop, SAMOS 2009, Samos, Greece, July 2009, pp.58–67.

[11] 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.

[12] J. Ohm, "Advances in Scalable Video Coding", Proceedings of the IEEE, vol. 93, no. 1, Jan. 2005.

[13] Y. Wang, A. R. Reibman, L. Shunan, "Multiple Description Coding for Video Delivery", Proceedings of IEEE, vol. 93, No. 1, Jan. 2005.

[14] 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.

[15] 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 .

[16] 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.

[17] 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

[18] 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.

[19] 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.

[20] 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.