Scalable Video Coding

Choupani, Roya

DOI
10.4233/uuid:1b75db0a-60da-4a6c-8c02-6cc2550c90d0

Publication date
2017

Document Version
Final published version

Citation (APA)
Choupani, R. (2017). *Scalable Video Coding*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:1b75db0a-60da-4a6c-8c02-6cc2550c90d0

Scalable Video Coding

Scalable Video Coding

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof.ir. K.C.A.M. Luyben;
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
dinsdag 13 Juni om 12:30 uur

door

Roya CHOUPANI
Master of Science in Computer Engineering
geboren te Kermanshah, Iran

This dissertation has been approved by the
promotor: Prof.dr. K. L. M. Bertels
co-promotor: Associate Prof. dr. Stephan Wong

Composition of the doctoral committee:

Rector Magnificus	Chairman
Prof.dr. K. L. M. Bertels	Delft University of Technology, promotor
Associate Prof. dr. Stephan Wong	Delft University of Technology, co-promotor

Independent members:

Prof. dr. Jarma Takala	Tampere University of Technology, Tampere, Finland
Prof. dr. Mehmet Tolun	Aksaray University, Aksaray, Turkey
Prof. dr. Luigi Carro	UFRGS University, Porto Alegre, Brazil
Prof. dr. ing Michael Huebner	Ruhr University, Bochum, Germany
Prof. dr. Said Hamdioui	Delft University of Technology

ISBN: 978-94-6186-798-8

To the warriors of the fight against ignorance

Abstract

With the rapid improvements in digital communication technologies, distributing high-definition visual information has become more widespread. However, the available technologies were not sufficient to support the rising demand for high-definition video. This situation is further complicated when the network resources such as the available bandwidth fluctuates, or packet losses occur during transmission. In this dissertation we present several video compression techniques which are capable of adapting with the varying network conditions. We address both challenges namely, the fluctuations in the available resources such as the bandwidth and processing power, and packet losses. These problems in turn translates into degradation of the perceived video playback as jitter, and delay before video playback starts. Hence, we concentrate on developing robust and fast adaptive video coding schemes necessary for handling the changes in the physical characteristics of the communication networks. We present a new multi-layer scalable video coding (SVC) method for optimizing the bit-per-pixel rate of the video which is robust against packet losses. The method reduces the quality degradation in presence of data loss by re-organizing the frames in a hierarchical structure and improving the video quality through decomposing each frame suitably to restrict the error propagation. Moreover, we present a solution for the quality degradation in video reconstruction when the video is scrambled for privacy protection. We also present two methods based on multiple description video coding (MDC) to handle packet losses in networks with a high rate of transmission error. The proposed methods are based on combining SVC with MDC through decomposing the video into spatial sub-streams in the first method, and SNR sub-streams in the second method. In both proposed methods, the error resilience of the video is increased. The proposed methods have the capability of being used as SVC methods where any data loss or corruption reduces the quality of the video in a minimized way, and except for the case when all descriptions are lost, the video streams do not experience jitter at playback. The proposed methods provide the feasibility of reducing data rate by scaling down the video whenever the connection suffers from a low bandwidth problem. We also propose Discrete Wavelet Transform (DWT)-based optimizations for MDC. A major drawback in MDC methods is their inefficiency in terms of bit-per-pixel which is a consequence of preserving correlation between decomposed video segments. We propose a method based on the self-similarity between DWT coefficients at different frequency levels to improve the coding efficiency of DWT-based MDC. In the proposed method, whenever a description is lost the coefficients at the delivered descriptions are utilized for estimating the missing data using self-similarity property.

Acknowledgements

First and foremost, I would like to express my grateful appreciation to Stephan Wong, Mehmet Tolun, and Koen Bertels for their unceasing support during this period. Their guidelines and consisting support provided me the chance to pursue my PhD research.
My special thanks go to my family for supporting and encouraging me, and to my friends Faisal Nadeem, Imran Ashraf, Assadollah Shahbahrami, Mitra Abasfard, Reza Shams, Mahmood Ahmadi and , Peyman Pouyan for their helps.
I am specially grateful to Helen Skinner, Hans Weijtmans, and Cahide Tolun who have assisted me along the way through moral and emotional supports and encouragements.
I would like to thank Cankaya University administrators for supporting this study. Finally I want to express my deep grateful feelings to the memory of Professor Stamatis Vassiliadis who provided this opportunity.

Contents

1 Introduction **3**
- 1.1 Adaptive and Error-robust Video Coding Techniques . . . 4
  - 1.1.1 Fundamentals of Video Coding . . . 4
  - 1.1.2 Adaptive Video Coding Methods . . . 7
  - 1.1.3 Error-Robust Video Coding . . . 15
- 1.2 Video Adaptation Challenges . . . 16
- 1.3 Research Directions . . . 17
- 1.4 Main Contributions . . . 18
- 1.5 Thesis Organization . . . 20

2 Minimizing Drift Error in SVC **21**

3 Optimizing MDC in Spatial Domain **35**

4 Using DWT for Optimizing MDC **73**

5 Conclusions **91**
- 5.1 Summary . . . 91
- 5.2 Future Research Directions . . . 93

Chapter 1

Introduction

Visual information has become a vital part of our daily lives. This information is generally accessed through computer networks with varying bandwidths and packet loss rates which degrade the quality of the delivered videos. Hence, adaptability in these communication conditions is an essential requirement for videos being transmitted with time constraints. Confronting the impacts of bandwidth variations and packet losses are the two challenges addressed in this research. Section 1.1 of this chapter introduces the current techniques of adapting videos with bandwidth variations, and improving their robustness against packet losses. In Section 1.2, we present the challenges in applying these techniques and their shortcomings. In Section 1.3, we present our approach of addressing the challenges considered in this research. Section 1.4 presents the main contributions of this thesis. Finally, in Section 1.5, we give the organization of the thesis.

Humans heavily rely on visual information. Therefore, many technologies were introduced to distribute video content. However, the available technologies were not sufficient to support high-definition video. Moreover, the demand for high-definition video is not expected to slow down and with the increased connectivity of humans on Earth, the available bandwidth is even further stretched despite continuous improvement in network technologies. Therefore, solutions were sought after to fit increasingly higher quality video content on the current network technologies leading to compression techniques. However, traditional video compression methods assume fixed properties of the video content, e.g., temporal and spatial resolution, i.e., frame rate and frame resolution, respectively. This is in contrast with the varying physical characteristics of the communication networks (e.g., bandwidth) as well as the display properties of the consumer's device(s). This is further exacerbated by upcoming streaming and on-demand services as they require the display of video content in real-time. Consequently, adaptive video coding is needed to deal with these (new) requirements.
In addition, communication networks are not ideal and leading to potential loss of packets that are used to transmit the video information. This in turn translates to degradation of the perceived video playback as jitter, delay before video playback starts, and frame losses. Hence, a robust and fast adaptive video coding scheme is necessary for handling the changes in the physical characteristics of the communication network, and diversity of the receiver end rendering power and characteristics of its display device.

1.1 Adaptive and Error-robust Video Coding Techniques

It is essential for videos to be capable of coping with the variations in the network bandwidth fluctuations and/or packet losses in video streaming. The bandwidth fluctuations which happen even in reliable networks where packets are delivered intact, reduce the network data rate and increase the video transmission delay. This delay is not tolerated in real-time video streaming applications, and hence, for faster transmission the video is adapted to the network bandwidth by lowering its quality. In this adaptation, the essential part of video data is preserved to reduce the degradation in video quality. On the other hand, in case of packet losses which happen in unreliable networks, we cannot select to preserve more essential data because packet loss is a random process. The main issue in this case is minimizing the impact of packet losses on the reconstructed video. In both cases however, video suffers from quality degradation. This degradation is the result of the dependency between encoded video frames. In order to describe how the quality degradation occurs during video adaptation, this section provides the fundamental concepts in digital video coding (Section 1.1.1). We also briefly introduce some adaptive (Section 1.1.2), and error-robust (Section 1.1.3) video coding techniques.

1.1.1 Fundamentals of Video Coding

Digital video coding is the fundamental technology for many applications, such as digital TV broadcasting, distance learning, video on demand, video telephony, and video conferencing. The problem of storing and transmitting digital video has been the topic of research for many years [3], [25], [26]. A typical video standard with a 720×576 frame resolution, 24 bits per pixel, and 30 frames per second, contains 37,324,800 bytes in one second, and about 128 gigabytes in one hour video. These numbers in 1080p (1920×1080 progressive scan) with the same frame rate are 186,624,000 bytes in one second, and about 650 gigabytes in one hour video. Considering the average data rate of the Internet services using ADSL connections which is about 10 Mbps[1], the videos described above should be compressed at least 35 and 170 times, respectively. These numbers are indications of the need for video compression in high ratios. Hence, the digital video coding problem is considered as investigating methods of representing a video in a size-efficient binary form and therefore it can be classified as an image compression problem in which redundancies found within and between images are exploited to reduce the overall size of images [27]. The feasibility of compressing video data however, depends on the amount of redundancy present. We distinguish the following types of redundancy:

- *Psychovisual redundancy*: The human visual system is more sensitive to the intensity of light rather than color. Therefore, a color model (YCbCr) is utilized that separates the brightness (Y) and color information (CbCr) in distinct components. The resolution of the CbCr components can be reduced (compared to the Y component), via a technique called downsampling, without the human eye being able to (subjectively) notice any difference. This reduction in resolution results in a greatly reduced need to store the ”unnecessary” information.

- *Spatial redundancy*: Spatial redundancy refers to the correlation between pixels within a single video frame. For instance the pixels in a column/row are highly correlated with the pixels in the adjacent column/row. Similarity between neighboring pixels is a result of having uniform areas where color and intensity are changing very slowly. This similarity can be exploited by a variety of techniques to achieve compression. The simplest example is predictive coding that 'predicts' the neighboring pixel value to be the same as itself. The correlation dictates that this

[1]The average actual download speed in the Netherlands has grown from 8.97 Mbps in the third quarter of 2009 to 15.6 Mbps in 2015, according to Akamai State of the Internet Report.

probability is high or at least the value is close and, therefore, the (value) difference should be small. In turn, the coding of the difference requires less bits than coding the original pixel value. A more complex example is the discrete cosine transform (DCT) that transforms a 2D array of pixels from the spatial domain into the frequency domain. Low-frequency coefficients represent the similarity of the pixels and high-frequency coefficients represent large value differences of the pixels (in the spatial domain). This coding method is combined with the fact that the human eye cannot notice large (pixel) value differences when they are closely located, i.e., more importance is placed on the low-frequency coefficients rather than the high-frequency coefficients. Consequently, high-frequency coefficients are less probable and if present, less precision can be used to represent them. The reduction of precision is achieved by the so-called quantization step, i.e., dividing the coefficients with pre-determined quantization factors. This leads to storing less information, thus, compression.

- *Temporal redundancy*: Temporal redundancy refers to the similarity of information between pixels from different, but still temporally close video frames. This would be exactly the same as spatial redundancy if there was no movement of the frame, i.e., panning, or of objects within the frame. Therefore, the same techniques in exploiting spatial redundancy can be utilized, but they must be preceded with one additional step - motion estimation. In this step, the motion is estimated first to improve the similarity of the to-be-encoded frame. The latter is usually divided into blocks when performing this step i.e. block-based motion estimation. Consequently, additional information must be stored, namely the motion vectors, but this is usually less than the 'additional' compression achieved.

- *Stochastic redundancy*: After the previously mentioned redundancies have been exploited to achieve compression, the resulting information (pixel differences or quantized DCT coefficients) can still exhibit a non-uniform distribution. For example, the value zero (no differences) is much more likely to be present than high difference values. This non-uniformity can be exploited by variable length coding (VLC) or arithmetic coding (AC) techniques to achieve further compression. For completeness, we have to mention here run-length coding techniques that compress strings of similar values into symbols as the first step before VLC or AC techniques are applied. This is especially useful when considering that the earlier mentioned quantization step results in a large amount of zeroes.

Both still image and video coding techniques utilize psychovisual, spatial, and stochastic redundancies for more effective compression. However, temporal redundancy is used only in video coding techniques as the frames of video are captured at different time instances.
The general procedure of redundancy elimination starts by dividing a frame into fixed-size blocks named macroblocks. Then the psychovisual redundancy is eliminated by transforming each block into YCbCr color space and downsampling Cb and Cr components (as explained above).[2] The Y component is further divided into four equal-sized blocks. Hence, after YCbCr transform and downsampling, six equal-sized blocks are obtained. [3] Subsequently, each block is motion compensated by finding its most similar area in a reference frame (motion estimation) and computing their differences (residues). These residues are DCT transformed to eliminate spatial redundancy. As we mentioned above, higher compression rates are achieved not only by eliminating redundancies, but also by reducing video quality through quantizing the DCT coefficients. Quantization is performed by dividing

[2]It should be noted that video coders provide the options of downsampling Cb and Cr in one or two dimension, downsampling Y component, or preserving all three components without applying any downsampling.

[3]More recent video coding standards such as H.264 use various block sizes such as $8 \times 4, 4 \times 4$, etc.

the DCT coefficient with a constant value (named quantization step size) and rounding the quotients to the nearest integer value to obtain a block of rounded and quantized coefficients named quantized coefficients matrix. The larger the quantization step size used in quantization, the smaller the quantized values are and hence, the compression rate is higher. However, quantization causes data loss and degrades the quality of the video. Therefore, the DCT coefficients corresponding to higher frequencies are quantized using larger quantization step sizes. After quantization, generally a large number of zeros are generated which are encoded using VLC for a better compression rate which eliminates stochastic redundancy. Two stages are considered for compacting the zero values for better compression. Given that the quantization step sizes for high and low frequency coefficients are not equal, we are expecting to have most of the zero values at the lower right side of the quantized coefficients matrix. Hence, in the first stage the coefficients are put in the order of low to high frequency by using a zig-zag scanning of the coefficients from upper-left to lower-right corners of the DCT coefficient matrix. Subsequently, the zero values at the end of the list are eliminated by inserting an end-of-block symbol (EOB) after the last nonzero coefficient. The second stage applies a run-length encoding to the list of quantized coefficients by grouping them as codewords of (run, value) where *run* is the number of zeros, and *value* is the succeeding non-zero coefficient. Since the codewords have a non-uniform distribution and their probability of occurrence are not equal, there exists a statistical redundancy. VLC is applied to the codewords to remove the statistical redundancy for further compression. Eliminating temporal redundancy in video coding helps reaching high compression rates however, it poses a major challenge in video adaptation as explained below.

Motion compensated video coding techniques require the reference frame(s) to be present before reconstructing a frame. The reconstruction of a frame using its preceding frame(s) creates a chain of frames that depend on each other. In the following sections, the term dependency chain is utilized in order to refer to this concept. Therefore, whenever the reference frame used for motion compensation is not available (due to packet losses, for instance), or has been modified the reconstruction fails or suffers from degraded quality, respectively. This quality degradation is propagated to the succeeding frames in the dependency chain to further deteriorate the quality. This accumulated error is called the drift error. In order to restrict how far the drift error can propagate the frame dependency chain boundaries must be set up. Video coding standards use grouping of the frames as a method to set up these boundaries. The sequence of frames in a video is divided up into groups of pictures (GOP). Each GOP starts with an intracoded-frame (I-frame) and contains some predictive frames (P-frame), and/or some bi-directional frames (B-frame). An I-frame is encoded independently from other frames and are not motion-compensated. P-frames are motion compensated and encoded by finding their differences with (a) previous reference frame(s) which is/are not necessarily their immediate preceding frame(s). However, the reference frames should be from the same GOP. B-frames are coded by comparing the values of each block with previous and next I- or P-frames, and by utilizing the average of these differences for coding. B-frames provide better coding efficiency than P- and I-frames. Since B-frames use their preceding and succeeding frames for motion compensation, a re-ordering of the frames is used for a faster decoding, although this re-ordering increases the required memory by the decoder. Figure 1.1 depicts the frame ordering in GOPs.

It is worth noting that I-frames are not motion-compensated and hence, their coding efficiency is low. Therefore, the larger the size of the GOPs, the better the compression efficiency will be. However, large GOP size means random access to the video frames will be more limited. Meanwhile, in case of a problem in reconstructing a frame, the error propagates to more frames. In addition, during the motion estimation of a block, if the difference of the block with previous frames(s) is larger than a threshold, it is intracoded. This means that a P-frame may contain one or some intracoded blocks. However, the general concept of dependency chain and the drift error is still valid in these cases.

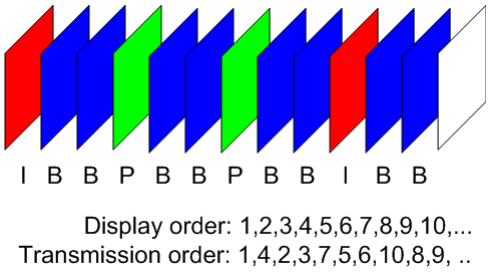

Figure 1.1: Ordering frames in GOPs.

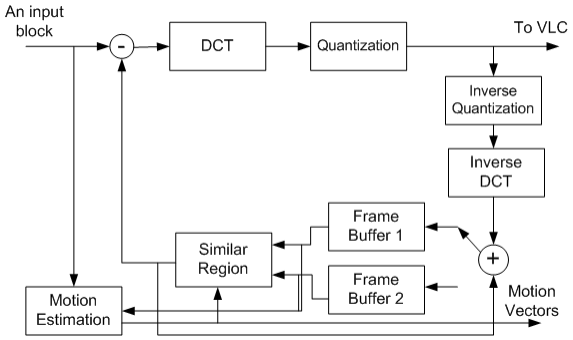

Figure 1.2: Structure of MPEG-2 encoder.

Figure 1.2 depicts the structure of MPEG-2 encoder where VLC is not shown. Each frame block is motion compensated, DCT transformed, and quantized. Since quantization causes data loss, the decoded block will differ from the original block slightly. Hence, the motion compensation is performed with decoded block which is stored in the frame buffer.

1.1.2 Adaptive Video Coding Methods

Adapting video according to the network conditions requires changing its spatial characteristic (resolution), temporal characteristic (frames per second), or visual quality (in terms of bit-per-pixel which is the number of bits of information stored per pixel) within a bounded time. The time restriction is considered as the time needed for a packet to be transmitted across a network from source to destination, or end-to-end delay. The real-time requirements of video transmission suggests the partitioning of video into at least two parts, e.g., one part to be transmitted to ensure a certain minimum quality within the worst set of timing constraints and another part to enhance the video when more processing time is available or when the network conditions improve. This type of adaptations require special video coding techniques. One of these techniques is scalable video coding (SVC). SVC methods enables fast adaptation of videos with the transmission network bandwidth or rendering capabilities of the recipient device. This adaptability however, comes with the cost of sacrificing coding efficiency to some extent. In SVC methods, the video stream is represented by a main bitstream which consists of several sub-streams. Each sub-stream contains partial information about the video, and adds to the spatial/temporal resolution, or bit-per-pixel quality of the video during reconstruction. Scaling video is fulfilled through including/droping these sub-streams. The video is reconstructed in its highest quality only by using all sub-streams. Hence, increasing the number of sub-streams provides more adaptation levels and a smoother video adaptation experience. The scalability however, is achieved

at the expense of increased complexity and reduced efficiency of video coding [24]. Although it is desirable to keep the coding efficiency of the sub-streams at the same level as non-scalable coding standards such as MPEGx or H.26x, the efficiency of SVC is lower because each sub-stream adds an overhead to the video. Therefore, a trade-off is sought between a smoother scalability (through a large number of sub-streams) and better coding efficiency.
State-of-the-art SVC methods implement the sub-streams in a multi-layered format [16], [21] where each sub-stream is considered as a layer. The first layer, named the base layer, contains the video in its lowest quality in terms of frame rate, spatial resolution, or bit-per-pixel. The remaining layers, called enhancement layers, add to the quality of the video and hence, a higher spatial/temporal resolution video, or quality video is obtained through increasing the number of layers decoded at the receiver side. The main drawback of the multi-layer video coding scheme used by SVC is the dependency between the layers so that a layer can be used in reconstructing of the video only if all previous layers are present. This implies that for downscaling a video, only the top most layer(s) should be dropped. Figure 1.3 depicts the encoding and decoding of video using SVC. As it is shown in Figure 1.3 SVC

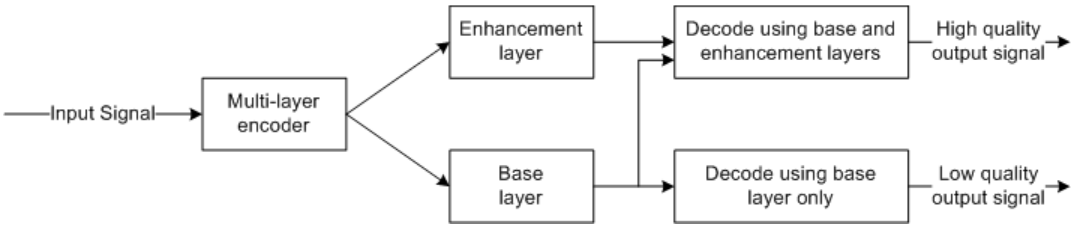

Figure 1.3: Block diagram of scalable coding of video.

assumes that adapting videos is carried out by dropping less significant layers (topmost enhancement layer(s)) and preserving more significant layers. Hence, the base layer should always be delivered.
In this section, the detailed concept of multi-layer video encoding for spatial scalability, temporal scalability, SNR scalability, and fine granularity scalability (FGS) is presented. Subsequently, some properties of the discrete wavelet transform (DWT) which are used for video scalability are presented.

Spatial Scalability

Spatial scalability is a technique to encode a video sequence into multi-layers at the same frame rate, but different spatial resolutions. The first layer (the base layer) is coded at the lowest spatial resolution by downsampling the frames. Downsampling with a rate of M is carried out by reducing high-frequency signal components with a digital low-pass filter and keeping only every M^{th} sample. Then the difference between upsampled base layer and the original frame is encoded as the enhancement layer. In case that the video is coded in more layers, this procedure is repeated on the base layer yielding a new base layer in lower resolution, and an enhancement layer [5], [32]. In spatially scalable video coding two strategies are followed for motion compensation. In the first strategy which is called single-loop encoding/decoding, the motion compensation at each layer is performed independently of the remaining layers. Hence, each layer can be decoded without waiting for the reconstruction of other layers. Figure 1.4 depicts a single-loop spatial scalability decoder. The advantage of single-loop spatial scalability is that the enhancement layer k does not need to be decoded after layer k-1 and hence, it has a lower coding and decoding complexity. In addition, since the single-loop spatial scalability decoder does not include the enhancement layer information into the prediction loop, the drift error does not occur when only the base layer is delivered. However, as the enhancement information

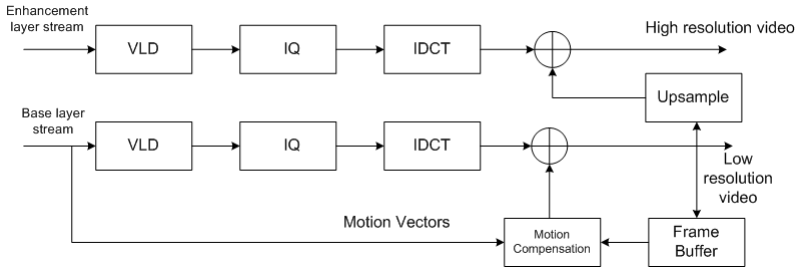

Figure 1.4: Single-loop Spatial Scalability Decoder.

of the previous frame is not utilized in the motion estimation/compensation of the current frame, this strategy causes the coding efficiency of the scalable coding to be low.
The second strategy utilizes the information from layers 1 up to k-1 for motion estimation/compensation and encoding layer k. Therefore, the reconstruction is carried out in multiple iterations and hence, the strategy is referred to as multi-loop encoding/decoding. Multi-loop encoding has a higher coding efficiency, but its complexity is higher and the encoding/decoding time is longer. The spatial scalability decoders defined in MPEG-2 and MPEG-4 use two prediction loops, one in the base layer and the other one in the enhancement layer [14],[40]. The MPEG-2 spatial scalable decoder uses a weighted combination of up-sampled reconstructed frames from the base layer and the previously reconstructed frame in the enhancement layer, while MPEG-4 spatial scalable decoder allows a bi-directional prediction using up-sampled reconstructed frame from the base layer as the backward reference and the previously reconstructed frame from the enhancement layer as the forward reference. The impacts of these two strategies on the efficiency of video coding, and the robustness of video against packet losses are discussed in "Summary on Video Adaptability" section.

Temporal Scalability

Temporal scalability is a technique to code a video sequence into two layers at the same spatial resolution, but different frame rates [6], [17]. The base layer is coded at a lower frame rate. The enhancement layer provides the missing frames to form a video with a higher frame rate. Coding efficiency of temporal scalable coding is high and very close to non-scalable coding [17]. Figure 1.5 depicts the structure of temporal scalability with two layers. Considering the two layer structure depicted in

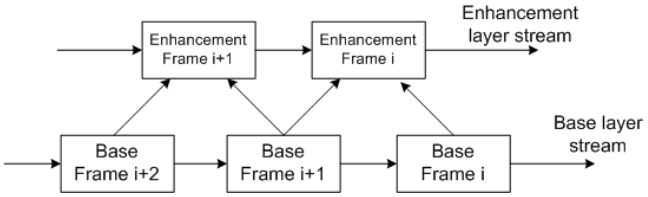

Figure 1.5: Typical Structure of a Temporal Scalability Decoder.

Figure 1.5, the enhancement frame i is the succeeder of the base frame i in the original sequence. Either enhancement frame i-1, or base frames i or i+1 can be used as a reference frame for enhancement frame i. Therefore, in compliance with the restrictions in video coding standards before H.264, only P-type predicted frames are used in the base layer. The enhancement layer predicted frames can be either P-type, or B-type referencing a P-type frame from the base layer or the enhancement layer.

Motion compensation in the based layer utilizes only the base layer information so no drift error is expected here. However, with moving some of the frames to enhancement layer(s), the distance between consecutive frames in the base layer is increased. This increase can cause a decrease in the coding efficiency.

Signal-to-Noise Ratio Scalability

Bit-per-pixel or signal-to-noise ratio (SNR) scalability is a technique to decompose a video sequence into multiple layers at the same frame rate and the same spatial resolution, so that each layer adds to the quality of the video. The decomposition can be performed in the pixel domain by putting more significant bits in the base layer and less significant bits in the enhancement layers. For instance, SNR scalability is defined as the re-quantization and subsequent re-encoding of the coding error from the previous layer in Annex 0 of H.263 [1]. H.263 defines the coding error as:

$$e(k) = f(k) - \hat{f}(k) \tag{1.1}$$

where $f(k)$ is the original frame, $\hat{f}(k)$ is the reconstructed frame, and $e(k)$ is the coding error. The coding error is then encoded using the same steps used for encoding the base layer but with a finer quantization [18]. SNR scalability is also applicable on the DCT coefficients. In this case the low frequency coefficients of the DCT are put in the base layer and the high frequency coefficients are put in the enhancement layer(s). The scalability characteristic of this scheme comes from utilizing the order of coding of the DCT coefficients. Given that the zig-zag scanning arranges the coefficients from low to high frequencies, a receiver can decide to receive the DCT coefficients partly when the bit rate of its network is low. This corresponds to eliminating high frequency content of the video for scaling down its quality and bit rate [19], [22]. An alternative method for decomposing DCT coefficients is using different quantization matrices. A coarse quantization is used to quantize the DCT coefficients at the base layer. Subsequently, a fine quantization is applied and the difference between the fine quantized and the coarse quantized DCT coefficients are used in the enhancement layer. The decoder adds the values from the base and the enhancement layers before performing inverse quantization. Figures 1.6 and 1.7 depict the SNR scalability encoder and decoder in MPEG-2 standard, respectively [23], [39].

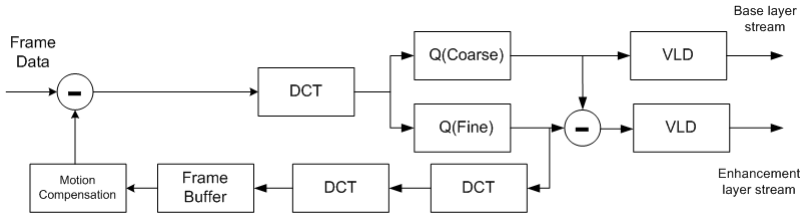

Figure 1.6: SNR Scalability Encoder in MPEG-2.

Fine Granularity Scalability

In video adaptation through dividing it into a set of smaller parts, a measure of the number of items comprising a unit (such as a frame or a macroblock) is called its *granularity* [38]. Granularity can also be considered as the precision with which rate can be controlled. Therefore, fine-grained scalability scheme permits rate to be added in small increments [15]. The scheme of gradual refining of a unit or

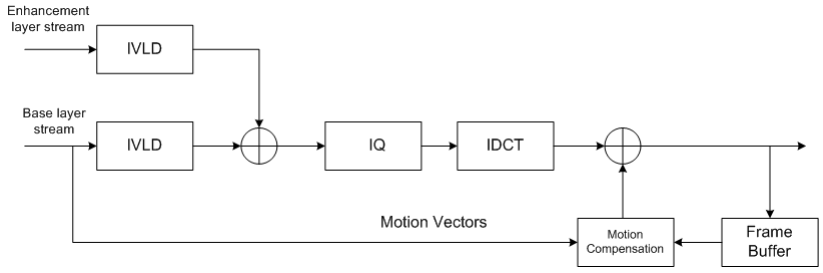

Figure 1.7: SNR Scalability Decoder in MPEG-2.

increasing the granularity of a unit is called Fine Granularity Scalability (FGS) [8], [9]. SVC defines FGS scheme for video content using a multi-layered format [16], [21]. A base layer which includes the minimal required data in order to reconstruct the video, and an enhancement layer which can be used fully or partially for enhancing the reconstructed video [7]. The enhancement layer can be truncated to match the available bandwidth. Figure 1.8 depicts a sample scenario for an FGS encoder. As depicted in Figure 1.8, the quantized DCT coefficients are inverse quantized and subtracted from

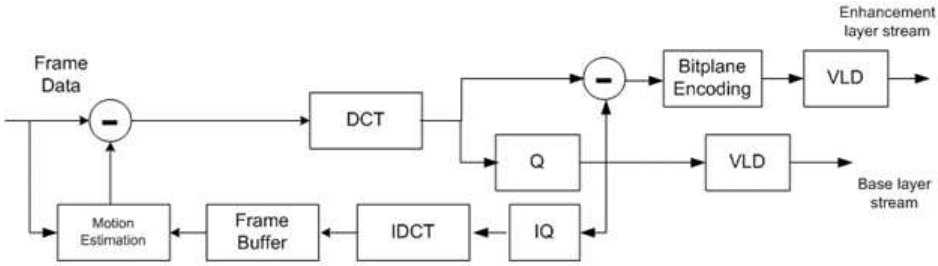

Figure 1.8: FGS applied to SNR Scalability.

the original DCT coefficients. These differences (residues) are included in the enhancement layer. The residues at the enhancement layer are re-ordered by zig-zag scanning and then they are bitplane encoded. A bitplane is defined as the bits with the same significance. Bitplanes are variable length encoded in the same way as the quantized DCT coefficients of a macroblock (Section 1.1.1). As an example lets assume the residues of the enhancement layer are 21, 19, 13, 11, and 14. Table 1.1 shows the values stored in each bitplane. Multi-layer coding method used in SVC on the other hand does

Table 1.1: Creation of the bitplanes for some sample values

Difference coefficient	21	19	13	11	14	Bit string (before VLC)
Bitplane 5 (Most significant)	1	1	0	0	0	11000
Bitplane 4	0	0	1	1	1	00111
Bitplane 3	1	0	1	0	1	10101
Bitplane 2	0	1	0	1	1	01011
Bitplane 1 (Least significant)	1	1	1	1	0	11110

not allow truncation of the data sent is each layer. Therefore, SVC layers are either included in the reconstructed video or dropped. The main drawback of FGS is its lower coding efficiency compared

to SVC which stems from the fact that motion estimation/compensation is carried out using the base layer only.
FGS is mainly used for SNR scalability of video however, the main video coding standards such as MPEG-4 and H.264 support its application to spatial and temporal scalability as well [28]. In case of spatial scalability, the frames are down-sampled to create the base layer. Then the base layer frames are up-sampled and their difference with the original frames are found. These differences are used as the enhancement layer data. In temporal FGS, the frames are grouped as the base layer and the enhancement layer frames. The enhancement layer frames are encoded by finding their difference with the base layer frames. This is similar to assuming that the reference frames in a motion compensated encoder are always placed in the base layer. The enhancement layer coding (zig-zag ordering of coefficients, bitplane, VLC) is the same as SNR FGS. Temporal FGS provides the possibility of dropping the frames of the enhancement layer if the available bandwidth is low (similar to temporal SVC), but also since the enhancement layer frames are bitplane encoded, it is possible to eliminate some of the bitplanes from the enhancement layer instead of eliminating the whole frames. FGS methods have the following characteristics:

- Since the enhancement layer can be truncated at any point, they provide a continuous scalability.
- Bitplane encoding used by FGS methods is more efficient that run-length encoding used in standard video coding methods [20].
- In order to avoid the drift error, the enhancement layer in FGS methods is not used for motion estimation/compensation. This feature reduces the coding efficiency of FGS methods.

Scalability using DWT

The DWT has also been used for providing scalability characteristic for a video stream since it allows localization in both the spacial and frequency domains [4], [31], [33]. Signals carry information which can be audio, image, etc. which changes with time. Besides, information contained in a signal is comprised of different frequencies, with each frequency having a (probably) different energy level. The Fourier transform has been used for many years to transform the representation of signals from time domain into frequency domain and vice-versa. Fourier transform uses sine and cosine functions as its basis functions for the transformation. However, sine and cosine functions are not limited in time, extending to infinity. The impact of this property is that the energy amount at each frequency can only be found for the whole period of the signal. On the other hand if the frequency measurement is carried out at a limited time, we can only find out the amount of energy at a given frequency in that period. This property is referred to as the frequency/time resolution of the Fourier transform. The wavelet transform uses basis functions which are limited in time (hence the name wavelet). Different basis functions have been used for the wavelet transform. The basis function ($\psi(x)$) can be translated and dilated as shown in Equation 1.2.

$$\psi(\frac{x-\tau}{s}), (\tau, s) \in R^+ \times R \tag{1.2}$$

The wavelet transform provides time and frequency resolution of the signal at the same time by dilating and translating the basis function. In addition, functions with spikes or discontinuities require fewer wavelets to represent compare to sine and cosine functions. This property makes wavelets more suitable for data compression. The continuous wavelet transform of a signal f(t) is given in Equation 1.3.

$$\Psi_f(\tau, s) = \frac{1}{\sqrt{s}} \int f(t)\psi^*(\frac{t-\tau}{s})dt \tag{1.3}$$

where ψ^* is the complex conjugate of the basis wavelet function. The one-dimensional discrete form of the wavelet transform is given in Equation 1.4.

$$\Psi_f(\tau, s) = \frac{1}{M}\Sigma_n f(n)\psi(\tau, s) \tag{1.4}$$

The two-dimensional DWT is used for image and video compression. To use the wavelet transform for image/video compression the wavelet coefficients are quantized and binary encoded. Generally (and depending on the content of the transformed data) many of the quantized DWT coefficients are zero. In order to attain a better compression rate, non-zero coefficients and their locations are stored. This is accomplished using various techniques such as the zig-zag scanning, or spatial oriented trees such as Embedded Zero-tree Wavelet (EZW) [29], [30]. During the storing or transmission of the transformed video data as a bitstream, these techniques ensure that more significant information is located at the initial bits. Therefore, truncating the bitstream of the DWT coefficients preserves more significant information and minimizes the video quality distortion. Truncating the DWT bitstream as mentioned above can provide SNR scalability where fewer bits in the bitstream implies more video quality degradation.
In the DWT-based compression, typically a visual data unit such as an image or a frame, is decomposed in a hierarchy of frequency sub-bands by filtering along one spatial dimension at a time to effectively obtain four frequency bands as shown in Figure 1.9. Here we have utilized "Low" and

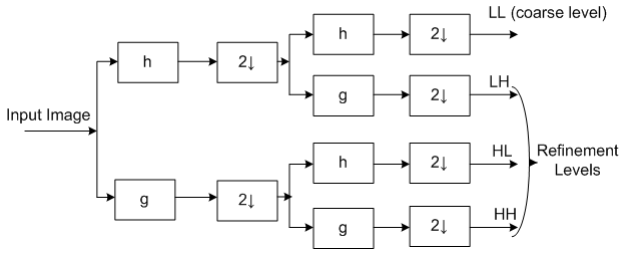

Figure 1.9: Decomposition into Frequency Sub-bands in Wavelet Transform.

"High" to indicate application of a low-pass and a high-pass filter, respectively. The filters ***h*** and ***g*** shown in Figure 1.9 decompose the image into independent frequency spectra of different bandwidths or resolutions [31], [35], producing different levels of detail and are commonly referred to as the analysis filters. The analysis filters ***h*** and ***g*** are low- and high-pass filters, respectively. The two dimensional DWT is performed by applying the low-pass and high-pass filters in horizontal and vertical directions. Here L and H stand for low and high frequency bands respectively. Hence, Low-High indicates applying low-pass and then high-pass filters to the data. The lowest sub-band, commonly referred to as Low-Low (LL), represents the information at coarser scales and is further decomposed and sub sampled to form another set of four sub-bands. This process can be continued until the intended number of decomposition levels is reached. Applying inverse of the DWT to reconstruct the original data is carried out through synthesis filters. The reconstruction is accomplished by upsampling the lower resolution images and passing them through synthesis filters. The various sub-band signals are recombined to reconstruct the original signal.Various types of analysis/synthesis filters have been proposed in image and video compression. The most basic analysis/synthesis filters are Haar filters proposed by Alfred Haar in 1910 [13]. Haar analysis filters decomposes a signal *x(n)* into two signals *c(n)* and *d(n)*. *c(n)* is the low-pass filtered signal and is obtained by finding the average of every value pair in *x(n)*. *d(n)* or difference signal, is the high-pass filtered signal and is obtained by finding the

difference of each pair of values. Equation 1.5 defines the Haar analysis filters.

$$\begin{aligned} c(n) &= 0.5 \times (2n) + 0.5 \times (2n+1) \\ d(n) &= 0.5 \times (2n) - 0.5 \times (2n+1) \end{aligned} \tag{1.5}$$

Haar synthesis filters are given in Equation 1.6.

$$\begin{aligned} y(2n) &= c(n) + d(n) \\ y(2n+1) &= c(n) - d(n) \end{aligned} \tag{1.6}$$

The Haar transform in approximately piecewise constant signals where *d(n)* is zero, is very successful. For more general signals which are not necessarily piecewise-constant more complex analysis/synthesis filters have been proposed. Among the proposed filters the most famous ones are Daubechies filters. Equations 1.7 and 1.8 define the Daubechies analysis and the synthesis filters, respectively.

$$\begin{aligned} c(n) &= h_0 \times (2n) + h_1 \times (2n+1) + h_2 \times (2n+2) + h_3 \times (2n+3) \\ d(n) &= h_3 \times (2n) - h_2 \times (2n+1) + h_1 \times (2n+2) - h_0 \times (2n+3) \end{aligned} \tag{1.7}$$

$$\begin{aligned} y(2n) &= h_0 c(n) + h_2 c(n-1) + h_3 d(n) + h_1 d(n-1) \\ y(2n+1) &= h_1 c(n) + h_3 c(n-1) - h_2 d(n) - h_0 d(n-1) \end{aligned} \tag{1.8}$$

where the multipliers are:

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \quad h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

Daubechies filters have the property that when the original signal *x(n)* is linear (such as *x(n)= an+b*), the output signal *d(n)* will be identically zero. A linear signal indicate that with the increasing value of n, there is a gradual increase in the value of the signal as well. This kind of gradual increase is a frequent case in images and videos corresponding to the effect of illumination on uniform areas, which results in a piecewise linear signal. The DWT of the signal provides a sparse matrix in these cases. This property is very important in data compression applications as the coefficients in high frequency bands become mostly zero.

Summary on Video Adaptability

Video adaptability is achieved by decomposing it into multiple sub-streams where each sub-stream improves the quality of the video in terms of its resolution (spatial), frame rate (temporal), or number of bits allocated to each pixel (SNR). The first sub-stream which presents the video in its lowest quality is called the base layer. It is necessary to add the remaining sub-streams in a pre-defined order. The general idea in video adaptation using SVC methods is presented in Section 1.1.2. There are however, three main drawbacks with SVC methods:

- SVC methods down-scale the video when it needs to be adapted. As a result of this adaptation, the frames reconstructed by the receiver become different from the frames used by the encoder. This difference inversely affects the quality of the succeeding frame(s) which use it as their reference frame. This quality degradation accumulates to the drift error.

- Encoding video in a multi-layer format reduces the coding efficiency.
- The frames should be decoded in the order of the layers in SVC. Hence, enhancement layers are decoded after the base layer, and in their correct order. Therefore, whenever the base layer is damaged or lost, the remaining layers become useless. This feature makes SVC methods very sensitive to data loss, although SVC techniques are not intended for unreliable networks.

As a consequence, improvements in SVC methods are necessary to make them robust against down-scaling reference frames, and packet losses, while improving the coding efficiency.

1.1.3 Error-Robust Video Coding

Packet loss or corruption during the video transmission can severely degrade the video quality. The impact of packet loss on multi-layer SVC coded videos is even worse if the affected packet contains the base layer data. This is due to the dependency of bit-streams (layers) in SVC which makes decoding upper layer bit-streams impossible when the lower layers are not present. When a video is down-scaled due to the network conditions, we decide to keep the lower layers and drop the upper layers. However, in case of communication errors such as packet losses, we cannot select the delivered packets. The error robust video coding methods either use error correction codes to handle the data loss, or decompose the video into bit-streams and transmit each bit-stream independently hoping that some of them will be delivered intact. The main difference between SVC methods and error robust methods in decomposing a video is that SVC bit-streams are dependent on each other but bit-streams created by error robust methods are independent and can be reconstructed without relying on the information from other bit-streams.
One of the error robust video coding methods is Multiple Description Coding (MDC) which divide the video data into some bit-streams called descriptions. Descriptions are generally transmitted separately over different network channels [37]. Typically, descriptions have the same importance and data rates, even though this is not a necessary requirement. Each description can be decoded independently from other descriptions and the loss of some of these descriptions does not affect the decoding of the rest [34], however, the accuracy of the decoded video depends on the number of descriptions received [10]. When one or some descriptions of a video are lost, the decoder estimates them by utilizing the spatial or temporal correlation in data. The correlation of the video data implies that a statistical redundancy is present which can be exploited for more effective coding [11]. The presence of this correlation is the source of reduced efficiency in MDC compared to single stream video coding, although it is necessary for estimating the missing data through interpolation.
The MDC can be applied to the transform coefficients of the video as well. When the decomposition of a video is performed on the transform coefficient estimating the missing data from the received descriptions becomes difficult as the coefficients are not correlated. An attempt to create a correlation between coefficients was made in [36]. In their work, they found 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:

$$\gamma_1 = 2^{-1/2}(S_1 + S_2)$$
$$\gamma_2 = 2^{-1/2}(S_1 - S_2)$$

with correlation coefficient $\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ known to the receiver end. Besides, an essential difference between decomposing video signals into subsets in spatial and transform domain is that the data items in spatial domain are of the same importance but the transform coefficients correspond to information

in different frequency range and hence are of different importance. Meanwhile, some parts of the transformed data are of vital importance. The low frequency or DC coefficient in DCT transform is an example for such a vital coefficient. In [2] for each macro block transformed by DWT, two streams at low and high rates are created. The low rate streams are created by truncating the symbols generated after zig-zag scanning the coefficients and putting them in an EZW tree. The method consists of two descriptions where each description contains high rate streams of some of the blocks and low rate streams of the rest. If both descriptions are received, high rate streams are used for reconstructing the video. However, if only one description is received, some blocks are reconstructed using their low rate streams. A balanced and an unbalanced form of decomposing the streams to create descriptions have been proposed by the authors as depicted in Figures 1.10 and 1.11. The most important drawback

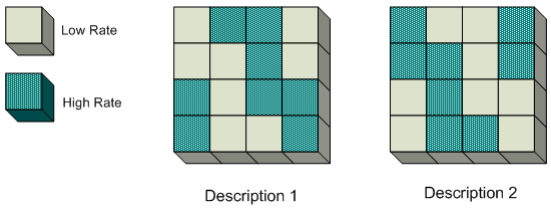

Figure 1.10: Balanced descriptions using low and high rate DWT blocks.

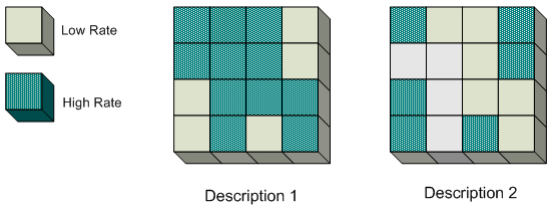

Figure 1.11: Unbalanced descriptions using low and high rate DWT blocks.

of error-robust video coding methods is the redundancy they add to the data in order to estimate the missing data. This redundancy is present even if the communication network has a very low error rate and most of the data packets are reliably delivered. The existence of this redundancy is the main difference between the error robust methods and the SVC methods. For instance, the method depicted in Figures 1.10 and 1.11 which transmits the block data in low quality in one description, and in high quality in the second description, corresponds to sending base and enhancement layers data in one description, and only the base layer in the second description.

1.2 Video Adaptation Challenges

Fulfilling the requirement for video transmission over the Internet as explained in Sections 1.1.2 and 1.1.3, necessitates considering some issues and dealing with a couple of challenges. In this section we first describe the issues which should be addressed in video coding and then we pose the challenges stemming from the requirements that motivated this research. We argued that video transmission requires adaptation to the fluctuations in the network bandwidth or receiver end rendering capabilities. In addition, we also argued that the adaptation necessitates the development of algorithms capable of fast processing, being robust against transmission errors, and minimizing overheads. These

requirements pose challenges such as:

- Video adaptation involves partial delivery of the video data. Motion compensated encoding of the frames however, requires the reference frames to be present for error-free reconstruction. Adapting video by down-scaling frames changes the reference frames. If the motion compensation utilizes the high-quality reference frames, using the low-quality reference frames at the receiver side will result in reconstruction error which accumulates to cause the drift error. Reducing or eliminating the drift error reduces the coding efficiency. The challenges in this issue are:
  - Optimizing the coding efficiency with respect to the expected bandwidth changes or packet loss rates dynamically, and through adapting video parameters such as GOP length,
  - The bit rate of a video depends on its content. The impact of the content of a frame can be determined from the frequency content distribution after the DCT transform. Coarse quantization of a frame with low frequency content creates less distortion than a frame with high frequency content. Therefore, the impact of bandwidth fluctuations on videos varies with their content. A challenge in video coding is optimizing video coding parameters with the frame content,
  - The impact of changes in a frame on the succeeding frames depends on the position of the frame in the GOP. Any change at a frame located at the beginning of a GOP affects the video quality more than a frame which is located close to the end of the GOP. Adjusting video coding and adapting parameters in accordance with the frame position is also a challenge.
- While packet loss and partial data delivery in scaled-down videos have similar impacts on the quality of the reconstructed videos, the inhibitions to prevent them in video encoders are quite different. The expected characteristic from a video stream is its adaptability with, and error resilience against the network conditions. Combining these improvisations in a single encoding technique is a challenge in video coding.
- Video coding, handling error and packet losses, and adapting videos during transmission require time-consuming processings. On the other hand, many video transmissions have time-restrictions for video delivery. To avoid the processing delay, video is coded in a way that minimizes the on-the-fly processing. However, this comes with the extra cost of lower efficiency and overheads. Hence, the encoder should provide the possibility of real-time video adapting and/or error handling for video streaming applications.

1.3 Research Directions

Video adaptation poses many challenges as summarized in Section 1.2. Reducing the video quality degradation is the main research direction of this research. As it is mentioned in Sections 1.1.2 and 1.1.3, the quality degradation occurs as a consequence of video adaptation if the reference frames used at the encoder and the decoder become different, and/or due to video data loss. In this thesis we address both issues jointly. Hence, the challenges addressed are:

1. Adapting video with network conditions and/or receiver device rendering capabilities while minimizing the quality degradation and optimizing the coding efficiency,

2. Developing error-robust video coding techniques encompassing minimum overhead.

Adapting video with network conditions and/or receiver device rendering capabilities while minimizing the quality degradation and optimizing the coding efficiency

An adaptive coding method implies that the coding method should be able to encode the video in a way that its data rate, or spatial and/or temporal characteristic can be altered when needed. Minimizing the quality deterioration of this adaptation can negatively affect the coding efficiency. Developing optimized methods for adaptive video coding which minimizes the quality deterioration requires analyzing the impact of each stage of the video coder and its parameters. Developing these methods is the first challenge addressed in this dissertation.

Developing error-robust video coding techniques encompassing minimum overhead

Packet loss is an inevitable characteristic of the current unreliable networks. This packet loss imposes video reconstruction from partially available data. In order to minimize the quality loss due to partial data availability, the reconstruction should not depend on any specific part of the video, but on the number of delivered parts. This requirement prohibits the utilization of the correlation between data parts for better compression. The challenge addressed here is developing encoding methods for error-resilient video coding with improved coding efficiency, and capable of avoiding large overheads.

In this research, we address both challenges together and propose methods which provide video adaptability and error resilience at the same while optimizing the coding efficiency. The third challenge of minimizing processing time which has been mentioned in Section 1.2 has not been directly addressed in this dissertation.

1.4 Main Contributions

The main contributions of this work regarding the challenge of *adapting video with network conditions and/or receiver device rendering capabilities while minimizing the quality degradation and optimizing the coding efficiency* are:

- A hierarchical structure for reducing the drift error when the transmission is over an unreliable network. The structure decomposed the frames of a GOP into low-pass and high-pass frames by pairwise temporal filtering the frames. This hierarchical structure provides the possibility of reducing the number of frames depending on each other in a dependency chain while preserving the length of the group of pictures. Since the drift error is the result of changes in reference frames, the size of the enhancement layer is selected based on the location of the frame in the frame dependency chain. A method for optimizing the quantization step size (and hence the enhancement layer size) for each frame, and GOP length is proposed which minimizes the total video degradation due to the drift error while optimizing the efficiency in terms of bit per pixel rate. The average improvement in term of PSNR value when both layers are delivered is 4.78(dB) while it is 3.70(dB) when only the base layer is delivered. The improvements in the robustness of the video in presence of burst errors are 3.52(dB) for the base layer only delivered videos where the average PSNR values are 22.32(dB) and 18.8(dB) for the proposed method and the sequential coding respectively, and 4.50(dB) when the base and the enhancement layers are delivered with the average PSNR values are 24.01(dB) and 19.51(dB) for the proposed method and the sequential coding respectively. *(Chapter 2)*

- In privacy protected video streaming the reconstructed video suffers from the drift error when the scrambled area is (partly) used as the reference area for (a) block(s) of succeeding frames. The drift error happens when the receiver is not authorized to access to the scrambled area and hence the motion compensated succeeding blocks which refer to it become invalid. Avoiding the utilization of the privacy protected area puts the burden of tracking their locations during motion estimation and increases the processing time. An amendment to the motion compensation algorithm has been proposed which eliminates the drift error when parts of the video is scrambled for privacy protection. The quality improvement in PSNR is 0.6dB while the processing time for scrambling and unscrambling the video is negligible. *(Chapter 2)*

- A scalable video coding algorithm has been proposed for SNR decomposition of the video. The proposed method improves the robustness of the video against transmission errors by creating descriptions after its decomposition. In order to correlate the decomposed video streams, a transformation has been defined. The significance of this method is that the reconstruction of video in a lower quality is possible regardless of which description is lost. This means that despite traditional multi-layer scalable coding techniques for SNR decomposition, the proposed algorithm creates descriptions which can be used independently for reconstructing the video and carry information of (almost) the same significance. Our proposed method imposes an average of about 41.2% redundancy to the video while the similar state-of-the-art method of Multiple Description Transform Method (MDTM) adds a redundancy of 45% in the same PSNR value. *(Chapter 3)*

- An improvement has been proposed for increasing the coding efficiency of temporal decomposing of video frames into two sub-streams. In the temporal decomposition of videos into multiple sub-streams, a frame and its reference may occur in different sub-streams. This situation can result in the drift error and quality degradation when packet losses happen in the network. Using the frames from the same sub-stream as reference frame can reduce coding efficiency. We propose an improvement which groups so that a frame and its most similar frame are placed in the same sub-stream. The proposed method improves the bit-per-pixel rates of the smaple videos such as "Foreman" to 1.181 bpp from 1.2212 bpp, and "Stefan" to 1.237 bpp from 1.291 bpp when no optimaztion is performed. *(Chapter 3)*

The main contributions regarding the challenge of *developing error-robust video coding techniques encompassing minimum overhead* are:

- An algorithm for decomposing video frames in spatial domain has been proposed. The proposed method creates four streams as four sub-sets of pixel values where each sub-set contains a common base layer and an enhancement layer. The common base layer is defined as the average of the four sub-sets, and the enhancement layer is found through the difference of each sub-set from the common base layer. The proposed method transmits each stream as a description where the correlation between the streams is utilized for estimating lost data in case of packet loss errors. In addition, the reconstructed video quality in case of one or some description losses has been improved by introducing a new interpolation method which minimizes the reconstruction error. An optimization algorithm has also been proposed for minimizing artifacts when the frame contains thin horizontal/vertical areas. The proposed method can reconstruct video when one description is lost with a PSNR value of upto 36.34dB which makes the data loss almost unnoticeable. *(Chapter 3)*

- A method to estimate the missing data due to packet losses has been proposed which utilizes the self-similarity between the DWT coefficients at different decomposition levels. The coefficients are transmitted on different streams using multiple descriptions. A new structure for organizing data in descriptions is proposed which combines lower coefficients of one sub-band with high coefficients of another sub-band in a description. By doing so, the proposed method can estimate the missing data using the self-similarity property of the coefficients in each sub-band in case of a description loss. The experimental results indicate that when two descriptions are delivered, the average PSNR values with and without using self-similarity index for reconstruction are 35.69dB and 33.55dB respectively. The average PSNR values when only one description is delivered are 34.38dB and 27.12dB for reconstruction with and without using self-similarity index respectively. *(Chapter 4)*

- An algorithm is proposed for decomposing the 3D wavelet transformed video data into multiple descriptions. The proposed method minimizes the distortion by selecting the optimum truncation of the DWT coefficients in each description. This optimization considers the bandwidth fluctuations of the communication channels, and the frequency content at each sub-band. For instance, if a frame contains more frequency content in the horizontal frequency sub-band, the corresponding description will have a higher bit-rate. The assignment of a description to a communication channel is performed by considering the dynamic changes in the bit-rate of the descriptions and the bandwidth of the channels. The algorithm utilized for this purpose, minimized the overall distortion of the video (R-D optimization). The bit-rate performance of the peroposed method using ”Foreman” sample sequence are 31.5dB in 100Kbits, 33.4dB in 200Kbits, and 34.0dB in 300 Kbits while state-of-the-art method of F-MDC displays 31.6dB, 32.8dB, and 33.5dB for the same bit-rates. (Chapter 4)

1.5 Thesis Organization

This dissertation has the following organization:
Chapter 2 includes the papers which concentrate on minimizing the drift error. The minimization is considered as an improvement to the SVC methods. In Chapter 3 improvements of MDC methods for error-robust video coding in spatial, temporal, and SNR decomposition of videos are presented. These techniques are combined with SVC methods for video adaptability. In Chapter 4 MDC methods are applied in DWT domain. The presented methods are combined with SVC as well.

Chapter 2

Minimizing Drift Error in Scalable Video Coding

High quality videos include a huge amount of data. Storing and communicating these large videos require efficient encoding with high compression rates which in turn necessitates reducing or eliminating redundancies in videos. One of the redundancies in videos is temporal redundancy which stems from the similarity between the consecutive frames (Section 1.1.1). In order to eliminate the temporal redundancy, only the differences between the current frame and its preceding frame(s) are encoded. However, this encoding makes video frames dependent on their preceding frames (reference frames) and as a result, any changes in the reference frames cause deviation from the original frame during the reconstruction. This deviation results in video frames quality degradation which accumulates to the drift error (Section 1.1.1. The reference frames of a video may change for different reasons. The most important reasons are as follow:

- Intentional changes by an end user or an application: For instance, an application may manipulate parts of frames to protect the identity of the people appearing in the scene, or to hide sensitive information such as car license plates. These manipulations are carried out through scrambling data using special algorithms. The assumption is that if the receiver is authorized to view the protected information, (s)he should be able to unscramble them.
- A frame may be lost due to a packet loss during transmission. This case happens frequently in wireless networks, especially when the receiver device is on a moving platform, such as a cell-phone being used in a vehicle.
- Frame changes due to down-scaling the video when the available network resources are insufficient. This can happen in all types of public networks where the resources are shared and the traffic load varies unpredictably.

In this chapter we first address the drift error due to the intentional changes in the reference frames. The application we considered here is privacy protection in videos through scrambling. We propose a method for scrambling the protected parts of the video using a private key. The algorithm provides the possibility of unscrambling the video whenever the private key is known. The proposed method completely eliminates the drift error even when the private key in not known.
In this chapter we also address video quality degradation due to the drift error when the video is scaled-down. We observed that the amount of degradation due to the drift error depends on two factors as explained below:

1. Any change in a frame negatively affects the succeeding frames using it as their reference frame. Since the quality degradations in the frames accumulate, the number of succeeding frames following the modified reference frame before a new GOP starts is an important factor in the overall quality degradation of the video.

2. Since the frames encoded in predicted or bidirectional modes (P- or B-frames) rely on the information of their reference frame, the amount of changes in the reference frame is an affective factor is the quality degradation. In SVC we drop the enhancement layer for down-scaling the video. Therefore, the relative sizes of the base and the enhancement layers are important.

Regarding the first observation, in our proposed method we tried to reduce the number of frames which depend on each other in a GOP while preserving the coding efficiency. We proposed utilizing a hierarchical structure for organizing the frames in a GOP to reduce the number of frames depending on each other in a GOP. Regarding the second observation, we proposed an adaptive SVC method which decides the size of the base layer with the position of the frame in a frame dependency chain. The details of the proposed methods and their experimental evaluations have been presented in the following research articles:

- R. Choupani, S. Wong, M.R. Tolun, Drift-free Video Coding for Privacy Protected Video Scrambling, 10th International Conference on Information, Communications and Signal Processing (ICICS 2015), Singapore, pp. 66-72

- R. Choupani, S. Wong, M.R. Tolun, Hierarchical SNR Scalable Video Coding with Adaptive Quantization for Reduced Drift Error, 10th International Conference on Computer Vision Theory and Applications (VISAPP 2015), Berlin Germany, pp. 117-123.

Drift-free Video Coding for Privacy Protected Video Scrambling

Roya Choupani
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, the Netherlands
e-mail: r.choupani@tudelft.nl

Stephan Wong
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, the Netherlands
e-mail: J.S.S.M.Wong@tudelft.nl

Mehmet Tolun
Department of Electrical Engineering
Aksaray University
Aksaray, Turkey
e-mail: mehmet.tolun@aksaray.edu.tr

***Abstract*—With video surveillance systems becoming ubiquitous nowadays, protecting people's privacy raises an increasingly serious concern. Video streaming with privacy protection requires modifying parts of the video content. This modification should provide the possibility of unprotected access to the video if the user is authenticated through a private key. However, any modification in the content of a video can result in a drift error and deteriorate the quality of the reconstructed video. In addition, it is required that the privacy protection does not adversely affect the computational complexity and the coding efficiency in terms of bitrate. In this work, we propose a drift-free method for scrambling the privacy protected regions of the frames while preserving the coding efficiency. Our proposed method provides the possibility of utilizing private keys for restricting unauthorized access to the private contents of the video with a small increase in the computational complexity of the encoder and decoder. The experimental results indicate that our proposed drift-free method can achieve a higher coding efficiency of 0.6 dB on average compared to similar methods.**

***Index Terms*—Privacy Protection in Video, Video Scrambling, Restricted Video Coding.**

I. Introduction

With the fast progress in multimedia technologies, video surveillance has obtained a widespread use. This excessive use of multimedia in general and video in particular, has given rise to many concerns about the privacy of individuals [1][2]. Among the solutions provided for the privacy protection in video is scrambling parts of the video frames corresponding to the private information such as the identity of the people in the scene [3][4][5][6]. The scrambling is performed by using a private key which can also be used for unscrambling the video. This means that the video should be decodable in both scrambled and unscrambled forms at the receiver side. Scrambling is performed by inverting the sign of AC coefficients [7], dividing the coefficients by a scrambling matrix generated at the server side using a private key [3], or by means of a seed number generating a random sequence which is used in performing XOR-operation on the coefficients [8]. The state-of-the-art video coding standards such as H.264, make use of Motion Compensated Temporal Filtering (MCTF) methods for eliminating temporal redundancy [9][10]. Hence, the inter-coded frame blocks are motion compensated before applying the Discrete Cosine Transform (DCT). If the area to be scrambled is used (even partially) as the reference for a block from a succeeding frame(s), the reconstructed video at the receiver will suffer from a frame quality degradation if the decoder does not have the necessary key to unscramble the reference frame. This degradation is accumulated in the subsequent frames and results in the drift error until an intra-coded frame (I frame) is reached. The drift error which is the result of mismatches between the reference frames at the encoder and the decoder, is one of the most important challenges in the privacy protection of video through scrambling. This paper addresses the drift error problem in privacy protected video coding and proposes a new solution for the problem. The paper is organized as follows: In Section II we review the related previous work and the state-of-the-art in privacy protection in videos. Section III introduces our proposed method, followed by the experimental results in Section IV. Finally in Section V we draw our conclusions and indicate possible directions of improvement.

II. Related Work

Apart from the methods which are based on the intra-coding of video frames without utilizing MCTF [11], the drift error caused by scrambling Region of Interest (ROI) areas is addressed in two different ways. The first group of methods is based on letting the drift error to happen either assuming that the scrambled frame range contains the entire GOP, or its effect is negligible.

In [12] the authors protect video content privacy by scrambling the entire frame by means of a context-aware middleware. The scrambling is carried out by by pseudo-randomly flipping the ac/dc coefficients of macro-blocks (MBs) in intra-coded frames only. Hence, when the key to unscramble the intra-coded frame is not available, the entire GOP becomes scrambled. In fact the reconstruction of the inter-coded frames of a scrambled GOP suffers from the drift error however, since these frames are not intended to be clearly decoded, the drift error becomes irrelevant here. Dufaux et al. [7] change the AC coefficients using a pseudo-random sequence generated by a seed number. They toggle the coefficients as:

$$qAC_{coef} = \begin{cases} -qAC_{coef} & \text{if RandomBit=1} \\ +qAC_{coef} & \text{otherwise} \end{cases}$$

978-1-4673-7218-3/15/$31.00 ©2015 IEEE

without considering the effect of the drift error.
The second group of methods are based on avoiding the drift error. In [4] the authors propose a restricted video coding scheme to avoid the drift error. The main idea in their proposed method is restricting the search area during motion estimation not to include the scrambled areas. In this way the scrambled areas are not used as reference for any MB and hence, the drift error is avoided when the user is not authorized to see the protected areas. They propose three different methods, Mode Restricted Intra Prediction (MRIP), Search Window Restricted Motion Estimation (SWRME), and Boundary Strength Restricted Deblocking Filtering (BSRDF), to handle the intra-prediction in I frames, and inter-frame prediction modes to avoid using scrambled areas for motion compensation. Tong et al. [6] improve the MRIP method proposed by Dai et al. [4] assuming that the most probable blocks to be utilized in intra-prediction mode are left and top blocks of the current block. Hence, if the left or top block of the current block is in the scrambled area, they forbid using intra 4×4 prediction mode for it. Wang et al. [13] use a similar method to scramble the video however, they reduce the bitrate overhead by choosing the prediction modes of the 4×4 blocks around the boundary of the privacy area instead of forbidding intra-prediction. In [14] a bit mask is proposed to indicate the location of blocks from the privacy area. Our proposed method is a drift-free scheme however, the encoder does not need to restrict the search areas during motion estimation. In this way, the complexity of the encoder is reduced while the coding efficiency is slightly improved. To justify this improvement, we can consider the case when the most similar area to a block partly overlaps with a scrambled area. However, if overlapping is not allowed during motion estimation, the residual values after motion compensation can be larger.

III. Proposed Method

Our proposed method utilizes the linearity characteristic of the cosine transform. The main issue addressed in the proposed method is eliminating the drift error in protecting the privacy of the people appearing in the video, while preserving the video coding efficiency. The methods proposed in the literature perform motion compensation in the original video, and then scramble the quantized and DCT-transformed residues. We propose performing motion compensation using scrambled frames. The block diagram of our encoder is depicted in Figure 1. The scrambling is carried out by post-multiplication of the 8×8 matrix of quantized coefficients by matrix $\boldsymbol{S}$ which is defined in Equation 1.

$$S = sign \begin{pmatrix} -||key||^8 + 0.5 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & -||key||^1 + 0.5 \end{pmatrix} \quad (1)$$

where ***key*** is a positive integer of eight digits representing the secret key being used for scrambling and unscrambling the privacy protected parts of the frames, and $||.||^r$ refers to normalizing the r^{th} digit of an integer number to a decimal number between 0 and 1 as shown in Equation 2.

$$\begin{aligned} For \quad & r = 1, .., 8 \\ & y = floor((x \pmod{10^r}) \times 10) \\ & ||x|| = \tfrac{y}{10} \end{aligned} \quad (2)$$

In fact the scrambling matrix is used for toggling the signs of the quantized DCT coefficients at the main diagonal position. The most significant digit of the ***key*** value toggles the sign of the DC value in the DCT coefficients. Changing the DC value scrambles the area completely however, the visual quality of the image degrades significantly. If a simple blurring of the protected area is sufficient, this digit should be less than 5. Given a block of motion-compensated coefficients, $\boldsymbol{B}$, the encoding process is as given in Equation 3.

$$EB = VLC(S \times Q(DCT(B))) \quad (3)$$

where *EB* is the encoded block, and Q and DCT refer to the quantization and the discrete cosine transform, respectively. Our assumption is that the locations of the blocks belonging to the privacy protected area are provided as shown in Figure 1. The decoding process can be carried out in two different cases as below:

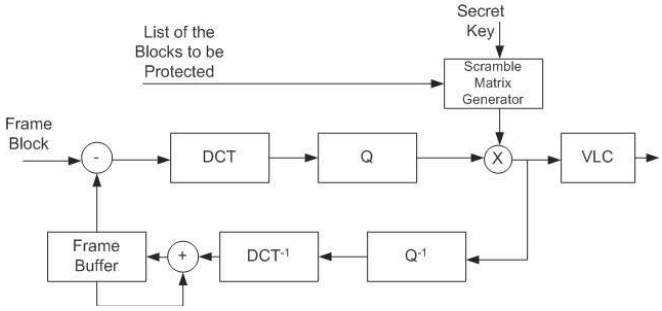

Figure 1. Frame coding block-diagram of the proposed method.

- The user does not provide the decoding key. In this case the normal steps of decoding blocks are followed. Since the scrambled frames have been used as (part of) the reference areas, in the decoded frames the privacy protected areas appear as scrambled.
- The user provides the decoding key which is used for creating matrix S. This matrix is used for unscrambling the protected areas as shown in Equation 4.

$$DB = Inv(DCT^{-1}(S)) \times DCT^{-1}(Q^{-1}(VLC^{-1}(EB))) \quad (4)$$

where DCT^{-1}, Q^{-1}, and VLC^{-1} are inverse DCT, inverse quantization, and inverse variable length coding, respectively. *DB* refer to the decoded blocks, and *Inv* is used to show the inverse of a matrix. An analytical proof of Equation 4 where we assume quantization is an invertible operation, is given in Equation 5. Replacing *EB* with its definition from Equation 3,

we have:

$$\begin{aligned} DB =& Inv(DCT^{-1}(S)) \times DCT^{-1}(Q^{-1}(VLC^{-1}(VLC(\\ & S \times Q(DCT(EB)))))) \\ DB =& Inv(DCT^{-1}(S)) \times DCT^{-1}(Q^{-1}(S \times Q(DCT(EB)))) \\ DB =& Inv(DCT^{-1}(S)) \times DCT^{-1}(S \times DCT(EB)) \\ DB =& Inv(DCT^{-1}(S)) \times DCT^{-1}(S) \times EB) \\ DB =& EB \end{aligned} \quad (5)$$

Despite the methods proposed in the literature which provide privacy protection through direct manipulation of the quantized coefficients, our proposed method uses matrix multiplication for modifying the coefficients. This lets us utilize the linear property of the DCT to design the decoder so that the un-scrambling is done as the last step and after applying inverse DCT to the coefficients. Hence, the proposed method has a minimal impact on the encoding and decoding procedures. For instance, the proposed method does not need to consider if a non-protected block is motion-compensated (partly) using a privacy protected block or not. The decoding process is depicted in Figure 2. As shown in Figure 2, in both cases

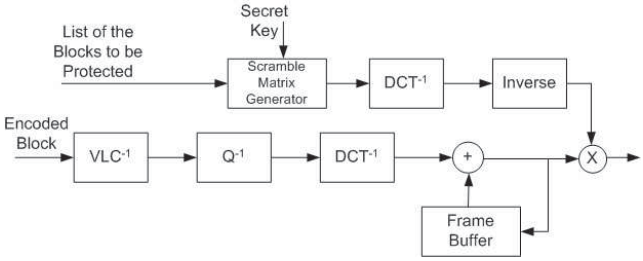

Figure 2. Frame coding block-diagram of the proposed method.

whether the decoding key is provided or not, the reference frames are unscrambled frames. Considering that the motion compensation in frame encoding is done using scrambled frames, no drift error is created by the proposed method. Moreover, there is no need to restrict the search area of during motion estimation process hence, the proposed method adds only minor modifications to the standard video coding methods. This means that the proposed method can be used with all state-of-the-are video coding standards.

IV. Experimental Results

In our proposed method we claim that the method does not suffer from the drift error yet its encoding performance in terms of rate distortion ratio is better than the state-of-the-art methods. Moreover, the encoding latency, and the coding complexity of the proposed method are less than the similar methods in literature. As benchmarks we are considering the method proposed by Tong et al [6].

A. Time Complexity Analysis

The proposed method uses a matrix multiplication operation to scramble the privacy related areas of the frames. In this subsection we analyze the processing cost of the scrambling operation with respect to the total frame encoding time. Since we assume only some parts of each frame is scrambled, we consider the parameter ρ as the probability of a block falling into the private area of the frame. Without considering computationally less intensive steps such as interpolation filtering, deblocking, and in-loop filtering, the encoding time of a block is given as shown in Equation 6.

$$T_{Encode} = T_{ME} + T_{DCT} + T_Q + \rho \times T_{MM} + T_{VLC} \quad (6)$$

where T_{ME} is the time spent for motion estimation, T_{DCT} is the time for applying the DCT, T_Q is the time for quantization, T_{MM} is the matrix multiplication time, and T_{VLC} is the time for variable length coding. The extra step added by the proposed method increases the time complexity of the encoder by $O(n^{2.373})$ [15] where *n* is the size of the matrix. However, the scrambling matrix in the proposed method is a diagonal matrix consisting of ± 1. This property reduces the matrix multiplication to eight sign changes. Moreover, considering the following scenarios for the privacy protected parts of a frame, we estimate $\rho = 0.1$.

- A security assistant/officer may want to have his/her face image to be hidden from un-authorized viewers.
- A driver may ask for the protection of his/her face from in videos.
- The license plates of the cars passing by an accident area should be hidden in a public video stream.

The DCT is performed as a matrix-vector product [16] as $y = Tx$. The fast transform is carried out by the factorization of T into a product of sparse structured matrices. Vashkevich et al. [17] presented a fast DCT algorithm which uses 32 multiplication and 81 addition for a 16-point data. The time needed for the DCT in comparison to the matrix multiplication used by the proposed method, considering that the motion estimation accounts for about 60% of the video coding time, and the fact that the scrambling is applied to a small portion of the frame blocks, the impact of the proposed method on the total video coding time is minor.

In decoding stage, the inverse matrix computation ($inv(DCT^{-1}(S))$) is performed only once as long as the key value remains the same. Therefore, without considering the motion estimation, the time complexity values of the encoder is valid for decoder too.

B. Bit-rate Overhead Analysis

In order to avoid the drift error we propose motion compensating the blocks without forbidding the partially scrambled area of the reference frames as explained in Section III. Although in general there will be a slight reduction in the coding efficiency compared to the case when video is encoded without scrambling, we claim that the proposed method provides better bit overhead saving than MRIP method proposed in [6]. This improvement is explained by considering the fact that the MRIP method restricts the search area while the proposed method searches everywhere including the areas searched by MRIP. In our first set of experiments we have compared the performance of the proposed method with MPEG encoder

without scrambling the frames. The results depicted in Figures 3 and 4 show the performance of our proposed method in two video sequences. Our experimental results in all test

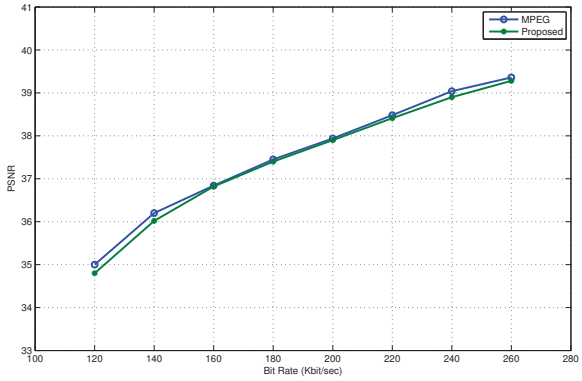

Figure 3. Rate-Distortion comparison of the proposed method and MPEG4 encoder using Foreman video sequence.

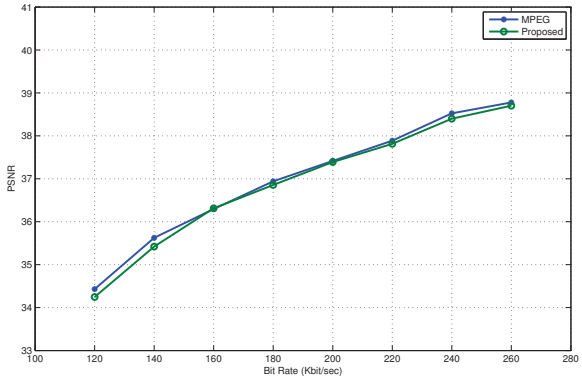

Figure 4. Rate-Distortion comparison of the proposed method and MPEG4 encoder using Container video sequence.

video sequences indicate that the performance of the proposed method is comparable with and only slightly less than MPEG4 video coding standard where scrambling is not applied.

In our second set of experiments we have compared the performance of our method with the case when the search area is restricted to non-scrambled areas. Figures 5 and 6 depict the rate distortion comparisons of the proposed method and the method proposed in [6] using '*Foreman*' and '*Container*' sequences, respectively. Our experiments with sample video sequences reveal that some of the blocks are motion compensated with areas which partially overlap with scrambled areas. This observation explains the improvement in coding efficiency of the proposed method which amount to 0.6 dB on average. Figures 7 and 8 depict the result of scrambling the privacy area and corresponding motion vectors, respectively. It is important to note that despite scrambling the face area, motion vectors do not show significant increase in size which justifies the

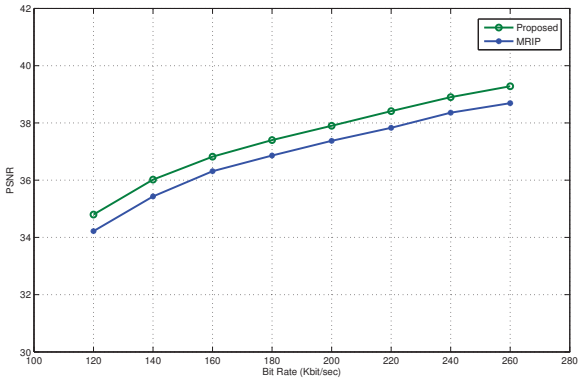

Figure 5. Rate-Distortion comparison of the proposed method and MRIP method using Foreman video sequence.

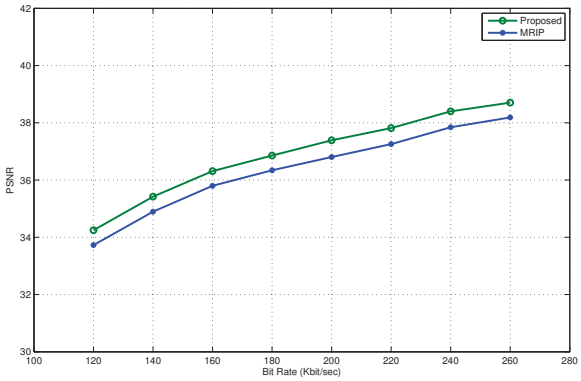

Figure 6. Rate-Distortion comparison of the proposed method and MRIP method using Container video sequence.

performance improvement of the proposed method. Our drift-free method has advantages over methods which include the whole frames in an entire GOP to avoid quality degradation due to the drift error [12]. In this case, the obvious advantage is that only privacy areas are scrambled in the relevant frames in our proposed method whereas, in these methods the entire frame, and some of the irrelevant frames (frames with no privacy importance) are also scrambled.

It should be noted that in our design, the decoder needs to know the location of privacy area or ROI. We assume this information is transmitted as a binary map which includes one bit for each macro-block. This corresponds to 396 bits per frame in CIF format. Since the main interest of this work is eliminating drift error due to scrambling, we have not considered the communication of the so-called binary map.

V. Conclusions

A new drift-free method for scrambling videos for privacy is proposed. The proposed method addresses the quality degradation due to the use of scrambled areas of a frame in motion compensating blocks of succeeding frames. Despite the methods proposed in the literature, our proposed method

Figure 7. Sample Frame from Foreman Video (left), with Scrambled Private Area (right).

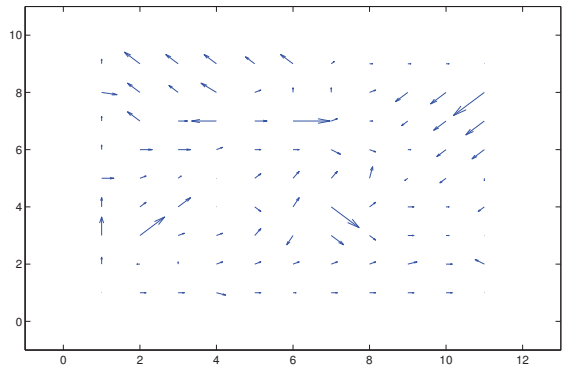

Figure 8. Motion Vectors of the Second Frame from Foreman Video Scrambled using the Proposed Method.

does not forbid utilization of the scrambled areas as reference areas. Besides, our method is capable of reconstructing the video both when the private key is available and not without any degradation in the frame quality. Our experimental results indicate that the performance of our method is comparable to standard video coding methods with only a very slight decrease in rate distortion ratio.

An important feature of the proposed method is that the scrambling and unscrambling steps are independent from the motion estimation/motion compensation and transformation of the residues. Therefore any encoding in the form of a matrix multiplication can be utilized. However, since the scrambling matrix consists of positive and negative ones, the performance degradation is limited. Moreover, the method is compatible with all video coding standards as it is applied after quantization step.

REFERENCES

[1] A. Becker, A. Arnab, and M. Serra. Assessing privacy criteria for DRM using EU privacy legislation. *Proceedings of the 8th ACM workshop on Digital Rights Management*, pages 77–86, 2008.

[2] J.R. Troncoso-Pastoriza, P. Comesaa, L. Prez-Freire, and F. Prez-Gonzlez. Videosurveillance and privacy: Covering the two sides of the mirror with DRM. *Proceedings of the Nineth ACM Workshop on Digital Rights Management*, pages 83–94, 2009.

[3] Y.H. Sohn, H.C. Huang, and S.Y. Wang. Video scrambling and fingerprinting for digital right protection. *International Symposium on Computer, Consumer and Control*, pages 471–474, 2012.

[4] L. Tong, F. Dai, Y. Zhang, and J. Li. Restricted H.264/AVC video coding for privacy protected video scrambling. *Journal of Visual Communications and Image Representation*, pages 479–490, 2011.

[5] H. Sohn, W. D. Neve, and Y. M. Ro. Privacy protection in video surveillance systems: Analysis of subband-adaptive scrambling in JPEG XR. *IEEE Transaction on Circuits and Systems for Video Technology*, 21(2):170–177, 2011.

[6] L. Tong, F. Dai, Y. Zhang, and J. Li. Restricted H.264/AVC video coding for privacy region scrambling. *Proceedings of 17th IEEE International Conference on Image Processing*, pages 2089–2092, 2010.

[7] F. Dufaux and T. Ebrahimi. H.264/AVC video scrambling for privacy protection. *IEEE International Conference on Image Processing*, pages 1688–1691, 2008.

[8] Y. Kim, S. H. Jin, T. M. Bae, and Y. M. Ro. A selective video encryption for the region of interest in scalable video coding. *IEEE Region 10 Conference (TENCON 2007)*, pages 1–4, 2007.

[9] H. Schwarz, D. Marpe, and T. Wiegand. Analysis of hierarchical b-pictures and mctf. *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1929–1932, 2006.

[10] ISO=IEC 144962. *Coding of audio-visual objects*. 1998.

[11] H. Sohn, W. D. Neve, and Y. M. Ro. Region-of-interest scrambling for scalable surveillance video using JPEG XR. *Proceedings of the 17th ACM international conference on Multimedia (MM 09)*, pages 861–864, 2009.

[12] J. Shen, Y.-J. Cai, and L. Luo. A context-aware mobile web middleware for service of surveillance video with privacy. *Multimedia Tools and Applications*, pages DOI 10.1007/s11042–014–2036–9, 2014.

[13] F. Kurugollu Y. Wang, M. ONeill. Privacy region protection for h.264/avc by encrypting the intra prediction modes without drift error in i frames. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2964–2968, 2013.

[14] F. Kurugollu Y. Wang, M. ONeill. Adaptive binary mask for privacy region protection. *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1127–1130, 2012.

[15] V. V. Williams. Multiplying matrices in $O(n^{2.373})$ time. *Stanford University*, 2014.

[16] R. J. Cintra and V. S. Dimitrov. The arithmetic cosine transform: Exact and approximate algorithms. *IEEE Transaction on Signal Processing*, 58(6):3076–3085, 2010.

[17] M. Vashkevich and A. Petrovsky. A low multiplicative complexity fast recursive DCT-2 algorithm. *CoRR*, abs/1203.3442, 2012.

Hierarchical SNR Scalable Video Coding with Adaptive Quantization for Reduced Drift Error

Roya Choupani[12], Stephan Wong[1] and Mehmet Tolun[3]
[1]*Computer Engineering Department, Delft University of Technology, Delft, The Netherlands*
[2]*Computer Engineering Department, Çankaya University, Ankara, Turkey*
[3]*Electrical Engineering Department, Aksaray University, Aksaray, Turkey*

Keywords: Scalable Video Coding, Rate Distortion Optimization, Drift Error.

Abstract: In video coding, dependencies between frames are being exploited to achieve compression by only coding the differences. This dependency can potentially lead to decoding inaccuracies when there is a communication error, or a deliberate quality reduction due to reduced network or receiver capabilities. The dependency can start at the reference frame and progress through a chain of dependent frames within a group of pictures (GOP) resulting in the so-called drift error. Scalable video coding schemes should deal with such drift errors while maximizing the delivered video quality. In this paper, we present a multi-layer hierarchical structure for scalable video coding capable of reducing the drift error. Moreover, we propose an optimization to adaptively determine the quantization step size for the base and enhancement layers. In addition, we address the trade-off between the drift error and the coding efficiency. The improvements in terms of average PSNR values when one frame in a GOP is lost are 3.70(dB) when only the base layer is delivered, and 4.78(dB) when both the base and the enhancement layers are delivered. The improvements in presence of burst errors are 3.52(dB) when only the base layer is delivered, and 4.50(dB) when both base and enhancement layers are delivered.

1 INTRODUCTION

The scalability property of video coding provides the possibility of changing the video quality if it is required by network conditions or display device capabilities of the receiver. The scalability property of video is provided by multi-layer video coding through decomposition of the video into smaller units or layers (Adami et al., 2007). The first layer which includes the video content in its lowest quality (in terms of resolution, frame rate, or bits-per-pixel) is called the base layer. All other layers add to the quality of the video, and are called enhancement layers (Segall and Sullivan, 2007),(Schwarz et al., 2006). The order of including the layers in multi-layer video coding is important and a higher level layer cannot be utilized when the lower level layers are not present (Lan et al., 2007). A significant number of video coding methods using scalable video coding (SVC) schemes have been reported in literature (Segall, 2007),(Ohm, 2005),(Schwarz et al., 2007a),(Abanoz and Tekalp, 2009) and a comprehensive overview paper on SVC methods is presented in (Adami et al., 2007) and (Wien et al., 2007). State-of-the-art video coding methods however, utilize motion-compensated temporal filtering (MCTF), where each inter-coded video frame is encoded by predicting the motion of every macro-block with respect to a reference frame and encoding the differences or residues. When an MCTF-based SVC method delivers only some of the encoded video layers, the reconstructed frames will be different than the encoded frames. The difference $\Delta I'$ between the encoded frame I and the reconstructed frame I' increases at the subsequent decodings based on imperfectly reconstructed reference frames. This error which accumulates until an intra-coded frame is reached, is called the drift error. The drift error is the result of selective transmission where some of the DCT coefficients are eliminated, and/or re-quantized which changes the original quantized DCT coefficients (Yin et al., 2002). The drift error can occur in multi-layer scalable video coding methods if the decoder does not receive all enhancement layer data (Lee et al., 2004). Improving the robustness of SVC methods against packet loss through data redundancy (Abanoz and Tekalp, 2009) or selective protection of layers (Xiang et al., 2009),(LOPEZ-FUENTES, 2011) reduces the bit rate performance of the encoder

Choupani R., Wong S. and Tolun M..
Hierarchical SNR Scalable Video Coding with Adaptive Quantization for Reduced Drift Error.
DOI: 10.5220/0005306001170123
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 117-123
ISBN: 978-989-758-089-5

(Wien et al., 2007). For instance, the enhancement layer(s) information can be used in the motion prediction loop of the encoder to improve the coding performance (Ohm, 2005). Consequently, the absence of the enhancement layer(s) at the decoder can contribute to the drift error.

Some video coding standards such as H.263 and MPEG4 prefer drift-free solutions where the encoder performs motion prediction using only the base layer information. This means that the reconstruction will be error free if only the base layer is delivered. However, these solutions are provided with a reduction in performance. Other approaches that attempt to optimize the coding efficiency while minimizing the drift error have been proposed in literature (Reibman et al., 2001),(Regunathan et al., 2001). In (Seran and Kondi, 2007), the authors report a coding method which maintains two frame buffers in the encoder and decoder. These buffers are based on the base layer, and the base and enhancement layers. They initially use the base and enhancement layer buffer for encoding and decoding. Their method measures the drift error based on the channel information. When the drift error exceeds a predefined threshold, the method switches to the base layer buffer, assuming that the base layer is always available to the receiver. A similar method reported in (Reibman et al., 2003) balances the tradeoff between compression efficiency and the drift error. The authors assume two coding parameters, namely the quantizer and the prediction strategy. By selecting the appropriate parameter based on the network conditions, they try to optimize the video coding process. In (Yang et al., 2002), a method is proposed to minimize the rate distortion by utilizing the distortion feedback from the receiver. The authors assume the base and the enhancement layer macro-blocks can be encoded in different modes. They optimize the coding by choosing the quantization step and the coding mode for each macro-block.

The main problem with these methods is that the decision about optimizing the encoder parameters is made by considering the average value of drift error. As a result, the same parameter values are applied to all frames of a group of picture (GOP). However, since the drift error cannot propagate beyond a GOP, each frame contributes to the accumulation of error with a different rate. For instance, the last frame in a GOP has no impact on error accumulation while the error happening in the first frame propagates until the end of the current GOP. In this paper, we address the video quality degradation due to the drift error in SVC. We consider adjusting the coding parameters according to the network conditions and the frame position in the GOP. We propose a method to improve the coding efficiency in terms of the R-D ratio, while reducing the drift error whenever the reconstruction is performed using the base layer only. Moreover, we consider measurements to make the encoded video robust against single and multiple frame losses.

2 OPTIMIZING VIDEO ENCODER PERFORMANCE BY MINIMIZING THE DRIFT ERROR

Video coding optimization and visual quality preservation have conflicting requirements. Motion-compensation techniques for instance are not robust against frame losses and apt to quality loss due to the drift error. Our proposed method for reducing the drift error while preserving the coding efficiency is based on the following observations:

- The dependency of a frame to its preceding frame creates a chain of frames that are dependent on each other (dependency chain). The drift error has a direct correlation with the number of video frames in a dependency chain. On the other hand, a longer GOP provides a better I-frames to P/B-frame ratio and hence a smaller bit-per-pixel rate. In (Goldmann et al., 2010) the video quality degradation due to the drift error is analyzed subjectively. Although the quality degradation varies with the spatial details and the amounts of local motion, the quality of video drops below fair for GOP lengths greater than 5. The result of this analysis is compatible with our observation.
- The drift error also has direct correlation with the mismatch between the original frames and the reconstructed frames. When some part(s) of a frame data is lost or corrupted, the other parts are used for the frame reconstruction. In multi-layer SVC, the receiver may reconstruct the video using the base layer, or the base layer and some of the enhancement layers. Hence, the size of the enhancement layer(s) should be adjusted with the maximum tolerable distortion rate of the video.

Based on the above observations, the proposed method reduces the number of dependent frames by introducing a dyadic hierarchical structure. Besides, the amount of data in the base and enhancement layers is adjusted adaptively as a function of the location of the frame in the dependency chain. An optimum GOP length is sought after to minimize the drift error while preserving the performance of the encoder.

The amount of data transmitted in the enhancement layer(s) and the quality degradation due to the drift error are inversely proportional and hence, an optimum balance should be found for the best performance and the least distortion. In this paper we consider only one enhancement layer.

2.1 R-D Optimization in Hierarchical Coding of Video Frames

In the proposed multi-layer SVC, different quantization parameters are used in the base and the enhancement layers. The motion compensated blocks, which we refer to as residues, are transformed using DCT and quantized using two different quantization step-sizes. A fine quantization which produces larger quantized coefficients (considering the absolute values), and a coarse quantization which results in smaller quantized coefficients. We use the coarse quantization results as the base layer. The difference between the fine quantized coefficients and the coarse quantized coefficients are considered as the enhancement layer. The encoding and decoding processes can be expressed as shown in Equations 1 and 2 where BL and EL represent the base layer and the enhancement layer bitstreams, respectively.

$$\begin{aligned} BL &= VLC(Q(DCT(Residues), QP_b)) \\ EL &= VLC(Q(DCT(Residues), QP_e) - \\ &\quad Q(DCT(Residues), QP_b)) \end{aligned} \tag{1}$$

Reconstruction using only the base layer (BL'), and the base and the enhancement layers (BEL') are shown in Equation 2.

$$\begin{aligned} BL'(Residues) &= IDCT(IQ(IVLC(BL), QP_b)) \\ BEL'(Residues) &= IDCT(IQ(IVLC(BL) + \\ &\quad IVLC(EL), QP_e)) \end{aligned} \tag{2}$$

where QP_b and QP_e are the base layer and the enhancement layer quantization parameters, respectively, and IVLC is the inverse of the variable length coding process. As it is shown in Equation 2, the reconstructed frame is obtained from inverse discrete transform of the base and the enhancement layers quantized residues. Whenever the enhancement layer is not delivered, the reconstructed frame is deviated from the encoded frame. This deviation is a function of the amount of data in the base and the enhancement layers, which are determined by the quantization parameters of these layers namely QP_b and QP_e, and the number of the frames in a dependency chain which indicates the propagation extent of the drift error. On the other hand, the bit rate of the base layer is a function of QP_b. Hence, for a given bit rate, the optimized coding efficiency and lowest rate distortion depend on the QP_b, QP_e, and GOP length parameters. The drift error can be largely reduced by utilizing a hierarchical dyadic organization of the frames in a group of pictures which restricts the maximum error propagation range to $\lceil \log_2 GOP \rceil$ (Schwarz et al., 2007b). Clearly, not all frames are used as a reference frame while some frames are used as reference for many frames. These observations lead us to adapt the quantization parameters QP_b and QP_e with the position of the frame in a GOP for each bit rate. This adaptation results in different distortion levels in the frames of a GOP while the average distortion is minimized. The rate distortion optimization in a GOP given the base layer and the enhancement layer quantization step-sizes is shown in Equation 3. We assumed the video contains only one enhancement layer however, it is readily extendable to include several enhancement layers.

$$\begin{aligned} &J(QP_b, QP_e, GOPlen, \rho) = \sum_{i=1}^{GOPlen} \ldots \\ &D_i(QP_b(i), QP_e(i)) + \lambda_i R_i(QP_b(i), QP_e(i), \rho) \end{aligned} \tag{3}$$

where J is an auxiliary function denoting the optimization process, GOPlen is the number of frames in a GOP, λ is the Lagrange multiplier. D_i is the distortion and R_i is the bit rate of frame i when quantization parameters $QP_b(i)$ and $QP_e(i)$ are used, respectively. The optimization is carried out for a given bit rate, ρ, and over a GOP. The summation in Equation 3 therefore, minimizes the total distortion of frames in a GOP, when their total bit rate is limited to ρ. The length of the dependency chain is a determining factor in the total distortion of the video due to the drift error. Therefore, the rate distortion problem depends on the quantization parameters of each frame in a GOP, and the GOP length. Since we arrange the frames of a GOP in a dyadic hierarchical structure, each GOP contains many dependency chains which should be considered in optimization process.

2.2 The Scalability Features of the Proposed Method

Signal-to-noise (SNR) scalability in the proposed method is provided as a multi-layer coding of the frames where the number of layers determines the granularity of the video with the main feature of having a different approach for handling the drift error. For instance, the fine granularity quality scalable (FGS) coding in MPEG-4 was chosen so that the drift error is completely omitted by using base layer frames as reference frames in motion compensation. It is obvious that the drift free coding of MPEG-4 comes with a reduction in coding efficiency. However, our approach is based on balancing the bit rate

with the distortion caused by the drift error. The quantization parameters after decomposing the frames into the base and the enhancement layers is adapted in a way that in the frames which serve as reference for a larger number of frames, the enhancement layer is smaller and hence the inaccuracy with the original frame when the enhancement layer is missing becomes smaller.

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 important restriction in the temporal scalability feature of the traditional methods is that the number of layers determine the achievable temporal scalability rate(s). This means that a continuous temporal scalability is not feasible in these methods whereas, this feature is provided in the proposed method as described below. In the proposed method, the hierarchical organization of the frames provide several dependency chains. Since eliminating a frame from end of a dependency chain does not cause any drift-free, we perform temporal down-sampling by removing these frames in each GOP. For instance, assuming a GOP of 16 frames (Figure 1) the dependency chains and order of the frames for elimination for temporal down-sampling is as below:

$$
\begin{array}{llllllll}
1 & \rightarrow & 2 & & & & & \\
1 & \rightarrow & 3 & \rightarrow & 4 & & & \\
1 & \rightarrow & 5 & \rightarrow & 6 & & & \\
1 & \rightarrow & 5 & \rightarrow & 7 & \rightarrow & 8 & \\
1 & \rightarrow & 9 & \rightarrow & 10 & & & \\
1 & \rightarrow & 9 & \rightarrow & 11 & \rightarrow & 12 & \\
1 & \rightarrow & 9 & \rightarrow & 13 & \rightarrow & 14 & \\
1 & \rightarrow & 9 & \rightarrow & 13 & \rightarrow & 15 & \rightarrow 16
\end{array}
$$

$$\textit{Frame elimination order}: 2,4,6,8,10,12,14,16,3,7,11,15,5,13,9,1$$

It is worth to note that the temporal scalability property of the proposed method is drift error free.

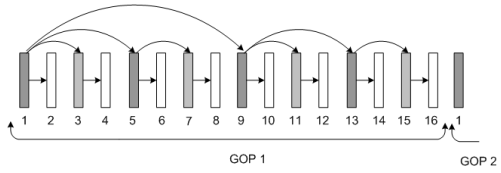

Figure 1: The Dependency Chains in the Dyadic Hierarchical Structure for Multi-layer SNR Scalable Video Coding.

3 EXPERIMENTAL RESULTS

The proposed method is experimentally verified using some video sequences. In order to verify the performance of our method, we need to determine the optimization parameters. Optimized quantization parameters for each frame is computed iteratively. As explained in Section 2.1 the QP for the base and the enhancement layer(s) are optimized to minimize the distortion due to the drift error for a given bit rate. The optimization variables are the quantization step size of each frame which is dependent on the position of the given frame in the frame dependency chain and the GOP length. Considering that the maximum length of a frame dependency chain in a dyadic hierarchical organization of the frames is $\log_2(GOP)$, we express the QP for each frame as shown in Equation 4.

$$
\begin{array}{ll}
QP_{StepE}(i) = QP_{StepE}(i) + & \Delta_{QP} \\
QP_{StepB}(i) = QP_{StepE}(i) + & (log_2(GOP) - Pos_i) \\
 & \times \Delta_{QP} + \tau_{QP}
\end{array} \quad (4)
$$

where QP_{StepB} is the quantization step size used at the base layer (lowest quality), QP_{StepE} is the quantization step size used for the highest quality quantization (base + enhancement), i refers to the current frame in the GOP, Pos_i is the number of frames dependent on the current frame (frame i) in the longest frame dependency chain, Δ_{QP} is the QP step size increment, and τ_{QP} is a constant used as the step size bias value. QP_{StepE} step size is incremented by adding the QP step size increment and then QP_{StepB} is optimized. Quantization matrices QP_b and QP_e are related to QP_{StepB} and QP_{StepB} as shown in Equation 5.

$$
\begin{array}{l}
QP_b = Q \times QP_{StepB} \\
QP_e = Q \times QP_{StepE}
\end{array} \quad (5)
$$

where Q is the default quantization table used by MPEG-4. To determine the optimum value for the step sizes, we iteratively tried different values of Δ_{QP} for GOP lengths of 8, 16, 32, and 64. The highest total quality in a GOP (minimum distortion) for a given bit rate is sought as the optimized quantization parameters which depend on the content of the frame in that GOP.

The proposed method is experimentally evaluated by comparing its performance against the following methods:

- Drift-free implementation where the base layer of the reference layer is used for motion prediction. Drift-free methods have the advantage of experiencing no distortion in terms of error accumulation when the enhancement layer is not delivered however, they suffer from coding performance.
- Hierarchical organizing the frames with a fixed quantization parameter optimized for the whole GOP. This experiment shows the gain we obtain by adaptively optimizing the quantization parameter which is the main contribution of the proposed method.

- The method proposed in (Yang et al., 2002) optimizes the rate distortion of SNR SVC video coder by determining the coding mode for each MB. Their assumption of using enhancement layer data of the reference frame for motion prediction of the current frame, and transmitting each frame in one packet are similar to our assumptions and hence makes a more realistic comparison possible.
- Verifying burst error effect. This experiment verifies the impact of single and burst errors when only base layer, and when both base and enhancement layers are delivered.

We measured the performance of the proposed method when the videos are scaled down and only the based layer is delivered. In this experiment, the videos are encoded using the proposed method with hierarchical frame organizations and adaptive quantization step size, and the sequential coding of the video with a fixed quantization step size. The proposed method outperforms the sequential video encoding by an average PSNR improvement of 2.86(dB). The PSNR values of the reconstructed frames for both methods have been depicted in Figure 2. The second

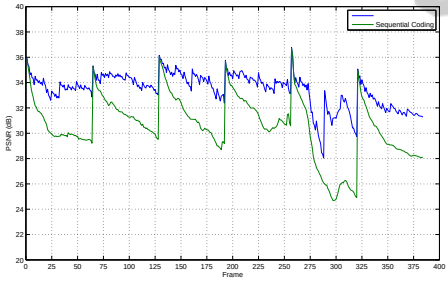

Figure 2: PSNR values of the reconstructed frames by the proposed method and the sequential coding method using base layer only.

set of experiments measures the performance of the proposed method compared to the drift-free method suggested in MPEG-4 (144962, 1998)(Peng et al., 2005), and the adaptive allocation method proposed in (Yang et al., 2002). The authors of (Yang et al., 2002) assume no data loss happens in the base layer. Therefore, the distortions feedback from the receiver are the result of losses at the enhancement layer and the drift error. Since our proposed method does not rely on the feedback from the receiver, we modified the method proposed in (Yang et al., 2002) to optimize for a given bit rate. We implemented their proposed low complexity sequential optimization method where the base layer and the enhancement layer are optimized sequentially, considering no error concealment and frame re-transmission in the network. We assume the videos are encoded for different bit-rates. Besides, a 10% frame loss is imposed in the transmissions where the position of the lost frames are randomly selected but are the same in all three methods. Figure 3 depicts the results of the comparison. The proposed

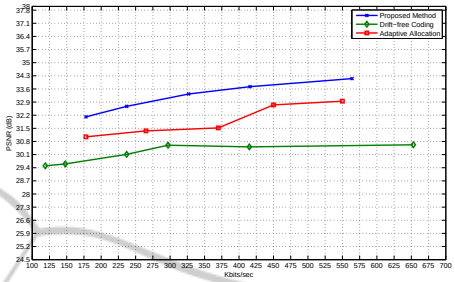

Figure 3: PSNR at different bit rates with 10% frame loss.

method provides better performance than the drift-free sequential coding with fixed quantization, and adaptive bit-rate allocation proposed in (Yang et al., 2002). The main reason for the better performance of the proposed method is the shorter dependency chains in the GOPs. Since the drift error results in more serious quality degradation when the lost frame is farther from the end of the dependency chain, the proposed method experiences a lower level of performance loss.

Our final experiment evaluates the robustness of the proposed method in presence of the frame loss. The experiment includes two cases. In case one several frames at random positions of a GOP are lost. The reconstructed videos when some frames are missing are evaluated by measuring the PSNR values of:

- the base layer of the delivered frames only where we assume the videos are scaled down,
- and the base and the enhancement layers, in which case we assume the videos are transmitted without scaling down.

The comparative results are illustrated in Figures 4 and 5. The second case for robustness evaluation is considered to measure the video quality degradation in presence of burst errors. A burst error is defined as a sequence of missing frames with a length of 5 to 10 frames. Figures 6 and 7 depict the results of the burst error experiments. The results of the experiments indicate that the proposed method outperforms the traditional video coding methods in presence of frame losses. The average PSNR values when both the base and the enhancement layers are delivered are 31.36(dB) and 27.66(dB) in the proposed method and the standard video coding respectively. The average PSNR values when only the base layer is delivered are 30.58(dB) and 25.80(dB) at the pro-

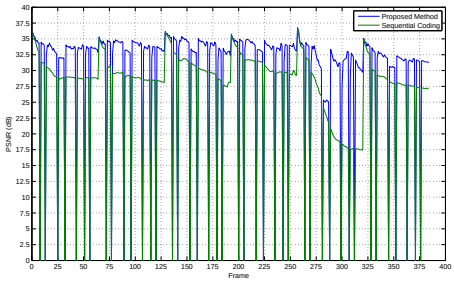

Figure 4: PSNR of the reconstructed frames using only the base layer in presence of single frame losses.

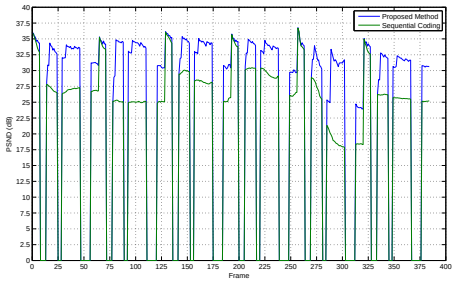

Figure 6: PSNR of the reconstructed frames using only the base layer in presence of multiple frame losses.

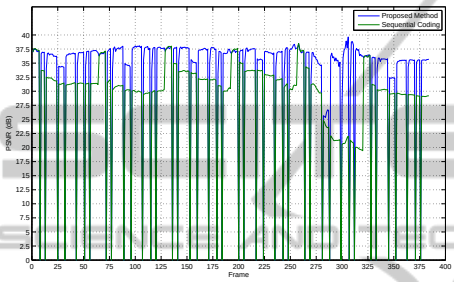

Figure 5: PSNR of the reconstructed frames using the base and the enhancement layers in presence of single frame losses.

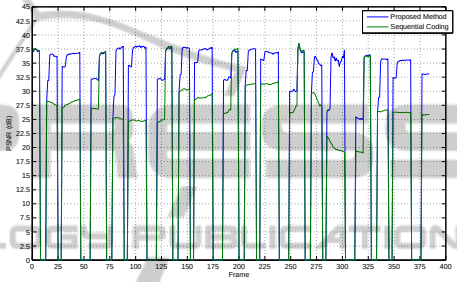

Figure 7: PSNR of the reconstructed frames using the base and the enhancement layers in presence of multiple frame losses.

posed method and the sequential coding respectively. This improvement can be associated with two effective factors. The first factor is the hierarchical structure of arranging the frames which makes the frame dependency chains shorter in the proposed method. The second factor which is valid when only the base layer is delivered is the adaptive quantization of the frames. We reconstruct the missing frames with a preceding intact frame having a higher level of accuracy in the reference frame. The effect of this factor is evident from the average PSNR values of the delivered frames where the difference in average PSNR value when both layers are delivered is 4.78(dB) while it is 3.70(dB) when only the base layer is delivered. The improvements in the robustness of the video in presence of burst errors are 3.52(dB) for the base layer only delivered videos where the average PSNR values are 22.32(dB) and 18.8(dB) for the proposed method and the sequential coding respectively, and 4.50(dB) when the base and the enhancement layers are delivered with the average PSNR values are 24.01(dB) and 19.51(dB) for the proposed method and the sequential coding respectively.

It is important to note that the optimization by the proposed method is carried out after motion estimation and the DCT steps of video coding and hence quite efficient in terms of processing time.

4 CONCLUSIONS

A new scalable video coding method for reducing drift error has been proposed. The proposed method utilizes the hierarchical organization of the video frames, and optimizes coding by adapting quantization step size of each frame according to its position in a GOP. The method is used for SNR, and temporal video scaling in presence of frame loss in noisy communication networks. The proposed method improves the performance of the SVC coder by relying on the observation that elimination of the drift error reduces the coding performance. Therefore, an optimization should be sought to reduce the distortion due to the drift error while preserving the quality of the transmitted video. The optimized video has a multilayer SVC format where the enhancement layer size is adaptively changed according to the network conditions and the frame position in GOP for minimum distortion. The improvement attained by the proposed method is at least 3.52(dB) in terms of PSNR values.

REFERENCES

144962, I. (1998). *Coding of audio-visual objects*.

Abanoz, T. B. and Tekalp, A. M. (2009). Svc-based scalable multiple description video coding and optimization of encoding configuration. *Signal Processing: Image Communication*, 24:691–701.

Adami, N., Signoroni, A., and Leonardi, R. (2007). State-of-the-art and trends in scalable video compression with wavelet-based approaches. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1238–1255.

Goldmann, L., Simone, F. D., Dufaux, F., Ebrahimi, T., Tanner, R., and Lattuada, M. (2010). Impact of video transcoding artifacts on the subjective quality. *International Workshop on the Quality of Multimedia Experience (QoMEX), Second*, pages 52–57.

Lan, X., Zheng, N., Xue, J., Gao, B., and Wu, X. (2007). Adaptive vod architecture for heterogeneous networks based on scalable wavelet video coding. *IEEE Transactions on Consumer Electronics*, 53(4):1401–1409.

Lee, Y.-C., Altunbasak, Y., and Mersereau, R. M. (2004). An enhanced two-stage multiple description video coder with drift reduction. *IEEE Transaction on Circuits and Systems for Video Technology*, 14(1):122–127.

LOPEZ-FUENTES, F. A. (2011). P2p video streaming combining svc and mdc. *International Journal of Applied Mathematics and Computer Science*, 21(2):295–306.

Ohm, J. (2005). Advances in scalable video coding. *Proceedings of the IEEE*, 93(1):42–56.

Peng, W.-H., Tsai, C.-Y., Chiang, T., and Hang, H.-M. (2005). Advances of mpeg scalable video coding standard. *Knowledge-Based Intelligent Information and Engineering Systems*, 3684:889–895.

Regunathan, S., Zhang, R., and Rose, K. (2001). Scalable video coding with robust mode selection. *Signal Processing: Image Communication*, 16(8):725–732.

Reibman, A., Bottou, L., and Basso, A. (2003). Scalable video coding with managed drift. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(2):131–140.

Reibman, A., Bottou, U., and Basso, A. (2001). Dct-based scalable video coding with drift. *IEEE International Conference on Image Processing (ICIP2001)*, 2:989–992.

Schwarz, H., Marpe, D., and Wiegand, T. (2006). Analysis of hierarchical b-pictures and mctf. *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1929–1932.

Schwarz, H., Marpe, D., and Wiegand, T. (2007a). Overview of the scalable video coding extension of the h.264/avc standard. *IEEE Transaction on Circuits and Systems for Video*, 17(9):1103–1120.

Schwarz, H., Marpe, D., and Wiegand, T. (2007b). Overview of the scalable video coding extension of the h.264/avc standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 17(9):1103–1120.

Segall, A. (2007). Ce 8: Svc-to-avc bit-stream rewriting for coarse grain scalability. *Joint Video Team, Doc. JVT-V035*.

Segall, A. and Sullivan, G. J. (2007). Spatial scalability. *IEEE Transaction on Circuits Systems for Video Technology*, 17(9):1121–1135.

Seran, V. and Kondi, L. (2007). Drift controlled scalable wavelet based video coding in the overcomplete discrete wavelet transform domain. *Journal of Image Communication*, 22(4):389–402.

Wien, M., Schwarz, H., and Oelbaum, T. (2007). Performance analysis of svc. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1194–1203.

Xiang, W., Zhu, C., Siew, C. K., Xu, Y., and Liu, M. (2009). Forward error correction-based 2-d layered multiple description coding for error-resilient h.264 svc video transmission. *IEEE Transaction on Circuits and Systems for Video Technology*, 19(12):1730–1738.

Yang, H., Zhang, R., and Rose, K. (2002). Drift management and adaptive bit rate allocation in scalable video coding. *IEEE International Conference on Image Processing*, 2:49–52.

Yin, P., Vetro, A., Lui, B., and Sun, H. (2002). Drift compensation for reduced spatial resolution transcoding. *IEEE Transaction on Circuits and Systems for Video Technology*, 12:1009–1020.

Chapter 3

Optimizing Multiple Description Video Coding In Spatial Domain

Data transmission using unreliable networks such as wireless networks is prone to packet losses. A re-transmission of the lost packets adds extra delay to the total transmission time and in case of video transmission, may cause jitter in its display at the receiver side. A more serious problem is due to the packet loss in communicating realtime data -such as video streaming- where re-transmissions are not possible. In case of the video streaming, the packet loss may result in the loss of a video frame and hence, the reconstruction of the succeeding frames becomes impossible (or subject to major quality degradation). Multiple description techniques are among the techniques used for handling the impact of packet losses (Section 1.1.3). The MDC methods decompose the video into independent streams named descriptions where each stream is transmitted (possibly) over an independent communication line. If one of these descriptions is lost or damaged, it is estimated/interpolated using the descriptions delivered intact. Hence, on one hand the decomposition of the video should preserve the correlations between the descriptions for reconstructing the video when packet losses occur. But on the other hand, preserving the correlations reduces the coding efficiency. Hence, the challenge in this case is finding suitable methods for decomposing the video. The video decomposition can be in one of the following ways:

- Spatial decomposition: In spatial decomposition, the pixels of a frame are put in different descriptions. The proximity of the pixels transmitted in different descriptions helps in interpolating the missing data however, since each description is encoded independently, the spatial redundancy of the video is not eliminated.

- Temporal decomposition: In temporal decomposition frames are transmitted using different descriptions. Since the descriptions are required to be independently decodeable, the frame being encoded and its reference frame(s) are transmitted in the same description. However, this preserve temporal redundancy partially and hence, reduces the coding efficiency.

- SNR decomposition: Here the assumption is that each independent description should increase the quality of the video by adding to color depth value of the frames.

We consider all three types of MDC of videos in this chapter. An important characteristic of the descriptions is that they can be reconstructed independently from each other. In SNR decomposition of videos, in order to create independent descriptions with the same importance, we have proposed a transform. The descriptions are further decomposed into multiple layers providing scalability inside

each description. Therefore, our proposed method is a combination of MDC and SVC methods. In spatial decomposition of the video frames, we distribute the pixels among descriptions. In case of a description loss, a weighted sum of the pixels from the delivered descriptions is used to estimate the missing pixel values. Our proposed method includes SVC combination with MDC as well. In temporal decomposing, generally the frames are distributed among different descriptions and motion compensation is performed inside each descriptions independently. However, since this decomposition increases the temporal distance between a frame and its reference frame, the coding efficiency is reduced. We proposed a new algorithm for assigning frames to different streams based on the similarity between the frames. The proposed method improves coding efficiency while preserving the balance in the size of the descriptions. The details of the proposed methods and their experimental evaluations have been presented in the following research articles:

- R. Choupani, S. Wong, M.R. Tolun, Multiple Description Coding for SNR Scalable Video Transmission over Unreliable Networks, Springer Journal of Multimedia Tools and Applications, Volume 69, Issue 3, (June 2012), pp 843-858.

- R. Choupani, S. Wong, M.R. Tolun, Multiple Description Scalable Coding for Video Transmission over Unreliable Networks (July 2009), International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (IC-SAMOS 2009), 20-23 July 2009, Samos, Greece.

- R. Choupani, S. Wong, M.R. Tolun, Optimized Multiple Description Coding for Temporal Video Scalability (June 2013), The Fifth International Conference on Wireless and Mobile Network (WiMo 2013), 7-9 June 2013, Konya, Turkey.

Multimed Tools Appl
DOI 10.1007/s11042-012-1150-9

Multiple description coding for SNR scalable video transmission over unreliable networks

Roya Choupani · Stephan Wong · Mehmet Tolun

Abstract Streaming multimedia data on best-effort networks such as the Internet requires measures against bandwidth fluctuations and frame loss. Multiple Description Coding (MDC) methods are used to overcome the jitter and delay problems arising from frame losses by making the transmitted data more error resilient. Meanwhile, varying characteristics of receiving devices require adaptation of video data. Data transmission in multiple descriptions provides the feasibility of receiving it partially and hence having a scalable and adaptive video. In this paper, a new method based on integrating MDC and signal-to-noise ratio (SNR) scalable video coding algorithms is proposed. Our method introduces a transform on data to permit transmitting them using independent descriptions. Our results indicate that on average 1.71dB reduction in terms of Y-PSNR occurs if only one description is received.

Keywords Scalable video coding · Multiple description coding · Multimedia transmission

R. Choupani · S. Wong
Computer Engineering Department, TU Delft, Delft, The Netherlands

S. Wong
e-mail: J.S.S.M.Wong@tudelft.nl

R. Choupani (✉)
Computer Engineering Department, Çankaya University, Ankara, Turkey
e-mail: roya@cankaya.edu.tr

M. Tolun
Computer Engineering Department, TED University, Ankara, Turkey
e-mail: mehmet.tolun@tedu.edu.tr

Published online: 14 June 2012

1 Introduction

With the steady increase in the Internet access bandwidth, increasingly more applications utilize streaming audio and video contents [27]. This trend has been further intensified by the appearance of small, powerful hand-held terminals (such as mobile phones, iPods, and tablet PCs) in the market. In streaming video applications, the servers normally have to serve a large number of users with different screen resolutions and network bandwidth and processing capabilities. Hence, an encoding method that makes use of a single encoded data for all types of bandwidth channels and displaying device capacities is of remarkable significance in multimedia applications. Scalable Video Coding (SVC) schemes are intended to be a solution for the Internet heterogeneity and receiving devices diversity problem by encoding the data at the highest quality but enabling the receiver to utilize it partially depending on its screen, memory, or processing capabilities, or the available bandwidth [15, 19, 26]. However, communication networks offer channels with varying bandwidth [8, 15] which together with the higher rate of frame loss or corruption in wireless networks becomes a complicated issue for video streaming. On the other hand, the main drawback of the currently available scalable video coding methods is that they are not suitable for non-reliable environments with a high rate of frame loss or corruption. This problem stems from the fact that the SVC methods are based on the Motion-Compensated Temporal Filtering (MCTF) scheme [11] where the frames are coded as the difference with a (generally prior) reference frame. In case of a reference frame loss or corruption, the whole chain of the motion compensated frames which depend on this reference frame becomes unrecoverable. To increase the error resilience of the video coding schemes, Multiple Description Coding (MDC) methods were introduced [12, 22, 25]. These methods improve the error resilience of the video by adding redundancy to the encoded data. In case a frame is lost or corrupted, redundancy is used to replace it with an estimated frame. Some researchers have considered the frame loss problem and have not addressed the scalability issue. Franchi et al. proposed a method to send a video by utilizing independent multiple descriptions. However, their method does not combine scalability features with multiple description coding and therefore does not deal with the bandwidth variation problem [10]. The combination of scalable video coding methods and multiple description coding has been addressed by some researchers recently [3, 15, 19]. In the proposed approaches the video data is partitioned into disjoint sets such as the group of odd and even frames in temporal MDC. These approaches take advantage of the correlation between the adjacent data items for estimating the lost data. However, when considering the signal-to-noise-ratio scalability, the assumption of correlated data is not valid, because the bits composing a pixel value cannot be interpolated from each other. An intuitive example is putting the more significant bits of a pixel value in one description and less significant bits in another one. The more significant bits cannot be estimated from the less significant bits in case that they are lost during the transmission. In this study, we propose a method which aims at expressing SNR scalable video coding scheme by multiple equivalent descriptions. In order to achieve this aim, we propose a transform which allows the data bits to have a contribution in each description. In this way, each description besides to conveying the most basic part of data values, is capable of refining the basic part of data. Our proposed method falls into the class of methods which combine MDC with SVC schemes. Our results

indicate that on average 1.71dB reduction in terms of Y-PSNR occurs if only one description is received. The remainder of this paper is organized as follows: Section 2 introduces the main multiple description coding methods. Section 3 describes the details of our proposed method. In Section 4, we introduce the theoretical base of our performance evaluation method and provide the experimental results and finally, in Section 5, we draw our conclusions.

2 MDC-based video coding techniques

Multiple descriptions have attracted a lot of attention as an error resilient way of encoding and communicating visual information over lossy packet networks. A multiple description coder divides the video data into several bit-streams called descriptions which are subsequently 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 these descriptions does not affect decoding of the others. The accuracy of the decoded video depends on the number of received descriptions. 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 recovered from the other description only. Descriptions are defined by constructing P non-empty *sets* summing up to the original signal f. Each set in this definition corresponds to one 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 increase the redundancy which results in a subsequent increase in the data size and reduced efficiency. Designing descriptions as partition, does not necessarily mean that there will be no redundancy in the data. In fact, designing the descriptions as partitions prevents extra bits to be added to the original data for error resilience but still a redundancy in the form of reduced coding efficiency exits. In case of a data loss, the correlation between the spatially or temporally close data can be used for estimating the lost bits. The estimation process is commonly referred

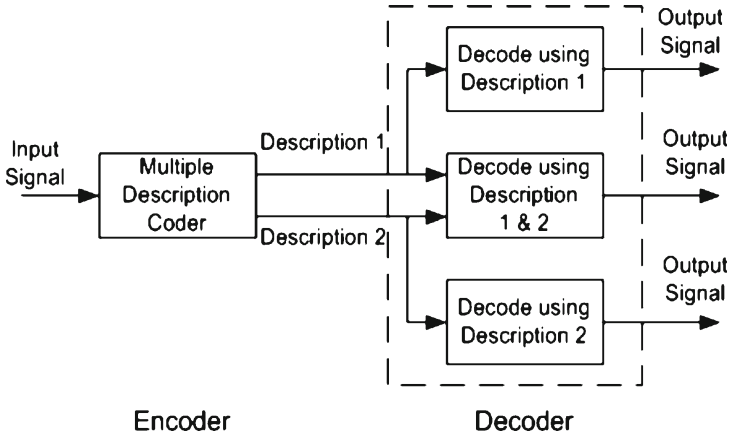

Fig. 1 Multiple descriptions coding block-diagram

to as error concealment and relies on the preserved correlation in constructing the descriptions. MDC schemes for video transmission can be classified as below:

– Multi-layer MDC schemes partition the video into one base layer and one or several enhancement layers [5]. 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 multi-layers puts the 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 corrupted. Repeating base layer bits in each description is the overhead added for a better error resilience. In [1] the authors propose to generate multiple scalable descriptions from a single SVC bit-stream by mapping scalability layers of different frames to different descriptions. Their scheme is intended for Peer-to-Peer (P2P) streaming over multiple multicast trees and features several encoding parameters, such as base layer rate of descriptions and overall redundancy, to optimize for mean rate-distortion performance of each description received over a packet loss network, range of extraction points of the SVC stream, and overall redundancy of their MDC scheme. In [9] the SVC is combined with MDC schemes, by sub-sampling in both horizontal and vertical directions which yields four subsequences. The authors use two approaches to combine the subsequences into two descriptions. In the first approach, each description is encoded by predicting one subsequence from the other using the inter layer prediction tools. The second approach exploits the redundancy between the subsequence with the hierarchical dyadic B frame prediction algorithm. The authors in [17] present a solution for the differences in the types of delivered services in H.264-based SVC combined with MDC by using optimization and control strategies. In [18] an algorithm is proposed to control the mismatch between the prediction loops at the encoder and decoder in MDC with motion-compensated predictions. They consider three cases when both descriptions received or either of the single descriptions is received.
– Forward Error Correction (FEC)-based MDC methods assume that the video is originally defined in a multi-resolution manner [16, 23]. This means if we have M levels of quality, each one is adding to the fidelity of the video with respect to the original one. This concept is similar to the multi-layer video coding method used by FGS scheme. The main difference, however, is that there exists 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. [14] used Unequal Loss Protection (ULP) to protect 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. In [28] a 2-D layered multiple description coding (2DL-MDC) for error-resilient video transmission over unreliable networks is used which encodes each group of pictures (GOP) using the SVC extension of H.264

into sub-streams. First dimension of encoding uses temporal scalability while the second dimension uses SNR scalability. Assuming that the temporal scalability takes priority over the SNR scalability, they put the base layer sub-streams in one group and the rest of the sub-streams in the other one and use FEC with ULP at each group. The first x packets from the first group and y packets from the second group are gathered in description one and the rest in description two. In [13] the authors combine SVC with MDC for video multicasting over P2P networks. Their proposed method uses one base layer and two enhancement layers for SVC. They use FEC with ULP to assign a higher priority to the base layer. 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 convenient 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 [3, 7, 20, 21, 29]. More efficient methods utilize MCTF which is aimed at removing the temporal redundancies of video sequences. In [4] MDC-SVC based on MCTF and 2D DWT is used for video streaming over P2P networks. The receiving peer can measure the channel conditions such as the packet loss rate and bandwidth of each sending peer's path in each GOP period and then calculates the optimal encoder parameters for that GOP through a post-encoding procedure. The resultant encoding parameters are sent to the sending peers through the feedback control channels. Also in [2] an adaptive P2P video streaming system with a flexible multiple description coding (F-MDC) framework is proposed, so that the number of base and enhancement descriptions, and the rate and redundancy level of each description can be adapted. They combine their F-MDC framework with SVC by using JPEG2000 based T+2D DWT which lets them truncate each code-block at any point of bit-plane codes.
– 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 domain 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 sub-sampled version of the signal, defines a description. Chang and Sang [5] utilize the even-odd splitting of the coded speech samples. For images, Tillo et al. [20] propose splitting the image into four sub-sampled 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.

Our proposed method falls into the group of multi-layer MDC schemes. We have proposed a transform to minimize the base layer size which is the main source of redundancy in these schemes. The proposed method allows us to split the video data into two descriptions although the method can be extended to 4, 8, and more descriptions by repeatedly applying the transform on the data.

3 Our proposed method

Our proposed method involves splitting video into two descriptions each representing video in a lower quality. In our previous work [6], we split each frame of the video spatially into four descriptions. In case of loss or damage in one of the descriptions we estimated the missing data from the remaining descriptions. The data belonging to one of the descriptions is not repeated in other descriptions and the redundancy introduced was in the form of inefficiency in motion compensation. We used the correlation between the adjacent pixels to estimate the missing data. Expressing SNR scalability however, is not feasible using the same method. The bits representing a pixel value do not show any correlation with each other. In SNR scalable video coding techniques, the video is split into two or more layers where the first layer, called the base layer, includes the most essential information and the remaining layers, called enhancement layers, improve the base layer data. The main drawback of these techniques is that the enhancement layers cannot be used whenever the base layer is damaged or lost. This means that when SNR scalability techniques are combined with MDC methods the base layer should be repeated in all descriptions which introduces a large redundancy and cause a decrease in bit rate efficiency. A second problem is that the importance levels of the enhancement layers are not the same. This characteristic arises from the fact that bits at different positions convey different values. Hence, descriptions with equal importance cannot be defined by simply distributing the bits between the descriptions. The solution proposed here defines the base layer in a way that each bit has a contribution in it. Figure 2 depicts the block diagram of our proposed method. The left side blocks refer to the MCTF video encoding where ME/MC indicate motion estimation/motion compensation, DCT is the discrete cosine transform, and Q refers to the quantization step. The right side blocks show the MDC coder proposed in our paper, and variable length coder (VLC). The output of our transform is sent to variable length coder where the descriptions are created. The process indicated by Description Encoder in the block diagram of Fig. 2 gets as input the quantized coefficients of the cosine transform and splits them into two descriptions. We propose a transform $\tau(.)$ to create the descriptions as specified in (1) where A is the original data, and B_1 and B_2 are the data transmitted in each description.

$$\tau(A) = [B_1, B_2] \tag{1}$$

The inverse of this transform reconstructs the original data value as indicated by (2).

$$\tau^{-1}([B_1, B_2]) = A \tag{2}$$

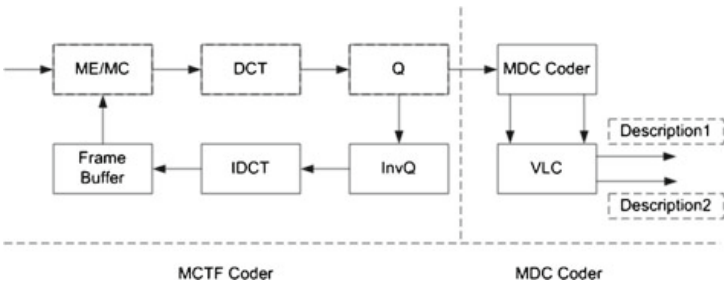

Fig. 2 Block-diagram of the proposed method

Hence in case of a damage or loss in one these descriptions, we should be able to reconstruct the original value partially as expressed in (3).

$$\begin{aligned} \tau^{-1}([B_1, null]) &= A' \text{ where } \quad |A - A'| < \epsilon \\ \tau^{-1}([null, B_2]) &= A'' \text{ where } \quad |A - A''| < \epsilon \end{aligned} \tag{3}$$

The error threshold value ϵ is determined by a tradeoff between efficiency and accuracy as described below. The proposed transform creates a base layer and an enhancement layer parts for each description. The base layer is repeated in both descriptions and hence introduces a redundancy to the coding. Each description D_i therefore can be given as:

$$D_i = a \bigoplus b_i \qquad for \quad i = 1, 2$$

where a is the base layer, b is the enhancement layer, and $\bigoplus$ is the operation of combining data from these layers. It should be noted that the reconstruction error rate in presence of damage or loss in one of the descriptions depends on the amount of information present in the enhancement layers. Hence a smaller enhancement layer tends to increase the accuracy. On the other hand, a smaller enhancement layer results in a large base layer which will increase the data redundancy. We have considered the following metrics in designing our transform:

- To minimize the redundancy, the base layer size should be minimized,
- Reconstruction error using the base layer only, should be minimized,
- The enhancement layer data size should be a function of the transmitted value.

The last item in the list above is the result of the observation that most of the quantized values are small numbers. Since enhancement layer data in split between the descriptions, reconstruction with one description only results in a large error when the base layer is small. Hence, we prefer an adaptive enhancement layer which grows with increasing data values. The above-mentioned metrics can be expressed mathematically as shown in (4) and (5):

$$Min(A - \tau^{-1}(\tau_B(A)) + \tau_B(A)) \tag{4}$$

$$Min(A - \tau^{-1}(\tau_B(A) \bigoplus \tau_{Ei}(A))) \qquad for \quad i = 1, 2 \tag{5}$$

where $\tau(.)$ is the intended transform, $\tau_B(.)$ is the base layer after the transform, $\tau_{Ei}(.)$ is the enhancement layer $\boldsymbol{i}$ after applying the transform, and $\tau^{-1}(.)$ is the inverse transform. Figures 3 and 4 depict the reconstruction error using the base layer only, and the reconstruction error using one description only, for an inverse quadratic and logarithmic functions respectively. We have used the proposed method with inverse-quadratic and logarithmic functions as transforms. Then we reconstructed the encoded value using different cases when one or both descriptions are received. The figures serve to verify the effectiveness of the proposed method in terms of the generated error.

In designing the transform we considered the issue of minimizing the reconstruction error for all cases of reconstruction using one description only, reconstruction using base layer only. The last case arises when the channels used for transmission of the descriptions suffer from the limited bandwidth problem and a down-scaled

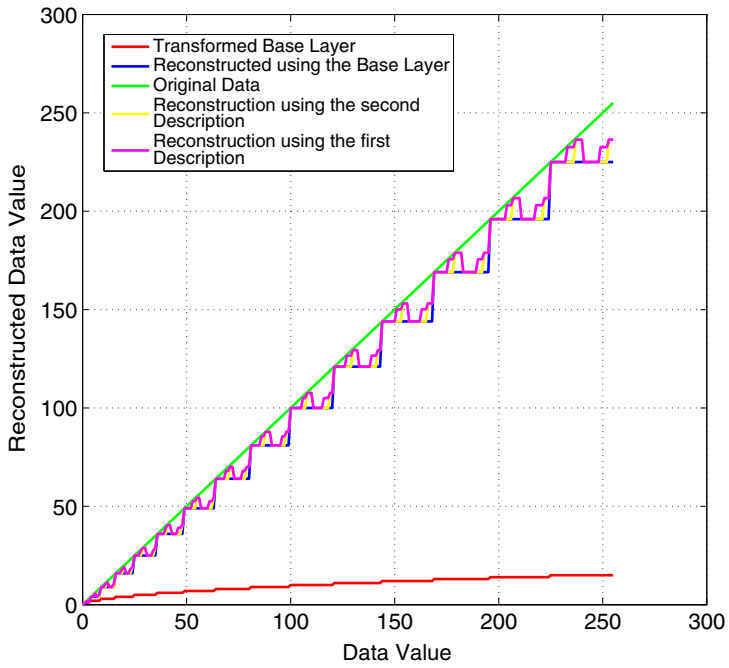

Fig. 3 Reconstruction error with inverse-quadratic transform

stream is received through each channel. The optimum solution considering the above-mentioned criteria is an inverse quadratic transform as given in (6):

$$Base = Trunc(\sqrt{Coef}) \qquad Enhancement = \sqrt{Coef} - Base \tag{6}$$

where Coef is the quantized DCT coefficients of the macro-blocks Fig. 2. The fraction part after applying the transform is divided into two parts and used as enhancements to the base layer data. The enhancement layer bits are coded separately. This feature provides the multi-layer scalability characteristic for each description. The

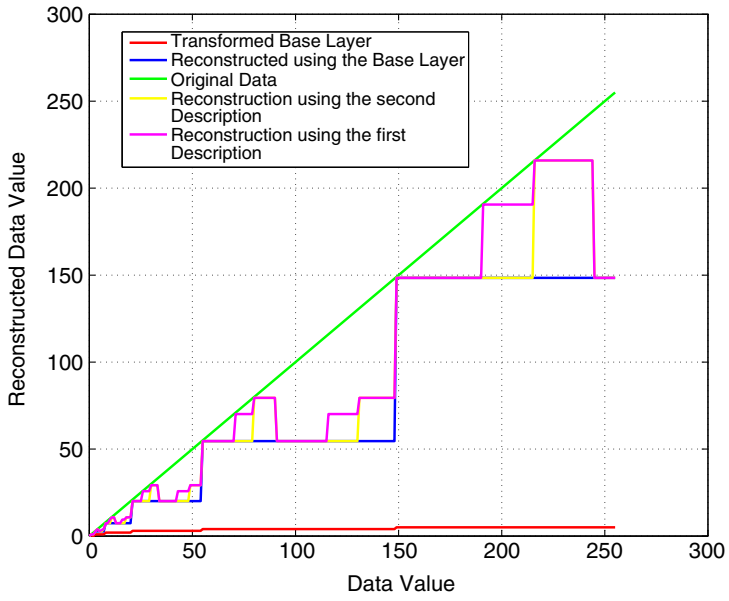

Fig. 4 Reconstruction error with logarithmic transform

descriptions go through entropy coding later on, so that each layer present in the descriptions is entropy coded separately.

In the following discussions we labeled the descriptions as D1 and D2. The fraction bits at position 2^{-1} and 2^{-4} are packed and entropy coded at the enhancement layers of D1 and the fraction bits at position 2^{-2} and 2^{-3} are packed and entropy coded at the enhancement layers of D2, respectively. In this way, we tried to balance the bit rate and accuracy of both descriptions. Algorithm 1 describes the reconstruction of the video when both descriptions are received or in case of failure in one of the descriptions. The proposed method provides the possibility of ignoring one or both enhancement layers data in each description in case of communication bandwidth restrictions. This scalability feature when the reconstruction is carried out using the base layer only, has not been considered in Algorithm 1.

Algorithm 1 Reconstructing video

if $BaseD1 = NULL$ **then**
  {If Description 1 is lost}
  $fraction \leftarrow EnhanceD2_1 \times (2^{-2}) + EnhanceD2_2 \times (2^{-3})$
  $Coef \leftarrow Round((BaseD2 + fraction)^2)$
else if $BaseD2 = NULL$ **then**
  {If Description 2 is lost}
  $fraction \leftarrow EnhanceD1_1 \times (2^{-1}) + EnhanceD1_2 \times (2^{-4})$
  $Coef \leftarrow Round((BaseD1 + fraction)^2)$
else
  {Both Descriptions are received}
  $fraction_1 \leftarrow EnhanceD1_1 \times (2^{-1}) + EnhanceD1_2 \times (2^{-4})$
  $fraction_2 \leftarrow EnhanceD2_1 \times (2^{-2}) + EnhanceD2_2 \times (2^{-3})$
  $fraction \leftarrow fraction_1 + fraction_2$
  $Coef \leftarrow Round((BaseD2 + fraction)^2)$
end if

4 Experimental results

For evaluating the performance of our proposed method, we have considered measuring Peak Signal to Noise Ratio of the Y component of YcbCr color space from the macro-blocks (Y-PSNR). Equations (7) and (8) describe PSNR used in our implementation mathematically.

$$PSNR = 20 \log_{10} \frac{\mathrm{Max}_I}{\sqrt{MSE}} \tag{7}$$

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} ||I(i, j) - I'(i, j)||^2 \tag{8}$$

where MSE is the mean square error, Max_I indicates the largest possible pixel value, I is the original frame, I' is the decoded frame at the receiver side, and m and n are number of rows and columns respectively. Y-PSNR is applied to all frames of video segments listed in Table 1 by comparing the corresponding frames of the original video segment and retrieved video using one or both descriptions from our

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

Sequence name	Resolution	Frame rate	Avg. Y-PSNR using D1 & D2 (dB)	Avg. Y-PSNR using D1 (dB)	Avg. Y-PSNR using D2 (dB)
Foreman	352 × 288	30	36.474	34.192	33.966
Stefan & Martin	768 × 576	30	34.078	32.472	32.105
City	704 × 576	60	34.643	32.671	31.978

proposed coding method. We place 32 frames in each GOP and a diadic hierarchical temporal structure has been used for motion compensated coding. Furthermore, we have imposed the same reference frame for all macro-blocks of a frame for simplicity although H.264 supports utilizing different reference frame for macro-blocks of a frame. As the proposed method has both error resilience characteristic through implementing multiple description coding, and scalable video coding, we have considered the following test scenarios.

– Measuring redundancy imposed by error resilience of MDC,
– Performance measurement when only the base layer is received,

Fig. 5 Retrieved frames using proposed method: *Upper-left* original frame, *upper-right* retrieved using both descriptions, *lower-left* retrieved using first description, *lower-right* retrieved using second description

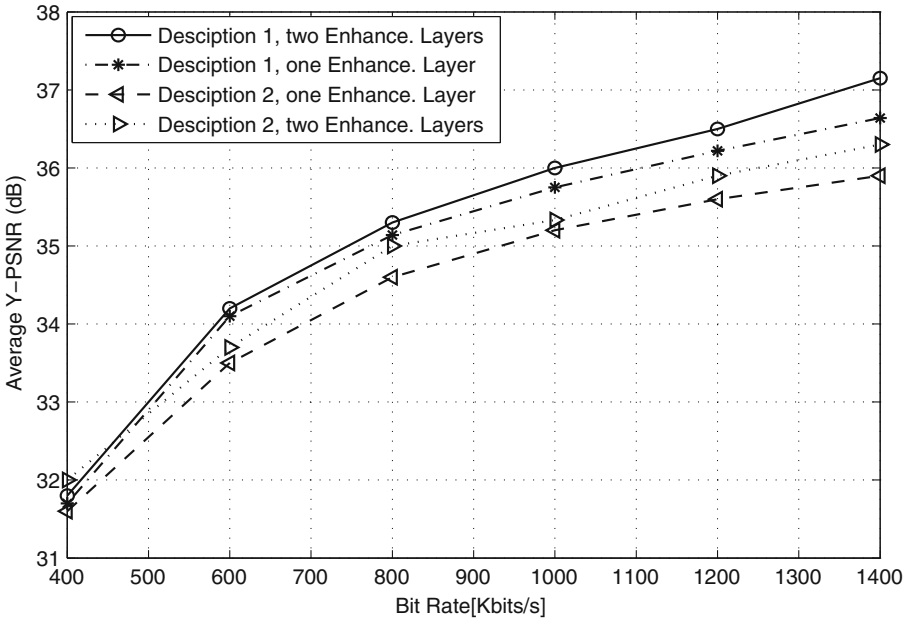

Fig. 6 Comparison of the rate distortion when one description is received

– Performance measurement when only one enhancement layer from each description is received,
– Performance measurement when only one description is received,
– Performance measurement when one description with one enhancement layer is received.

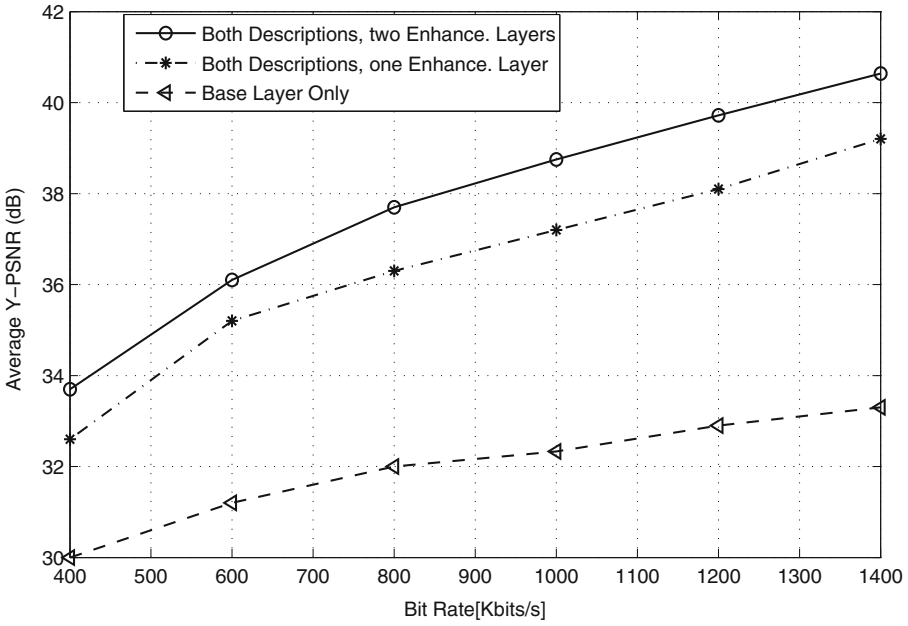

Fig. 7 Comparison of the rate distortion when both descriptions are received

Table 2 Redundancy added by the proposed method

Sequence name	Target bit rate (Kbits/s)	Bit rate using the proposed method (Kbits/s)	Redundancy percentage (%)
Foreman	1,000	1,292	29.2
Stefan & Martin	1,000	1,372	37.2
City	1,000	1,341	34.1

Figure 5 depicts a frame of the first test video and its corresponding reconstructions using one description, and using both descriptions. The visual inspection of the retrieved frames also indicates that the proposed method provides acceptable results even in presence of transmission error. The redundancy caused by repeating base layer information in both descriptions is partially compensated by organizing the data bits as mentioned in Section 3. To optimize the distortion with respect to the bit rate, the enhancement layer in each description has been entropy coded separately. The optimization is related to the observation that a large number of coefficients after quantization are small integers. Hence, entropy coding encodes the base layer more efficiently after applying the transform. Since the redundancy is arising from the repetition of the base layer in both descriptions, the total performance improves. Meanwhile, this feature provides the flexibility of having scalability at each description.

In our testing scenario, we have considered transmission over a packet loss network. The bitstreams of the two descriptions are separated in packets of maximal size of 1,500 bytes for compatibility with the maximum frame size of Ethernet. For each description, separate packets are created. If the packet is lost, we consider that the corresponding description is not available for reconstructing the block and hence, the block is reconstructed using the other description. In the second scenario we assume that as a result of bandwidth fluctuations, the receiver can receive the data in a description partially. This means that the enhancement layer of the data in a description is dropped. Figure 6 compares the rate distortion of the proposed method when one or two enhancement layers are received within a single description. With

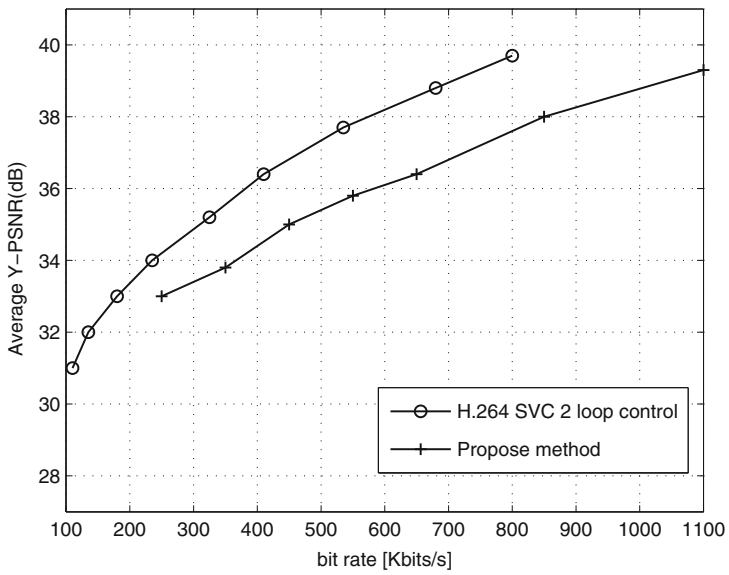

Fig. 8 Comparison of H.264 performance with the proposed method when both descriptions are received

Table 3 Performance comparison of the proposed method with MDTC in terms of Y-PSNR (dB)

Proposed method	MDTC	Loss (%)
29.11	29.03	10
27.87	28.53	20

only one description received, the video quality is still acceptable with an average PSNR reduction of less than 2 dB. Figure 7 depicts the case when both descriptions are received in down-scaled form. The extreme case of receiving base layer only is computed by considering a duplication in data. Table 2 shows the average bit rates and the related data inflation percentage due to the redundancy added by the proposed method. The higher rate of redundancy in Stefan and City sequences can be related to their higher spatial detail and amount of movements.

A comparison with SNR scalability of H.264 standard has been given in Figure 8. The video sequence 'City' has been used for comparison in CIF spatial resolution, at a temporal rate of 15 fps, and 16 frames in each GOP. The coarse grain quality scalable mode with three layers has been utilized. The multi-layer structure of H.264 encoder allows it to minimize redundancy which is present in our proposed method which is optimized for a noisy channel. Meanwhile, it should be noted that the results presented in Figure 8 for H.264 standard are obtained from a single description while our proposed method is tested with two descriptions which imposes a redundancy of 34.1% in 'City' video sequence. We have also compared our method with the multiple description transform method (MDTC) method proposed in [18] which compresses the video using SNR scalability, duplicates the base layer so that it appears in both descriptions, and alternates blocks (i.e., GOBs) of the enhancement layer between the two descriptions and hence has a similarity with our proposed method. For comparison we use 'Foreman' QCIF video sequence with 144 Kbps and 7.5 fps. The comparison is for 10% and 20% of frame losses. The results of the comparison are given in Table 3. Despite having almost similar PSNR performance, it is worth noting that in 144 Kbps the redundancy imposed by our proposed method is 41.2% whereas the redundancy rate is 45% in the method proposed in [18].

5 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 combined with signal to noise ratio (SNR) scalable video coding and hence it has the capability of being used as a scalable coding method where any data loss or corruption is reflected as reduction in the quality of the video. The multi-layer structure of data in each description provides the feasibility of reducing data rate by scaling down the video whenever the connection suffers from a low bandwidth problem. In order to measure the performance of the proposed coding method, distortion rate imposed by data loss and scaling down for rate efficiency, have been utilized. Except for the case when all descriptions are lost, the video streams do not experience a major quality loss at play back. Utilizing the motion compensated temporal filtering structure of video coding standards, we managed to preserve the compatibility of the proposed method with major standards such as H.264. Our

proposed method is based on SNR scalability of video coding standards, however, a reasonable extension of the work is going to be its combination with temporal and spatial scalabilities.

References

1. Abanoz TB, Tekalp AM (2009) SVC-based scalable multiple description video coding and optimization of encoding configuration. Signal Process Image Commun 24(9):691–701
2. Akyol E, Tekalp AM, Civanlar MR (2007) A flexible multiple description coding framework for adaptive peer-to-peer video streaming. IEEE Select Topics Signal Proc 1(2):231–245
3. Andreopoulos Y, van der Schaar M, Munteanu A, Barbarien J, Schelkens P, Cornelis J (2003) 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
4. Ardestani MR, Shirazi AAB, Hashemi MR (2011) Low-complexity unbalanced multiple description coding based on balanced clusters for adaptive peer-to-peer video streaming. Signal Process Image Commun 26(3):143–161
5. Chang SK, Sang L (2001) Multiple description coding of motion fields for robust video transmission. IEEE Trans Circuits Syst Video Technol 11(9):999–1010
6. Choupani R, Wong JSM, Tolun MR (2009) Multiple description scalable coding for video transmission over unreliable networks. In: Embedded computer systems: architectures, modeling, and simulation. 9th international workshop, SAMOS 2009. Samos, Greece, pp 58–67
7. Choupani R, Wong JSM, Tolun MR (2011) Scalable video transmission over unreliable networks using multiple description wavelet coding. In: The 7th International Conference on Digital Content, Multimedia Technology and its Application (IDCTA2011). Busan, Korea, pp 5–10
8. Conklin G, Greenbaum G, Lillevold K, Lippman A, Reznik Y (2001) Video coding for streaming media delivery on the internet. IEEE Trans Circuits Syst Video Technol 11:269–281
9. Folli M, Favalli L (2008) Scalable multiple description coding of video sequences. GTTI'08—Sessione elaborazione dei segnali
10. Franchi N, Fumagalli M, Lancini R, Tubaro S (2003) A space domain approach for multiple description video coding. In: ICIP 2003, vol 2, 253–256
11. Girod B (1987) The efficiency of motion-compensated prediction for hybrid video coding of video sequences. IEEE J Sel Areas Commun 5(7):1140–1154
12. Goyal VK (2001) Multiple description coding: compression meets the network. IEEE Signal Process Mag 18(5):74–93
13. López-Fuentes FA (2011) P2P video streaming combining SVC and MDC. Int J Appl Math Comput Sci 21(2):295–306
14. Mohr AE, Riskin EA, Ladner RE (2000) Unequal loss protection: graceful degradation of image quality over packet erasure channels through forward error correction. IEEE J Sel Areas Commun 18(6):819–828
15. Ohm J (2005) Advances in scalable video coding. Proc IEEE 93(1):42–56
16. Puri R, Ramchandran K (1999) Multiple description source coding using forward error correction codes. IEEE Conf Signal Syst Comp 1:342–346
17. Reguant VD, Prats FE, De Pozuelo RM, Margalef FP, Ublergo GF (2008) Delivery of H264 SVC/MDC streams over Wimax and DVB-T networks. In: Proc. Int. Symp. Consumer Electronics, ISCE, pp 1–4
18. Reibman AR, Jafarkhani H, Wang Y, Orchard MT, Puri R (2002) Multiple-description video coding using motion-compensated temporal prediction. IEEE Trans Circuits Syst Video Technol 12(3):193–204
19. Schwarz H, Marpe D, Wiegand T (2007) Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans Circuits Syst Video Technol 17(9):1103–1120
20. Tillo T, Olmo G (2004) A novel multiple description coding scheme compatible with the JPEG2000 decoder. IEEE Signal Process Lett 11(11):908–911

21. Tillo T, Grangetto M, Olmo G (2007) Multiple description image coding based on Lagrangian rate allocation. IEEE Trans Image Process 16(3):673–683
22. Venkataramani R, Kramer G, Goyal VK (2003) Multiple description coding with many channels. IEEE Trans Inf Theory 49(9):2106–2114
23. Wang Y, Lin S (2002) Error-resilient video coding using multiple description motion compensation. IEEE Trans Circuits Syst Video Technol 12(6):438–452
24. Wang Y, Orchard MT, Vaishampayan V, Reibman AR (2001) Multiple description coding using pairwise correlating transforms. IEEE Trans Image Process 10:351–366
25. Wang Y, Reibman AR, Shunan L (2005) Multiple description coding for video delivery. Proc IEEE 93(1):57–70
26. Wien M, Schwarz H, Oelbaum T (2007) Performance Analysis of SVC. IEEE Trans Circuits Syst Video Technol 17(9):1194–1203
27. Wu D, Hou Y, Zhu W, Zhang YQ, Peha J (2001) Streaming video over the internet: approaches and directions. IEEE Trans Circuits Syst Video Technol 11(3):282–300
28. Xiang W, Zhu C, Siew CK, Xu Y, Liu M (2009) Forward error correction-based 2-D layered multiple description coding for error-resilient H.264 SVC video transmission. IEEE Trans Circuits Syst Video Technol 19(12):1730–1738
29. Xuguang Y, Ramchandran K (2000) Optimal subband filter banks for multiple description coding. IEEE Trans Inf Theory 46(7):2477–2490

Roya Choupani received her BSc. and MSc. degrees in computer engineering in 1998 and 2002 respectively. Since 2006 she has started her PhD study at the Computer Engineering department of Delft University of Technology, in Delft—the Netherlands. Her research interests are video coding, data communications and networking.

Stephan Wong was born in Paramaribo, Suriname on October 20th, 1973. He obtained his PhD from the Delft University of Technology in December 2002 after which I started as an assistant professor at the same university. His PhD thesis entitled "Microcoded Reconfigurable Embedded Processor" describes the MOLEN polymorphic processor, organization, and (micro-)architecture. His research interests include: Reconfigurable Computing, Distributed Collaborative Computing, High-Performance Computing, Embedded Systems, Hardware/Software Co-Design, Network Processing.

Mehmet Tolun main research interests are: Artificial Intelligence and Expert Systems, Machine Learning Algorithms and Symbolic Inductive Learning. His Inductive Learning Algorithm-ILA-was the subject of many MS theses and two Ph.D. dissertations. Mehmet Tolun was past Dean of the EMU Faculty of Engineering for nearly 4 years. During his term of office two of the engineering programs at EMU have been applied for ABET "substantial equivalency". He contributed to IEEE as the IEEE Turkey Section Chair, Computer Society Chapter Chair, and GOLD Affinity Group Chair. He is the recipient of the IEEE Third Millennium Medal. He is currently a Senior Member of the IEEE.

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

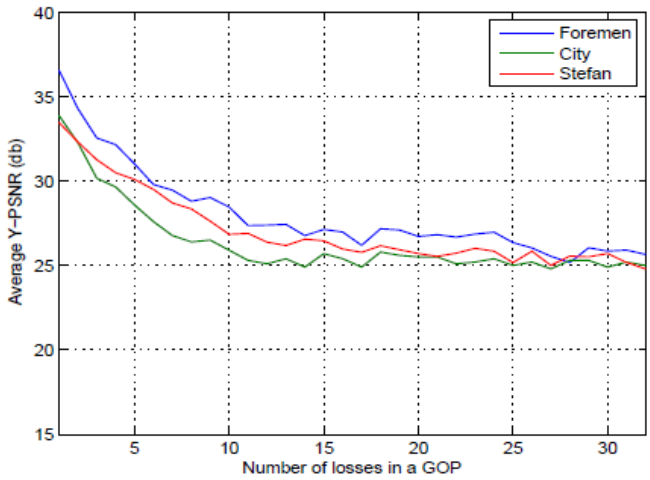

Fig. 3. Multiple Description Schemes with a) 9 Descriptions, b) 16 Descriptions

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.

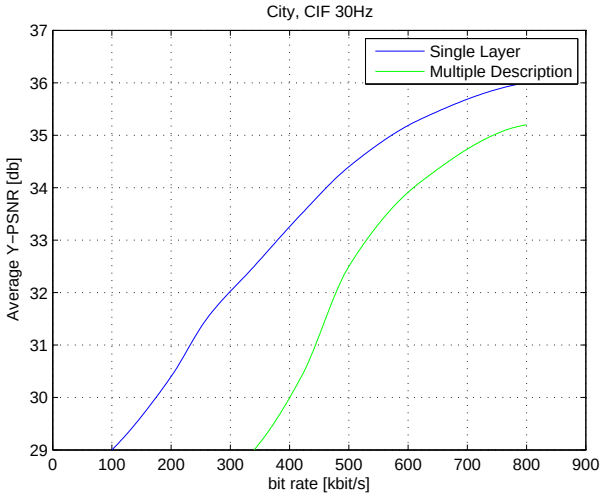

Fig. 4. Coding efficiency comparison between single layer and our proposed method using City video segment

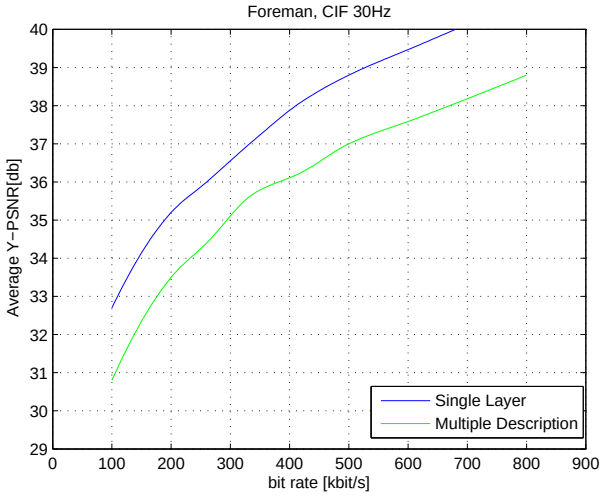

Fig. 5. Coding efficiency comparison between single layer and our proposed method using Foreman video segment

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.

Optimized Multiple Description Coding for Temporal Video Scalability

Roya Choupani[1,2], Stephan Wong[2], Mehmet Tolun[3]

[1] Computer Engineering Department, Cankaya University, Ankara-Turkey
roya@cankaya.edu.tr
[2] Computer Engineering Department, Delft University of Technology, Delft-the Netherlands
{rchoupani , j.s.s.m.wong }@tudelft.nl
[3] Computer Engineering Department, TED University, Ankara-Turkey
mehmet.tolun@tedu.edu.tr

Abstract

The vast application of video streaming over the Internet requires video adaptation to the fluctuations of the available bandwidth, and the rendering capabilities of the receiver device. On the other hand, the available video coding standards are designed for optimum bit rate which makes them susceptible to packet losses. A combination of video adaptation methods and error resilient methods can make the video stream more robust against networking problems. In this paper, an optimization for combining scalable video coding with multiple description coding schemes have been proposed. Our proposed method is capable of creating balanced descriptions with optimum coding efficiency.

Keywords: Scalable Video Coding, Multiple Description Coding, Video Coding

1 Introduction

Human's perception of his surrounding world is essentially dependent on visual information. This dependency has reached a higher level with the progress in advanced technologies, particularly in communications networks which makes it possible for video to be widely utilized in our daily life. Video as a sequence of frames, however, involves a huge amount of data. Hence, the storage and communicating video requires very large capacities which make video compression a necessity. However, the variations in the physical characteristics of the communication networks and the rendering capabilities of the receiver display device require adaptations to be made to the compressed video. Meanwhile, these adaptations should be fast to be applicable in real-time video streaming while preserving the quality of the video as much as possible. Adaptability of video to the transmission bandwidth or displaying capabilities of the recipient device is the objective of scalable video coding (SVC) methods. This adaptability however, does not require long processing and is performed by utilizing only some parts of the video data and simply ignoring the remaining parts in a flexible way. Meanwhile, this adaptability can not only handle the bandwidth fluctuation of the communication channels but also enables video rendering on older devices by allowing them to utilize the bit-stream partially to display the video in lower quality. The flexibility however, comes with the cost of sacrificing coding efficiency to

some extent. SVC however is not capable of handling the packet loss problems because almost all video coding standards are based on eliminating temporal redundancy by encoding the differences between consecutive frames instead of the frame itself. This scheme can reduce the coded video size in a very high rate however, it creates a dependency chain between the frames. A frame cannot be decoded if its previous (reference) frames is not available. This characteristic requires utilization of error concealment methods such as multiple description coding. In this paper we introduce an optimized method for decomposition of a video into multiple descriptions. Our proposed method has scalability and error resilience properties. In the following sections we introduce the basic concepts of scalable coding of video and multiple description coding. Then we describe our proposed method details followed by experimental evaluation results.

2 Scalable Video Coding

In SVC methods, the video stream is represented by a main bit-stream which consists of several sub-streams. Each sub-stream represents video in a lower spatial resolution, lower temporal resolution, or lower bit-per-pixel quality [2]. The reconstructed video by using all sub-streams is in its highest quality. In order to reconstruct the video in lower spatial, resolution, or bit-per-pixel quality, some sub-streams from the main bit-stream are left out. To adapt the data size to the changes in the bit rate of the communication channel, a unit in a video stream such as a frame or a macro-block, is divided into a set of smaller parts. A measure of the number of items comprising a unit is called its granularity [25]. The first item of this set contains the basic and coarsest part of the data and the remaining items contain refinements to the basic item [24],[23]. The scheme of gradual refining of a unit or increasing the granularity of a unit is called Fine Granularity Scalability (FGS) [23], [22]. It is clear from the definitions 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 communication channel. The FGS scheme defines the video content in a multi-layered format [21], [20]. A higher quality for a video is achieved through increasing the number of layers decoded at the receiver side. This scheme leads to placing the layers comprising a video in an ordered sequence where the base layer is always at the first position. The base layer contains the minimum data required while remaining layers include refinements to the data carried by the base layer. This makes scalability possible, as a receiver can receive some of these layers and ignore the rest depending on its current bit rate capacity. Scalability in video is achievable through signal-to-noise ratio (SNR), spatial, and temporal changes. Bit-per-pixel or signal-to-noise ratio scalability is a technique to decompose a video sequence into two layers at the same frame rate and the same spatial resolution, but different quantization accuracy. The decomposition can be performed in pixel domain by putting more significant bits in the base layer and less significant bits in the enhancement layers. The decomposition can also be performed in the DCT domain. In this case the low frequency coefficients of the DCT are put in the base layer and the high frequency coefficients are put in the enhancement layer(s) [19], [18]. Spatial scalability is a technique to code a video sequence into multi-layers at the same frame rate, but different spatial resolutions [1]. The first layer (the base layer) is coded at the lowest spatial resolution. The base layer is created by down-sampling the frames. The difference between up-sampled base layer and the original frame is coded as the enhancement layer. In case that the video is coded in more layers, this procedure is repeated on the base layer yielding a new

base layer in lower resolution, and an enhancement layer [17], [16]. This strategy of creating multi-layer spatial SVC, makes layer k dependent on all layer from 1 up to k-1. An important consideration for coding efficiency is motion compensation in each layer. Two strategies are followed for motion compensation. Temporal scalability is a technique to code a video sequence into two layers at the same spatial resolution, but different frame rates [15], [14]. The base layer is coded at a lower frame rate. The enhancement layer provides the missing frames to form a video with a higher frame rate. Coding efficiency of temporal scalable coding is high and very close to non-scalable coding [14]. Figure 1 depicts the structure of temporal scalability with two layers.

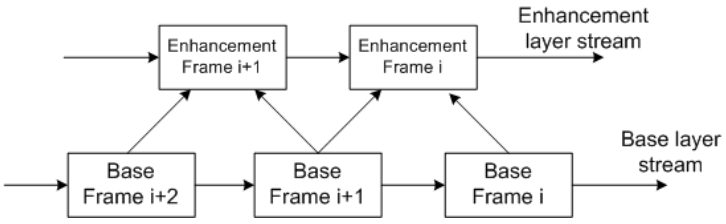

Fig. 1. Typical Structure of a Temporal Scalability Decoder.

Considering the two layer structure depicted in Figure 1, the enhancement frame i is the successor of the base layer frame i in the original sequence. Enhancement frame i-1, base frames i or i+1 can be used as a reference frame for enhancement layer frame i. Therefore, complying with the restrictions in video coding standards before H.264, only P-type predicted frames are used in the base layer [3]. The enhancement layer predicted frames can be either P-type, or B-type referencing a P-type frame from the base layer or the enhancement layer. Motion compensation in the based layer utilizes only the base layer information so no drift error is expected here [4]. However, with moving some of the frames to enhancement layer(s), the distance between consecutive frames in the base layer is increased. This increase can cause a slight decrease in the coding efficiency.

3 Multiple Description Coding

A multiple description coder (MDC) for video coding divides the video data into some bit-streams called descriptions which are then transmitted separately over different network channels [13]. Generally, descriptions have the same importance and data rates, even though this is not a necessary requirement. Each description can be decoded independently from other descriptions. This means that the loss of some of these descriptions does not affect the decoding of the rest [12]. The accuracy of the decoded video depends on the number of received descriptions [9]. Figure 2 depicts the basic framework for a multiple description encoder/decoder with two descriptions.

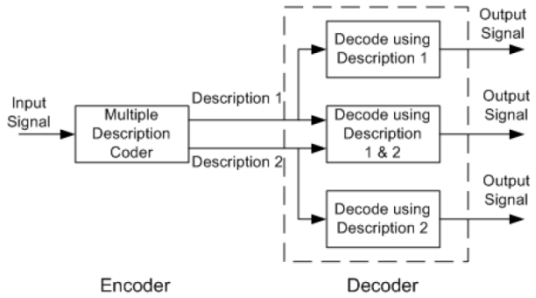

Fig. 2. Multiple descriptions coding block-diagram.

In case of a failure in one of the channels, the output signal is reconstructed from the other description. In contrast to descriptions, in a multi-layer coded video, layer i cannot be decoded if layer i-1 is not present [7]. This means that in order to decode a multi layer video using m layers out of a total of n layers, the available layers should be the lowest layers. However, the descriptions utilized for decoding a video are not necessarily from any order since the main goal of MDC is delivering video (although in a lower quality) when parts of video data are lost [8]. In order to reconstruct video in presence of data loss or corruption, redundancy should be added to the bit-stream. This redundancy is in the form of repeated bits (duplicated blocks), or inefficiency in the encoder when the bit-stream is encoded by a rate below channel capacity. When a frame or a block of a frame is missing, the decoder estimates it by utilizing its adjacent data that was received correctly. The adjacency can be in the spatial or temporal domain. Recovering data completely or partially when some parts of data is lost and masking the data loss effect is called error concealment. MDC schemes are among the techniques that are commonly utilized for error concealment. Even if the descriptions are designed as non-overlapping sets, or partitions, it does not necessarily mean that there is no redundancy in the data. Given that each partition is encoded independently from other partitions, the spatial or temporal correlation between the data in different partitions is not utilized and hence the redundancy is not eliminated [6]. On the other hand, the preserved spatial or temporal correlation can be used for estimating the lost data for error concealment [5]. This helps to create a scalable video resilient to packet losses [10][11].

4 Optimizing Temporal MDC

Decomposing a video into several descriptions by putting the frames in different descriptions is the main idea utilized in temporal scalability with multiple descriptions. For instance, using two descriptions, the odd numbered frames are put in the first description while the even numbered frames are assigned to the second description. The main drawback of this scheme is that in order to make descriptions independent from each other, a frame and its reference frame should be in the same description. Hence, the reference frame may not be necessarily the most similar frame to the current frame, and the most similar frame may have been assigned to the other description. This drawback reduces the coding efficiency because the temporal redundancy is not completely eliminated. Meanwhile, the decomposition of a video sequence into multiple descriptions should create balanced descriptions. This requirement is based on the assumption that the transmission networks used for delivering descriptions can be subject to bandwidth fluctuations and data loss. In presence of data loss, the video is

reconstructed using the delivered descriptions. Hence, to minimize the video quality degradation in all cases, the required condition is the dependency of the reconstructed video quality on the number of delivered descriptions regardless of which description is lost. In this section we present an optimization to solve this problem.

The proposed method presented here assumes only two descriptions (D1 and D2) however, it is readily extendable to more number of descriptions. Meanwhile, in our proposed method, we have assumed that each frame can have only one reference frame. Assuming that a GOP includes n frames, the proposed method starts by encoding frames F_1 and F_2 by using intra-frame coding. These two frames are the first frames of each of the descriptions in our proposed method. This assumption can be relaxed by a few minor changes in our proposed method. The method starts by considering frames F_3 and F_4. Since these frames are encoded by using inter-frames coding, their differences with their reference frames are computed. The proposed method considers reference frames from both descriptions and finds the differences for F_3 and F_4. The differences are summed up for each frame as given in Equation 2.

$$Diff_{total} = \sum_{i\in\{blocks\}} MAD_{v,w}(B_i, B_{Ri}) \qquad (1)$$

$$MAD_{s,t}(F_1, F_2) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}|F_1(i,j) - F_2(i+s, j+t)| \qquad (2)$$

where MAD is mean absolute difference Equation 1, B_i and B_{Ri} are a block from the current frame and its most similar area from the reference frame respectively, and *v,w* indicates the amount of displacement by the current block to reach to its most similar area in the reference frame (motion vector). The proposed method assigns each frame to the descriptions with smaller total difference as computed in Equation 1. Figure 3 depicts the decomposition of a GOP with 16 frames into two description with optimized assignment of the frames to descriptions as proposed in our method.

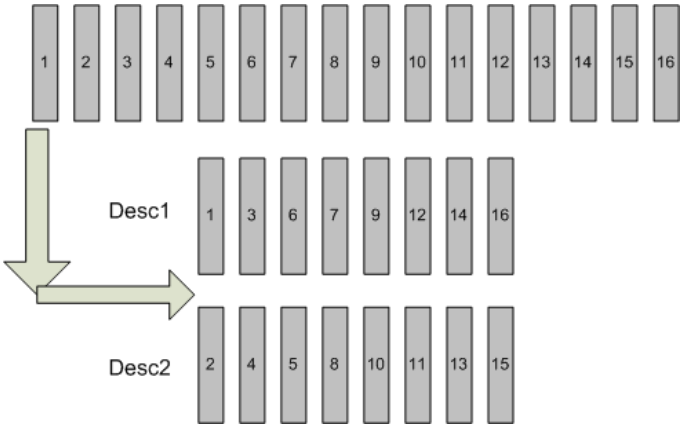

Fig. 3. A sample decomposition of a GOP into two descriptions in the proposed method.

As depicted in Figure 3 the worse case of having frames with consecutive sequence numbers puts only two adjacent frames in a description. Algorithm 1 defines how the assignment of the frames to the descriptions is carried out.

Algorithm 1. Assigning frames to descriptions in the proposed method

1. Encode the frames 1 and 2 of a GOP using intra-frame coding and assign them to description 1 and 2 respectively
2. While NOT end of the GOP DO
    a. Get next two frames (F_i and F_j)
    b. Find the motion compensated difference of each block of F_i and F_j in description 1 and 2 as Diff(i,1), Diff(i,2), Diff(j,1), Diff(j,2)
    c. IF Diff(i,1) + Diff(j,2) < Diff(i,2) + Diff(j,1)THEN
        Assign frame F_i to description 1 and F_j to description 2
    d. ELSE
        Assign frame F_i to description 2 and F_j to description 1

As shown in Algorithm 1, the proposed method preserves balance in the creation of the descriptions by grouping the frames of a GOP in pairs and assigning each frame of a pair to one of the descriptions. In case of extending the method to multiple descriptions, the grouping will be in n where n is the number of descriptions. In case that the size of GOP is not divisible by n, some descriptions will contain one frame less than the others (worst case).

5 Experimental Results

To evaluate the proposed method experimentally we have utilized the sequences 'Foreman', 'Stefan', and 'City'. The specifications of the sequences are given in Table 1.

Table 1. The Video Sequences Utilized in Experimental Evaluations

Sequence Name	Resolution	Frame Rate	Length (frames)
Foreman	352 × 288	30	300
Stefan	768 × 576	30	300
City	704 × 576	60	600

The comparison is considered to measure the effect of the proposed optimization. Therefore, as a benchmark, the frames of a GOP of 32 frames are decomposed as odd and even number frames into two descriptions. Subsequently, the same frames are decomposed into descriptions with the optimization proposed in our method. The coding efficiency in terms of bit-per-pixel is computed for each sequence for with optimization and without optimization cases as provided in Table 2.

Table 2. Performance Comparison of the Proposed Method with Odd-Even Decomposition of Video into Descriptions.

Sequence Name	Optimized using the Proposed Method (bpp)	No Optimization (bpp)
Foreman	1.181	1.2212
Stefan	1.237	1.291
City	1.062	1.076

As it is shown in Table 2, in all three sequences the bit-per-pixel values have been improved, although the improvements does not result in a major reduction in bit-per-pixel rates. Besides to the increased coding efficiency, the proposed method preserves the balance in the descriptions in terms of the number of frames, by decomposing the GOPs evenly, and avoids creating large timing gaps between adjacent frames in a description.

Our second experiment compares the coding efficiency of the odd-even decomposition of the frames with the coding efficiency the proposed method at different bit rates. Figure 4 depicts the results of our comparison using Foreman video sequence. The comparison indicates that the impact of the proposed method is higher in high bit rates. The closeness of the results in low bit rates is the result of the fact that in low bit rates much of the differences between the frames which correspond to high frequencies are eliminated. It is also important to note that the main goal of combining MDC with SVC is avoiding jitter in video transmission when a sudden bandwidth fluctuation occurs at high bit rates. Hence, the proposed method suitability for these types of applications are verified once more.

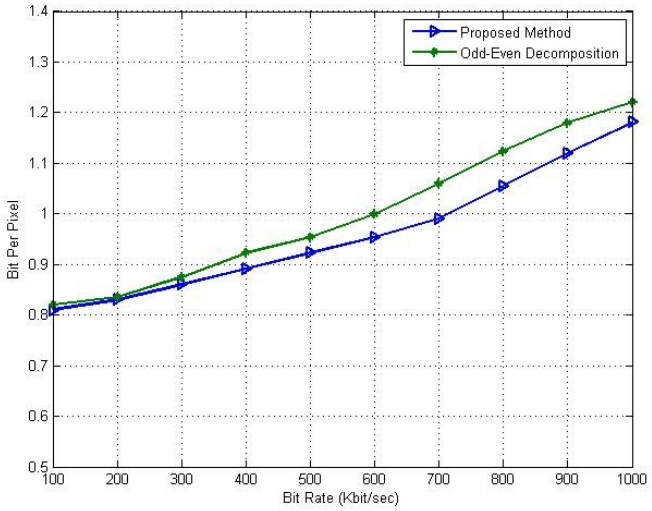

Fig. 4. Odd-Even Temporal Decomposition vis-à-vis Proposed Method Decomposition

6 Conclusion

A new method for handling the data loss during the transmission of video streams has been proposed. Our proposed methods are based on combining SVC with MDC where the video is decomposed into temporal sub-streams. In our proposed method, the error resilience of the video is increased. The proposed method has the capability of being used as scalable coding methods in which any data loss or corruption is reflected as reduction in the quality of the video. However, except for the case when all descriptions are lost, the video streams do not experience jitter at play back. In our method, an improvement is proposed for the well-known method of temporal decomposition of video into multiple descriptions by putting odd and even numbered frames in different descriptions. This method improves the coding efficiency of the odd-even temporal decomposition method by grouping the frames in pairs and assigned to the descriptions so that the differences with the reference frames are minimized. The

proposed method preserves the balance in the descriptions and avoids creating large timing gaps between adjacent frames in a description.

References

1. A. Segall and G. J. Sullivan, “Spatial scalability”, IEEE Transaction on Circuits Systems for Video Technology, 17(9):1121–1135, 2007.
2. H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the h.264/avc standard”, IEEE Transaction on Circuits and Systems for Video, 17(9):1103–1120, 2007.
3. C. Hewage, H. Karim, S. Worrall, S. Dogan, and A. Kondoz, “Comparison of stereo video coding support in mpeg-4 mac, h.264/avc and h.264/svc”, Proceeding of the 4th Visual Information Engineering Conference, pages 25–27, 2007.
4. R. Choupani, S. Wong, and M. Tolun, “A drift-reduced hierarchical wavelet coding scheme for scalable video transmissions”, First International Conference on Advances in Multimedia (MMEDIA), pages 68–73, 2009.
5. R. Choupani, S. Wong, and M. Tolun, “Multiple description scalable coding for video transmission over unreliable networks”, Embedded Computer Systems: Architectures, Modeling, and Simulation, 9th International Workshop, Samos-Greece, pages 58–67, 2009.
6. N. Franchi, M. Fumagalli, R. Lancini, and S. Tubaro. A space domain approach for multiple description video coding. ICIP 2003, 2:253–256, 2003.
7. T. Tillo and G. Olmo, “A low complexity pre-post processing multiple description coding for video streaming”, IEEE International Conference on Information and Communication Technologies (ICTTA 2004), 2004.
8. E. Akyol, A. M. Tekalp, and M. R. Civanlar, “A flexible multiple description coding framework for adaptive peer-to-peer video streaming. IEEE Journal of Selected Topics in Signal Processing, 1:231–245, 2007.
9. V. K. Goyal, “Multiple description coding: Compression meets the network”, IEEE Signal Processing Magazine, 18:74–94, 2001.
10. R. Choupani, S. Wong, and M. Tolun, “Multiple description coding for SNR scalable video transmission over unreliable networks”, Multimedia Tools and Applications, (June 2012), doi: 10.1007/s11042-012-1150-9
11. R. Choupani, S. Wong, and M. Tolun, “Unbalanced multiple description wavelet coding for scalable video transmission”, Journal of Electronic Imaging, 21 (4), 043006 (October 04, 2012), doi:10.1117/1.JEI.21.4.043006
12. R. Venkataramani, G. Kramer, and V.K. Goyal, “Multiple description coding with many channels”, IEEE Transaction on Information Theory, 49:2106–2114, 2003.
13. Y. Wang, A. R. Reibman, and L. Shunan, “Multiple description coding for video delivery”, Proceedings of IEEE, 93:57–70, 2005.
14. H. Katata, N. Ito, and H. Kusao, “Temporal-scalable coding based on image content”, IEEE Transaction on Circuits and Systems for Video Technology, 7:52–59, 1997.

15. M. Domanski, A. Luczak, and S. Mackowiak, “Spatial-temporal scalability for mpeg video coding”, IEEE Transaction on Circuits and Systems for Video Technology, 10:1088–1093, 2000.
16. W. Tan and A. Zakhor, “Real-time internet video using error resilient scalable compression and tcp-friendly transport protocol”, IEEE Transaction on Multimedia, 1(2):172–186, 1999.
17. E. J. Delp, P. Salama, E. Asbun, M. Saenz, and K. Shen, “Rate scalable image and video compression techniques”, Proceedings of the 42nd Midwest Symposium on Circuits and Systems, pages 635–638, 1999.
18. R. Mathew and J. F. Arnold, “Layered coding using bit-stream decomposition with drift correction”, IEEE Transaction on Circuits and Systems for Video Technology, 7:882–891, 1997.
19. D. Wilson and M. Ghanbari, “Exploiting interlayer correlation of SNR scalable video”, IEEE Transaction on Circuits and Systems for Video Technology, 9:783–797, 1999.
20. H. Jiang, “Experiment on post-clip FGS enhancement”, ISO/IEC JTC1/SC29/WG11, MPEG00/M5826, 2000.
21. W. Li and Y. Chen, “Experiment result on fine granularity scalability”, ISO/IEC JTC1/SC29/WG11, MPEG99/M4473, 1999.
22. H. Gharavi and M. H. Partovi, “Multilevel video coding and distribution architectures for emerging broadband digital networks”, IEEE Transaction on Circuits and Systems for Video Technology, 6:459–469, 1996.
23. M. Ghanbari, “Two-layer coding of video signals for VBR networks”, IEEE Journal of Selected Areas Communications, 7:771–781, 1989.
24. A. Puri and T. Chen, “Multimedia Systems, Standards, and Networks”, Marcel Dekker, New York, 2000.
25. L. Weiping, “Overview of fine granularity scalability in mpeg-4 video standard”, IEEE Transaction on Circuits and Systems for Video Technology, 11(3):301–317, 2001.

Chapter 4

Using Discrete Wavelet Transform for Optimizing Multiple Description Video Coding

Multiple description coding decomposes video into multiple streams. A major problem with MDC methods is their low efficiency compared to the single streams of video. The reason for this low efficiency is after decomposition the correlation is not removed and hence, the redundancies are not eliminated. A transform such as DWT can precede this decomposition to improve the coding efficiency. DWT decomposes the data into multiple sub-bands and in multiple levels and hence, frequency contents are computed in different scales (Section 1.1.2). The DWT decomposed sub-band coefficients can be transmitted over different network channels by considering each sub-band data as a description in MDC coding. Since the DWT coefficients are uncorrelated the redundancies are eliminated to some extent. Two main drawbacks of applying MDC methods in DWT domain are as follows:

- The DWT separates the low frequency content of video frames from their high frequency content. In most DWT-based compression algorithms, this feature is utilized by preserving low frequency contents and eliminating high frequency contents. Besides, small coefficients are eliminated in most cases (independently from the frequency sub-band that they belong to). However, the DWT coefficients are uncorrelated and hence, estimating a missing group of coefficients (or a sub-band coefficients) using a delivered group (or sub-band) is difficult.

- The amount of data at each DWT sub-bands depends on the video content. Hence, decomposing the video based on the sub-bands will create non-uniform distribution of data over the descriptions. This problem will result in higher rate of packet losses in some sub-bands due to unbalanced load put on channels.

We address both problems in our proposed methods. We utilize the self-similarity of the DWT coefficients to solve the first problem. The DWT coefficients show similar patterns at different decomposition levels in the same sub-band. For instance, vertical edges cause large coefficients in horizontal sub-band in all frequencies because the edges are not sharp in most images. We have used this property of the DWT coefficients to estimate missing data when each sub-band data is transmitted over a different network channel. Our solution for the second problem is the dynamic adjustment of the descriptions with the channel rates. The proposed method assigns the description with the smallest data size to the channel with lowest data rate. Since both the description size and the channel rate may

change dynamically, our method considers the frequency contents of the frames, and the feedbacks coming from the receiver. The details of the proposed methods and their experimental evaluations have been presented in the following research articles:

- R. Choupani, S. Wong, M.R. Tolun, Using Wavelet Transform Self-Similarity for Effective Multiple Description Video Coding, 10th International Conference on Information, Communications and Signal Processing (ICICS 2015), Singapore, pp 122-127
- R. Choupani, S. Wong, M.R. Tolun, Unbalanced multiple description wavelet coding for scalable video transmission (October 2012), SPIE Journal of Electronic Imaging (JEI), volume 21, issue 4.

Using Wavelet Transform Self-Similarity for Effective Multiple Description Video Coding

Roya Choupani
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, the Netherlands
e-mail: r.choupani@tudelft.nl

Stephan Wong
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, the Netherlands
e-mail: J.S.S.M.Wong@tudelft.nl

Mehmet Tolun
Department of Electrical Engineering
Aksaray University
Aksaray, Turkey
e-mail: mehmet.tolun@aksaray.edu.tr

***Abstract*—Video streaming over unreliable networks requires preventive measures to avoid quality deterioration in the presence of packet losses. However, these measures result in redundancy in the transmitted data which is utilized to estimate the missing packets lost in the delivered portions. In this paper, we have used the self-similarity property if the discrete wavelet transform (DWT) to minimize the redundancy and improve the fidelity of the delivered video streams in presence of data loss. Our proposed method decomposes the video into multiple descriptions after applying the DWT. The descriptions are organized in such a way that when one of them is lost during transmission, it is estimated using the delivered portions by means of self-similarity between the DWT coefficients. In our experiments, we compare video reconstruction in the presence of data loss in one or two descriptions. Based on the experimental results, we have ascertained that our estimation method for missing coefficients by means of self-similarity is able to improve the video quality by 2.14dB and 7.26dB in case of one description and two descriptions, respectively. Moreover, our proposed method outperforms the state-of-the-art Forward Error Correction (FEC) method in case of higher bit-rates.**

***Index Terms*—Multiple Description Coding, Video Transmission Error, Discrete Wavelet Transform, Self-Similarity.**

I. Introduction

Multiple Description Coding (MDC) methods are utilized for improving the error robustness of data transmission over unreliable networks. MDC methods provide error robustness by decomposing a certain video into various descriptions and transmitting each description over preferably an independent network channel [18]. The descriptions should be encoded in such a way that each stream is decodable independently [16]. Moreover, each delivered description should improve the quality of the reconstructed video regardless of the location of the delivered description data in the original video [9]. This decomposition should be optimized in such a way that the quality of the reconstructed video is maximized in case of a loss of one or more descriptions, and simultaneously maintaining the optimal coding efficiency by minimizing the redundancy in descriptions. The possibility of reconstructing the video (although in lower quality) when some of the descriptions are lost is provided by including redundant data in descriptions. Besides, the coding efficiency is reduced due to the elimination of the correlation between data during the decomposition. Minimizing the redundancy and improving the coding efficiency on the other hand, deteriorates the quality of video and results in distortions when some of the description are not delivered. Hence, a tradeoff between the encoder performance in terms of bit-rate and the imposed distortion is sought by adjusting coding parameters according to the channel conditions. In the present work we address the problem of minimizing the inaccuracy of the reconstructed video in presence of data loss or corruption. Our approach to the problem is based on utilizing the correlation present in order to estimate/iterpolate the missing data. Our proposed method utilizes the self-similarity feature of the Discrete Wavelet Transform (DWT) to estimate the missing data. We have presented a review of related works, the details of our proposed method, and its experimental evaluation in the following sections.

II. Related Work

A significant number of video coding methods using MDC schemes have been reported in literature [4][15][3][1]. A comprehensive overview paper on MDC methods is presented in [16]. Improving the robustness of MDC methods against packet loss through data redundancy [1] or selective protection of descriptions [21][10] reduces the bit-rate performance of the encoder [20]. In [13] the authors propose an algorithm to control the mismatch between the prediction loops at the encoder and decoder in multiple description (MD) video coders with motion-compensated predictions. They consider three different cases; one in which both descriptions are received and other two when either of individual descriptions is received. In [1] the authors propose to generate multiple scalable descriptions from a single SVC bit-stream by mapping scalability layers of different frames to different descriptions. Their scheme is intended for P2P streaming over multiple multicast trees and features several encoding parameters, such as base layer rate of descriptions and overall redundancy. They aim to optimize the mean rate-distortion performance of each description received over a packet loss network, range of extraction points of the SVC stream, and overall redundancy of their MDC scheme.
DWT-based MDC methods are also utilized together with SVC [11][2][6][5][12]. 3D DWT with MCTF is used in MDC methods where they either directly perform MCTF on the input video sequence before the spatial transform, or in the

978-1-4673-7218-3/15/$31.00 ©2015 IEEE

wavelet subband domain which is often referred to as in-band MCTF. Decomposing the DWT coefficients into independent descriptions are generally based on the spatial oriented trees introduced in [14]. In [15] the decomposition is carried out by dividing the DWT coefficients at each level into blocks of equal sizes, and obtaining the descriptions by distributing the blocks among them. However, to create balanced descriptions, the authors encode each block in both low and high distortion rates. Each description then contains low distortion coded versions of some of these blocks, and high distortion versions of the rest. The redundancy added in this way makes the method robust against the packet losses where they replace the missing low distortion blocks of the lost description with their high distortion version from the delivered description.
In [7], the authors propose a method which uses the scalability features of 3D DWT through the application of a t+2D wavelet transform [19] to each GOP. Subsequently, the authors divided the wavelet coefficients into three descriptions by utilizing a modified zigzag scanning methodology. Finally, based on the required quality and the date rate of each network channel, the descriptions were scaled by optimizing their threshold values. All missing data are replaced with zeros before reconstruction.

III. Self-Similar Descriptions

The presented method for combining MDC with SVC in wavelet domain has the robustness in terms of the network errors, and flexibility in terms of the bandwidth usage changes. In case of error when one or some descriptions are lost, the video is reconstructed by estimating the lost coefficients using the delivered coefficients before applying an inverse DWT transform. The estimation of the lost coefficients is performed by utilizing the self-similarity of data after applying DWT transform which is explained as follows. After applying DWT, most of the coefficients in the high frequency bands have very small absolute values. These small values are replaced by zeros after the quantization step [8], [17]. The discrete wavelet transform however, has the extra characteristic of self similarity. If we consider a multi-layer decomposition of an image using DWT, where the lower levels correspond to higher frequencies and higher levels correspond to lower frequencies, we can easily observe a decrease of energy when moving from a higher level to a lower level. Furthermore, if coefficients at a low level contain small energy, their corresponding coefficients at the same spatial orientation at a higher level will also contain low energy. This similarity between the coefficients at similar spatial locations of a multi-layer wavelet decomposition is called self similarity characteristic. This characteristic is a property of natural images since the object boundaries in these images are not completely sharp and are reflected at different frequency levels. The self-similarity characteristic of the DWT can be exploited to interpolate the missing data providing better bit error rate in video streams.

A. Organizing DWT Coefficient in Self-Similar Descriptions

In the method presented here, the wavelet coefficients as depicted in Figure 1, are decomposed into three descriptions. The wavelet transform is repeated twice and the low frequency part of the coefficients is repeated redundantly in each description. The wavelet coefficient content of the descriptions are as given in Table I. The labels LLLL, LLHL, HL, LLHH, HH, LLLH, and LH refer to the group of wavelet transform coefficients as depicted in Figure 1. The low frequency coefficients (LLLL) are repeated in each description hence always a minimum level of fidelity in the reconstructed video is guaranteed. The reconstruction in presence of error or loss of a description is carried out by estimating the missing coefficients with the corresponding coefficients in other sub-bands. For each description, we have computed a parameter termed as similarity coefficient (ζ) which indicates the average ratio of the low frequency sub-band coefficients to their corresponding high frequency coefficient. Besides, a new scheme for structuring the descriptions is proposed. The new scheme provides the facility of utilizing the self-similarity characteristic of DWT for estimating the coefficients of the missing description. In our proposed method, we define three coefficient groups as:

TABLE I
THE COEFFICIENTS INCLUDED IN EACH DESCRIPTION.

Description Number	Coefficients Included
Description 1	LLLL, LLLH, LH
Description 2	LLLL, LLHH, HH
Description 3	LLLL, LLHL, HL

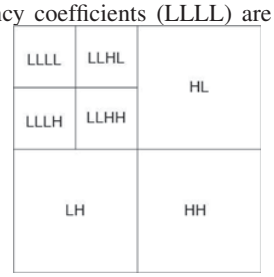

Fig. 1. 2D Wavelet transform coefficients.

Group 1 LLLL, LLLH, LH
Group 2 LLLL, LLHH, HH
Group 3 LLLL, LLHL, HL

The main idea in the proposed method is that when some of the coefficients in a coefficient group are lost, they can be estimated by means of the existing self-similarity. However, if we decompose the coefficients in a way that each coefficient group is transmitted in one description, in case of a loss or corruption in the description, the whole coefficient group is lost. Therefore, estimating the values of the lost coefficients by means of self-similarity will not be possible. However, if each description contains coefficients from different coefficient groups, then in case of a description loss, the coefficients can be estimated from the delivered description. The new organization of the coefficient groups in the descriptions is presented in Table II.

TABLE II
THE COEFFICIENTS INCLUDED IN EACH RE-ORGANIZED DESCRIPTION.

Description Number	Coefficients Included
Description 1	LLLL, LLLH, HH
Description 2	LLLL, LLHH, HL
Description 3	LLLL, LLHL, LH

The self-similarity between LLHL and HL for instance can be utilized to estimate the coefficients when one of these groups is lost. In case that HL is not available, LLHL can be up-sampled for an estimation, and when LLHL is lost, HL is down-sampled to obtain an approximation for LLHL. A similar method is used for (LLLH, LH) and (LLHH, HH) coefficient groups. As mentioned above, the proposed method ensures that the groups of self-similar coefficients are always transmitted in distinct descriptions In this way, by assuming that only one description is lost, the reconstructor will receive one coefficient group completely while, the remaining two coefficient groups are received partially. The partial coefficient groups can be completed by either up-sampling or down-sampling the available coefficients.

B. Reconstructing the Video

The frame reconstruction is presence of data loss in one or two descriptions is explained below, and the trivial case of no packet loss is not explained. We have observed that the self-similarity in a description is directly proportioned to the frequency content of the macro-block. This means that although there exists a strong correlation between the DWT coefficients at a sub-band, the ratio of the low frequency coefficients to the high frequency coefficients varies with the content of the block. Hence, we define a self-similarity index for each macro-block which is computed for the coefficients at (LH, LLLH) group as given in Equation 1. A similar method is used for computing the similarity index in other sub-bands.

$$\xi = \frac{1}{m^2}\left\| \frac{[\uparrow LLLH]_{ij}}{[LH]_{ij}} \right\|, \quad [LH]_{ij} \neq 0 \tag{1}$$

where $[LL]_{ij}$ indicates the matrix element at ij position, and the division of LLLH by LH is an element-wise division. Besides, m^2 is the number of non-zero coefficients in $[LH]$, $\uparrow$ is used to represent up-sampling operation, and ξ is the self-similarity index. $\|A\|$ is the matrix norm as defined in Equation 2.

$$\|A\| = \sum_i \sum_j |A_{ij}| \tag{2}$$

Self-similarity index for each description is encoded and transmitted with the current description and its following description in a circular manner. This means that the similarity index of (LLLH, LH) coefficient group is transmitted in descriptions 1 and 2, the similarity index of (LLHH, HH) coefficient group is transmitted in descriptions 2 and 3, and finally the similarity index of (LLHL, HL) coefficient group is transmitted in descriptions 3 and 1. The similarity index values are rounded to nearest integer and an upper limit of 8 has been considered for their values (similarity index values greater than 8 are considered as 8).

1) Case 1: One Description is Lost: Assuming description 2 is lost the decoder should estimate the low frequency coefficients at LLHH and high frequency coefficients at LH. The similarity index of (LLHH, HH) coefficient group is included in description 3 as well. Therefore, down-sampling coefficients at HH and multiplying them by their corresponding similarity index provides the estimation of the missing coefficients. Similarly, the missing coefficients at LH are estimated by up-sampling LLLH coefficients and multiplying them by the similarity index included in description 1.

2) Case 2: Two Descriptions are Lost: Assuming description 2 and 3 are lost the decoder should estimate the low frequency coefficients at LLHH and LLHL, and high frequency coefficients at LH and HH. The similarity indices and coefficients LLLH and HL transmitted in description 1 are utilized to estimate LLHH and LH. As a result, description 2 is estimated from the coefficients and similarity indices delivered with description 1. However, description 3 is cannot be estimated by the proposed method and its coefficients are replaced with zeros.

IV. EXPERIMENTAL RESULTS

The proposed method is experimentally verified using several video sequences. In order to verify the performance of our method, we considered two cases of packet losses as below:

- Only one description is lost. In this case the information in the delivered descriptions is utilized for reconstructing the video.
- Two descriptions are lost. Since part of the coefficients belonging to the adjacent description is in the delivered description, our proposed method is able to reconstruct one of the lost descriptions.

In both experimental cases mentioned above we repeated the experiments by changing the missing description. The GOP length has been fixed to 32 frames. The DWT transform is applied twice as depicted in Figure 1.

To emphasize the important impact of the self-similarity feature of the DWT in reconstruction in presence of data loss, we have compared the reconstructed video when the missing description is estimated using self-similarity feature, and the same video when the coefficients at the missing description are replaced with zeros. Figures 2 and 3 provide the comparative results for one description loss in low and high bit rates respectively which are averaged over the blocks of each frame. The experimental results provided in Figures 2 and 3 indicate that the self-similarity based estimation of the missing data is more effective in higher bit-rates which can be related to the fact that in low bit-rates the higher frequency coefficients are mostly zeros. Despite the fact that the results are not much different in low bit-rates, in high bit-rates we can see an average improvement of 2.14(dB) in terms of PSNR values in Figure 3.

Our next experiment is evaluation of the method when one description is lost. In [21] the authors combine layered coding

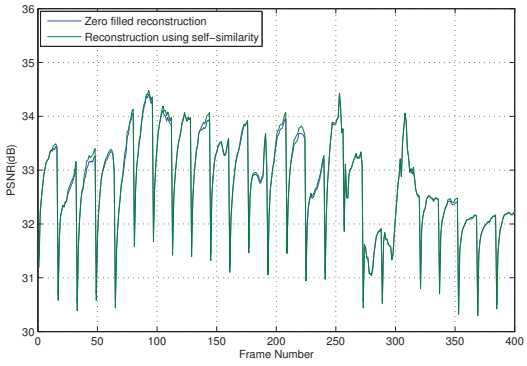

Fig. 2. PSNR values of the reconstructed frames by replacing missing coefficients with zero, and estimating using self-similarity in low bit-rate.

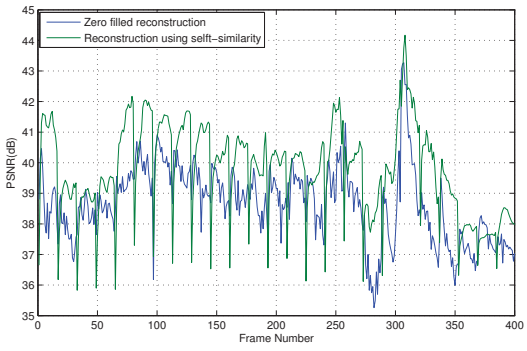

Fig. 3. PSNR values of the reconstructed frames by replacing missing coefficients with zero, and estimating using self-similarity in high bit-rate.

MDC methods for error-resilient video transmission over unreliable channels. They used unequal loss protection to provide the base layer with the highest level of channel error protection through the use of Forward Error Correction (FEC) coding. In order to cope with the network congestion which is main cause of packet losses, they have considered erasure codes for data protection. The FEC code creates redundancy in the transmitted video which makes the method proposed in [21] similar to our proposed method as our proposed method repeats the low frequency coefficients in all descriptions. The authors in [21] divide a bitstream into two portions where the first portion (b_1) consists of the base layer and is further divided into sub-bitstreams. The second portion (b_2) includes the enhancement layers and is also divided into sub-bitstreams. A description is created by including x sub-bitstreams from b_1 and y sub-bitstreams from b_2. As it is assumed the descriptions are transmitted over channels with different probability of data loss, they are protected against packet losses in an unbalanced way through FEC. Besides, in [1] the authors propose a SVC method which decomposes the video into multiple descriptions. Their method combines video segments coded at high and low rates and transmits the high rate segments from one stream together with the low rate segments of the other streams in each description. The low rate coefficients are used for reconstructing the missing description(s) in a lower quality. The authors also propose a Multiple-Objective Optimization (MOO) framework for selection of the best encoding configuration to achieve the best tradeoff between redundancy and reliability. Since their proposed method includes redundancy in each description to attain a better quality level in presence of packet losses, we have considered their method as a benchmark to compare the performance of our MD video coder.

Figure 4 depicts the comparative performance of the proposed method and the methods proposed in [1] and [21] . We have assumed that only one description is lost and later on reconstructed by using the redundancy available in other descriptions. When the self-similarity feature of the DWT

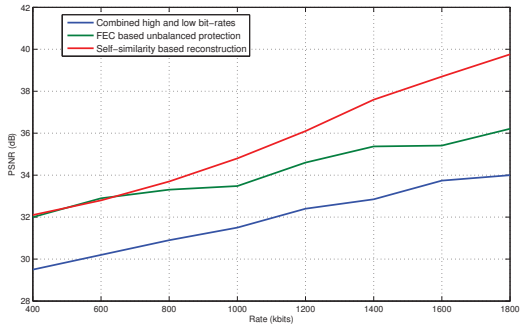

Fig. 4. PSNR values of the reconstructed frames by the proposed method, the combined high-low rate coding method [1], and FEC based unbalanced protection [21] when one description is lost.

is utilized, our proposed method outperforms the other two methods in high bit-rates. One important feature of the method proposed in [1] is that the redundancy is proportional to the intended total bit rate while in our proposed method, the low frequency part of the coefficients are repeated in all descriptions almost independently from the bit rate.

In our next experiment we have assumed that two out of three descriptions are lost during transmission. For fair performance analysis and comparison, we have modified the proposed methods given in [21] and [1] in order to include three descriptions. Figure 5 depicts the result of the reconstruction versus average PSNR value. The performance of the proposed method is considerably better in presence of high packet losses such as the case depicted in Figure 5. This result indicates that the proposed method is suitable for transmitting high rate videos over unreliable networks. The experimental results indicate that the proposed method outperforms the traditional video coding methods in presence of frame losses. When two descriptions are delivered, the average PSNR values with and without using self-similarity index for reconstruction are 35.69(dB) and 33.55(dB) respectively. The average PSNR values when only one description is delivered are 34.38(dB) and 27.12(dB) for reconstruction with and without using self-similarity index respectively. The redundancy imposed by repeating the low frequency coefficients of the DWT can be minimized by increasing the number of times the DWT is applied to frame

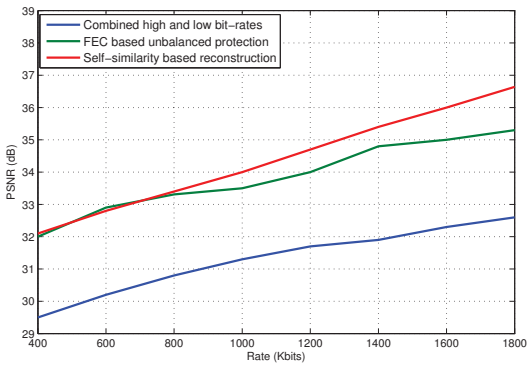

Fig. 5. PSNR values of the reconstructed frames by the proposed method, the combined high-low rate coding method [1], and FEC based unbalanced protection [21] when two descriptions are lost.

data. Moreover, a better performance of the proposed method in case of higher bit-rates indicates that the proposed method is more suitable for streaming over unreliable networks of high bandwidths.

V. Conclusions

A new DWT based video coding method for transmitting video over unreliable networks is proposed. The frame blocks are decomposed into three descriptions after applying the DWT transform. The proposed method improves the performance of the existing MDC methods by utilizing the self-similarity feature of the DWT. The experimental results indicate that the proposed method outperforms the existing methods when the video bit rate and the packet loss rate are high. The redundancy added by repeating the low frequency data in each description can be minimized by increasing the number of times that the DWT applied. However, the number of DWT levels should be optimized with the number of self-similarity index values which are transmitted in descriptions. Besides, the optimization can be performed by considering the available bandwidth of the underlying network.

References

[1] T. B. Abanoz and A. M. Tekalp. SVC-based scalable multiple description video coding and optimization of encoding configuration. *Signal Processing: Image Communication*, 24:691–701, 2009.

[2] N. Adami, A. Signoroni, and R. Leonardi. State-of-the-art and trends in scalable video compression with wavelet-based approaches. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1238–1255, 2007.

[3] E. Akyol, A. M. Tekalp, and M. R. Civanlar. A flexible multiple description coding framework for adaptive peer-to-peer video streaming. *IEEE Journal of Selected Topics in Signal Processing*, 1:231–245, 2007.

[4] M.R. Ardestani, A. A. Beheshti Shirazi, and M. R. Hashemi. Low-complexity unbalanced multiple description coding based on balanced clusters for adaptive peer-to-peer video streaming. *Signal Processing: Image Communication*, 26:143–161, 2011.

[5] M. Biswas, M. R. Frater, and J. F. Arnold. Multiple description wavelet video coding employing a new tree structure. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(10):1361–1368, 2008.

[6] S. Cho and W. A. Pearlman. A full-featured, error-resilient, scalable wavelet video codec based on the set partitioning in hierarchical trees (SPIHT) algorithm. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(3):157–171, 2002.

[7] R. Choupani, S. Wong, and M. Tolun. Scalable video transmission over unreliable networks using multiple description wavelet coding. *The 7th International Conference on Digital Content, Multimedia Technology and its Application (IDCTA2011)*, pages 5–10, 2011.

[8] M. L. Comer, K. Shen, and E. J. Delp. Rate-scalable video coding using a zerotree wavelet approach. *Proceedings of the Ninth Image and Multidimensional Digital Signal Processing Workshop*, pages 162–163, 1996.

[9] V. K. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18:74–94, 2001.

[10] F. A. Lopez-Fuentes. P2P video streaming combining SVC and MDC. *International Journal of Applied Mathematics and Computer Science*, 21(2):295–306, 2011.

[11] A. Mavlankar and E. Steinbach. Multiple description video coding using motion-compensated lifted 3d wavelet decomposition. *IEEE International Conference on Acoustics, Speech, and Signal Processing,(ICASSP '05)*, 2:65–68, 2005.

[12] N. Mehrseresht and D. Taubman. A flexible structure for fully scalable motion-compensated 3-d dwt with emphasis on the impact of spatial scalability. *IEEE Transaction on Image Processing*, 15(3):740–753, 2006.

[13] A.R. Reibman, H. Jafarkhani, Y. Wang, M.T. Orchard, and R. Puri. Multiple-description video coding using motion-compensated temporal prediction. *IEEE Transaction on Circuits and Systems for Video Technology*, 12:193–204, 2002.

[14] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):243–250, 1996.

[15] T. Tillo, M. Grangetto, and G. Olmo. Multiple description image coding based on lagrangian rate allocation. *IEEE Transaction on Image Processing*, 16(3):673–683, 2007.

[16] R. Venkataramani, G. Kramer, and V.K. Goyal. Multiple description coding with many channels. *IEEE Transaction on Information Theory*, 49:2106–2114, 2003.

[17] Q. Wang and M. Ghanbari. Scalable coding of very high resolution video using the virtual zerotree. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(5):719–727, 1997.

[18] Y. Wang, A. R. Reibman, and L. Shunan. Multiple description coding for video delivery. *Proceedings of IEEE*, 93:57–70, 2005.

[19] M. Weeks and M.A. Bayoumi. Three-dimensional discrete wavelet transform architectures. *IEEE Transactions on Signal Processing*, 50(8):2050–2063, 2002.

[20] M. Wien, H. Schwarz, and T. Oelbaum. Performance analysis of SVC. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1194–1203, 2007.

[21] W. Xiang, C. Zhu, C. K. Siew, Y. Xu, and M. Liu. Forward error correction-based 2-d layered multiple description coding for error-resilient H.264 SVC video transmission. *IEEE Transaction on Circuits and Systems for Video Technology*, 19(12):1730–1738, 2009.

Journal of Electronic Imaging 21(4), 043006 (Oct–Dec 2012)

Unbalanced multiple description wavelet coding for scalable video transmission

Roya Choupani
Delft University of Technology
Delft, The Netherlands
and
Çankaya University
Ankara Turkey
Email: roya@cankaya.edu.tr

Stephan Wong
Delft University of Technology
Delft, The Netherlands

Mehmet Tolun
TED University
Ankara Turkey

Abstract. *Scalable video coding and multiple description coding are the two different adaptation schemes for video transmission over heterogeneous and best-effort networks such as the Internet. We propose a new method to encode video for unreliable networks with rate adaptation capability. Our proposed method groups three dimensional discrete wavelet transform coefficients in different descriptions and applies a modified embedded zero tree data for rate adaptation. The proposed method optimizes the bit-rates of the descriptions with respect to the channel bit rates and the maximum acceptable distortion. The experimental results in the presence of one description loss indicate that on average the videos at the rate of* 1000 Kbit/s *are reconstructed with Y-component of peak signal to noise ratio (Y-PSNR) value of 36.2 dB. The dynamic allocation of descriptions to the network channels is optimized for rate distortion minimization. The improvement in term of Y-PSNR achieved by rate distortion optimization has been between 0.7 and 5.3 dB in different bit rates.*
 [DOI: 10.1117/1.JEI.21.4.043006]

1 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 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.[2–4] The quality increase is accomplished through the gradually 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 (FGS) scheme defines the video content in a multilayer format where the existence of at least one layer, or the base layer, containing the most basic data is required. The remaining layers, called enhancement layers, increase the quality of the video.[5] A higher quality video is attained by increasing the number of layers decoded at the receiver side. Adapting video streams at the transmitter can also be done by considering the current state of the network channel. In video streaming applications such as peer-to-peer networks, a feedback mechanism is used to inform the transmitter about the current state of each channel. This information is used for optimizing the data rates with respect to the video quality.[6] In this paper, we propose a method for optimizing the channel data rates which combines SVC with multiple description coding (MDC). We assume the network channels have different and variable characteristics. Therefore, the data rate of each description should be adjusted to satisfy the maximum acceptable video distortion with respect to the available bit rate for each channel. Furthermore, we are assuming the channel bit rates and the requested video quality are available to the transmitter. Meanwhile the proposed encoding method provides the possibility of video scaling at the receiver side through truncating parts of the coded stream for signal-to-noise ratio (SNR) or temporal scale-down. This work is an extension to our paper published in the 7th International Conference on Digital Content, Multimedia Technology and its Applications.[7] The remainder of this paper is organized as

Paper 12087 received Mar. 8, 2012; revised manuscript received Aug. 24, 2012; accepted for publication Sep. 10, 2012; published online Oct. 4, 2012.

0091-3286/2012/$25.00 © 2012 SPIE and IS&T

follows: Section 2 summarizes the related work on SVC using MDC. Section 3 presents the details of our proposed method. Section 4 describes the experimental results of our method. We draw our conclusions in Sec. 5.

2 Related Work

A multiple description coder divides the video data into several bit-streams called descriptions which are then transmitted separately over the network.[8] 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.[9] The accuracy of the decoded video depends on the number of received descriptions.[10] 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 recovered from other descriptions. Descriptions are defined by constructing P nonempty sets through partitioning the original signal f so that these sets sum up to 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 of the video. Repeating a signal sample in multiple descriptions is also a way of assigning higher importance to parts/signals of the video. The duplicate signal values increase the redundancy and hence the efficiency is reduced. Even if the descriptions are designed as nonoverlapping sets, or partitions, it does not necessarily mean that there is no redundancy in the data. Given that each partition is encoded independently from other partitions, the spatial or temporal correlation between data in different partitions is not utilized and redundancy is not eliminated.[11] However, the preserved spatial or temporal correlation can be used for estimating the lost bits,[12] which is commonly referred to as error concealment.

FGS-based MDC schemes partition the video into one base layer and one or several enhancement layers.[13] The base layer can be decoded independently from enhancement layers but it provides only the minimum spatial, temporal, or SNR quality.[14] 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 the 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 corrupted.[15] Repeating base layer bits in each descriptor is the overhead added for a better

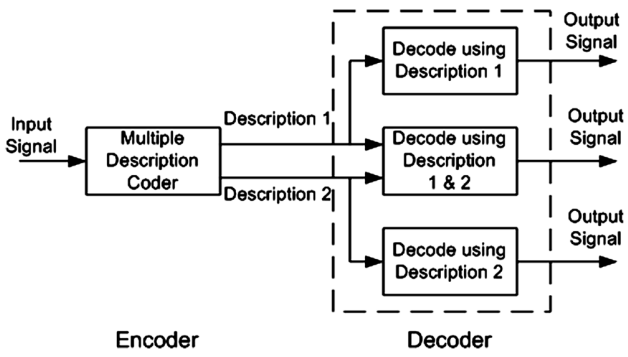

Fig. 1 Multiple descriptions coding block-diagram.

error resilience. In Ref. 16, the authors propose to generate multiple scalable descriptions from a single SVC bit-stream by mapping scalability layers of different frames to different descriptions. Their scheme is intended for Peer-to-Peer (P2P) streaming over multiple multicast trees and features several encoding parameters, such as base layer rate of descriptions and overall redundancy. They tried to optimize mean rate-distortion performance of each description received over a packet loss network, range of extraction points of the SVC stream, and overall redundancy of their MDC scheme. In Ref. 17, the SVC is combined with MDC schemes by subsampling in both horizontal and vertical directions yielding four subsequences. The authors used two approaches to combine the subsequences into two descriptions. In the first approach, each description is encoded by predicting one subsequence from the other using the inter-layer prediction tools. The second approach exploits the redundancy between the subsequence with the hierarchical dyadic B frame prediction algorithm. The authors in Ref. 18 present a solution for the differences in the types of delivered services in H.264-based SVC combined with MDC by using optimization and control strategies. In Ref. 19, an algorithm is proposed to control the mismatch between the prediction loops at the encoder and decoder in MDC with motion-compensated predictions. They consider three cases when both descriptions received or either of the single descriptions is received. In Ref. 20, an optimization algorithm for minimizing rate distortion for P2P networks is proposed. The method considers each sub-bitstream of a frame is considered as a description. The optimization is based on skipping or selection of an MDC packet so that the constraints of limited downloading bandwidth are met while the total distortion reduction is maximized.

In forward error correction (FEC)-based MDC methods, it is assumed that the video is originally defined in a multiresolution manner. This means if we have M levels of quality, each level increases the fidelity to the original video. This concept is very similar to the multilayer video coding method used by an FGS scheme. The main difference, however, is that there exists 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 MDC schemes 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. In Ref. 21, the authors used unequal loss protection (ULP) to protect 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.[22] Error resilience can also be provided by repeating some important parts of data as suggested in Ref. 23. The proposed MDC model in Ref. 23 utilizes a nondyadic hierarchical B-picture structure with four levels. It duplicates the key frames in each group of pictures (GOP) for providing error resilience. Hence, each description has its copy of the key frames. However, since the frame organization is hierarchical, the duplicated frames are at the lowest level.

If a frame in a higher level is lost, it is reconstructed from its counterpart frames from the second description.

discrete wavelet transform (DWT)-based video coding methods are convenient for applying multiple description coding.[24] 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.[25,26] 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.[7,24,27–29] More efficient methods utilize motion compensated temporal filtering (MCTF) which is aimed at removing the temporal redundancies of video sequences. In Ref. 30, MDC-SVC based on MCTF and 2D DWT is used for video streaming over P2P networks. The receiving peer can measure the channel conditions such as the packet loss rate and bandwidth of each sending peer's path in each GOP period and then calculates the optimal encoder parameters for that GOP through a post-encoding procedure. The resultant encoding parameters are sent to the sending peers through the feedback control channels. Also, in Ref. 6 an adaptive P2P video streaming system with a flexible multiple description coding (F-MDC) framework is proposed. They intended to adapt the number of base and enhancement descriptions, and the rate and redundancy level of each description. They combine their F-MDC framework with SVC by using JPEG2000 based t + 2D DWT allowing each code-block at any point of bit-plane codes to be truncated. In Ref. 31, the authors created descriptions by partitioning the transform domain of the signal into maximally separated sets. They applied their method to the Internet transmission of subband/wavelet-coded images and scalable motion compensated three-dimensional (3-D) subband/wavelet-coded video.

Our proposed method falls into the group of DWT-based methods; however, we have organized the coefficients in different description and minimized the common redundant part in them. Meanwhile, our proposed method provides the feasibility of having scalable transmission over each channel by encoding the coefficient of each description by using a modified EZW zig-zag scanning. In addition, assuming that the frequency content of a video is not distributed equally in all different directions (horizontal, vertical, and diagonal), we propose a method for dynamic allocation of channels to descriptions for minimizing the rate distortion of the video.

3 Proposed Method

Our proposed method involves using the scalability features of 3-D discrete wavelet transforms through the application of the t + 2D wavelet transform[32] to each GOP. Then by means of a modified zig-zag scan the wavelet coefficients are grouped in three descriptions. Finally, based on the required quality and the date rate of each network channel, the descriptions are scaled by optimizing their threshold values. In general the proposed method addresses MDC for unbalanced dynamic network channels and optimizes rate distortion over theses channels. The main features of our proposed method are as below:

(1) Defining a MDC scheme for scalable video transmission and error concealment which optimizes the rate distortion.

(2) The frequency content of the video is also considered in rate distortion optimization. Since descriptions represent the wavelet coefficients in horizontal, vertical, and both horizontal and vertical directions, a video having more frequency content in horizontal/vertical direction will have more nonzero coefficients in its corresponding description. Hence the system will assign this description to the channel with highest data rate.

(3) The dynamic channel rate changes and video content changes are combined in optimizing descriptions for transmission over each channel.

(4) Our proposed method defines descriptions in a way that each one contains the coefficients belonging to the frequency information in one direction and organizes them in an EZW tree. This provides the feasibility of truncating the bit stream in each description at any point for scalability.

The following subsections describe each step in details.

3.1 *Applying t + 2D Wavelet Transform*

The multiresolution analysis is based on the concept of vector spaces. For each vector space, there exists another vector space of higher resolution which contains all vector spaces of the lower resolution. The vector spaces and wavelets are closely related as the scaling functions of the wavelets are the basis for these vector spaces. An image can be assumed as a vector space where higher resolution image correspond to vector spaces of higher resolution. The original image or signal can be reconstructed from the subsampled images by means of conjugate mirror filters.[33] The conjugate mirror filters and ortho-normal wavelet basis are closely related. The scaling functions and wavelets are determined by low-pass and high-pass filters, respectively. By defining wavelets so that the scale is a power of 2 and the time an integer multiple of scale, an orthogonal basis as given in Eq. (1) is obtained:

$$w_{j,k} = 2^{j/2} w(2^j t - k), \tag{1}$$

where j and k indicate scale and time, respectively and $w(\cdot)$ is the time-shifted and time-scaled wavelet. There is a large class of wavelet functions which have an orthogonal basis and are called the orthogonal wavelets. The simplest orthogonal wavelet is the Haar wavelet. Haar filters have the property of being conjugate mirror filters having finite impulse response in the univariate case and no other filter bank has this property.[34]

In the proposed method, the frames of each GOP pass through a Haar lifting stage.[35] The splitting and prediction steps of the lifting process are repeated in several stages in a hierarchical structure. The general view of the hierarchical structure and the applied lifting method are depicted in Fig. 2. The frames in a group of pictures (GOP) are organized in pairs where 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

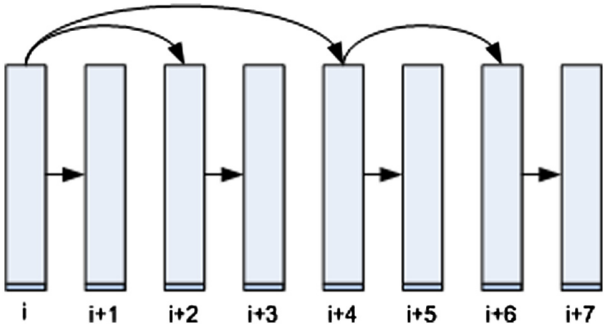

Fig. 2 Proposed hierarchical wavelet lifting structure.

pair. This procedure is repeated at the following levels of the tree hierarchy. A serious problem with SVC schemes is a quality degradation which is accumulated frame by frame and is referred to as drift error. Drift error is the result of selective transmission where some of the DCT coefficients are eliminated and/or requantized, which changes the original quantized DCT coefficients.[36] The drift error is reduced considerably by applying the proposed hierarchical structure. 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 in a GOP is limited to the tree height and therefore, the GOP size and hence tree height should be determined in a tradeoff with the maximum tolerable drift error. This structure reduces the drift error in a logarithmic manner. The proposed structure falls in the group of nondelay methods where no frame needs to be buffered till the arrival of the following frame(s) for decoding. This makes the decoder implementation simple, with minimal memory requirement. After applying the Haar lifting to each GOP and organizing the frames in hierarchical structure, each frame of the GOP undergo three levels of 2D wavelet transform.

3.2 *Description Creation*

The wavelet coefficients of each description are quantized and coded using a modified EZW scheme which is proposed to arrange the coefficients. The zig-zag scanning used in the original EZW starts with a threshold value T. It scans the coefficients from low frequency to high frequency and assigns a positive (P), a negative (N), a zero (Z), or zero tree (t) symbol to a coefficient if it is greater than T, less than $-T$, between $-T$ and T, or the whole branch rooted at that coefficient contains values between $-T$ and T, respectively. In the next iteration the threshold is reduced and the procedure is repeated. Here the emphasis is put on low frequency values, so that if the stream is truncated high frequency symbols will be deleted, and coefficients with large magnitude because the threshold starts with a large value initially. Meanwhile, the self-similarity property of the wavelet transform makes it possible to represent a whole sub-tree by a single symbol (t) which reduces the total bit rate required. The modified EZW scheme follows the low frequency to high frequency order in transmission of information, and preserves the self-similarity between the coefficients. In our modified EZW, the coefficients are grouped in three descriptions and EZW is utilized for encoding the coefficients in each description. In order to distribute the low and high frequency information between three descriptions evenly, each sub-tree rooted at the low frequency part of the coefficients is used as a description. Given that the self-similarity is among the coefficients of a subtree, reducing the size of the coded data by replacing a sub-tree with a (t) symbol is still possible. The differences between the original zero tree and the proposed modified zero tree are two fold:

(1) The zig-zag scanning is performed in each sub-tree separately and according to the order depicted in Fig. 3.

(2) Each sub-tree corresponds to one of the descriptions. The low frequency part of the coefficients is repeated in all descriptions.

3.3 *Rate Distortion Optimization*

The descriptions obtained after zig-zag scanning the wavelet coefficients and organizing them in zero-trees, are transmitted over network channels with different bit-rate characteristics. Our assumption is that the transmitter is aware of the channel rate changes and the video quality requirement of the receiver. In this part, we show how the descriptions are optimized in the presence of the bit rate restrictions. The optimization is defined as:

Minimize $R_{\text{total}} = R_1 + R_2 + R_3$ subject to

$$R_i \le R_i^T \quad \text{for } i = 1,\ 2,\ 3 \quad Q(R_1, R_2, R_3) \ge Q^T,$$

where R_i is the bit rate of the network channel i allocated to one of the descriptions, R_i^T is the bit rate threshold of channel i which indicates the max data rate available to the allocated

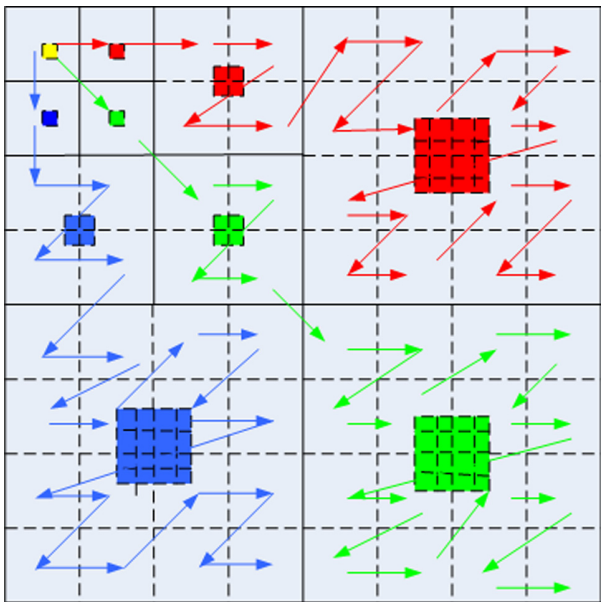

Fig. 3 Zig-zag scanning of DWT coefficients at each description. Each description is illustrated in a different color.

description, $Q(R_1,\ R_2,\ R_3)$ is the quality obtained after receiving three descriptions with data rates R_1, R_2, and R_3, respectively. $Q(R_1,\ R_2,\ R_3)$ is computed by joining the coefficients from three descriptions and applying IDWT as shown in Eq. (2),

$$Q(R_1,\ R_2,\ R_3) = \mathrm{IDWT}(R_i^1 \bigoplus R_i^2 \bigoplus R_i^3), \qquad (2)$$

where $\bigoplus$ is used to indicate the merging of the coefficients at the descriptions. The frame obtained then is compared to the original frame using a peak signal-to-noise-ratio (PSNR) metric and hence the value returned by $Q(R_1,\ R_2,\ R_3)$ is the computed PSNR value. Q^T is the minimum required quality. We measure the video quality using the average PSNR of the frames in a GOP. The PSNR is defined in Eq (3):

$$\begin{aligned} \mathrm{PSNR} &= 20\log_{10}\frac{\mathrm{Max}_I}{\sqrt{\mathrm{MSE}}} \\ \mathrm{MSE} &= \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|I(i,\ j) - I'(i,\ j)\|. \end{aligned} \qquad (3)$$

The minimization problem is solved in an iterative form using the steepest descent algorithm. Steepest descent is used for finding the nearest local minimum of a function by starting at a point P_0 and iteratively approaching the minimum point by moving from the current point P_i to point P_{i+1} along $-\nabla f(P_i)$. In our implementation, we reduce the data rate of each description by $\mathrm{d}R_i^j$ at each iteration, where the subscript indicates the description and the superscript is the current iteration. Given that the number of descriptions is limited in our method, we consider the optimum descent direction $-\nabla f(P_i)$ by evaluating the reconstructed video quality in each case as described in Algorithm 1. Our aim is finding the optimum data rates of descriptions which satisfy the network requirements and minimize the distortion.

The assumption of the proposed algorithm is that the least distortion in video quality after applying a rate reduction to the descriptions should be chosen first. This corresponds to the fastest rate reduction with smallest quality degradation which follows the steepest descent method.

3.4 *Scalability Feature of the Proposed Method*

As our proposed method is based on discrete wavelet transform, multilayer 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 is achieved by organizing the wavelet coefficients in an EZW spatial tree structure.

Temporal scalability in traditional video coding methods is achieved through placing some of the frames in the base layer and the rest in the enhancement layer(s). 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 reduction in the compression efficiency or causing any drift error problem. A higher rate of scalability is possible by eliminating the second frames at the next level, etc.

Algorithm 1 Rate distortion optimization algorithm.

$j \leftarrow 1$

$R_1^1 \leftarrow R_1^T$

$R_2^1 \leftarrow R_2^T$

$R_3^1 \leftarrow R_3^T$

WHILE $Q(R_1, R_2, R_3) \geq Q^T$ **DO**

Compute($\mathrm{d}R_1^j$)

Compute($\mathrm{d}R_2^j$)

Compute($\mathrm{d}R_3^j$)

$Q_1 \leftarrow Q(R_1 - \mathrm{d}R_1^j, R_2, R_3)$

$Q_2 \leftarrow Q(R_1, R_2 - \mathrm{d}R_2^j, R_3)$

$Q_3 \leftarrow Q(R_1, R_2, R_3 - \mathrm{d}R_3^j)$

$m\mathrm{Index} \leftarrow \mathrm{MinIndex}(Q_1, Q_2, Q_3)$

Update($R_{m\mathrm{Index}}, \mathrm{d}R_{m\mathrm{Index}}$)

$j \leftarrow j + 1$

END

4 Experimental Results

For evaluating the proposed method we considered two sets of experiments. The first set is developed to verify the efficiency of our rate distortion optimization algorithm. The second set evaluates the performance of the method in presence of data loss.

4.1 *Rate Distortion Optimization Evaluation*

We tested the performance of the rate distortion optimization by considering three channels for transmitting three descriptions. The channels have varying data rates which are known by the transmitter. Table 1 gives the data rates of each description used in our test cases and the bit rates over each description after optimization. The values provided in Table 1 are the average values of all frames of the test videos listed in Table 2. An important observation in the results of our experiments is that when the data rate of one of the channels is low, the algorithm tends to eliminate it as the optimum rate is achieved by using the remaining two descriptions. However, considering the probability of data loss in these channels, this will affect the error concealment feature of multiple description coding scheme. As an implementation measurement, we do not allow the optimization algorithm to reduce the data rate of a description beyond its enhancement layer. This means each description contain at leat the base layer of the DWT coefficients. The third row in Table 1 indicates the case where one of the channels has a very low bit rate and the algorithms would tends to

Table 1 Assumed channel rates and their corresponding optimized description rates.

Description 1 rate, Kbit/s	Description 2 rate, Kbit/s	Description 3 rate, Kbit/s	Optimized rate 1, Kbit/s	Optimized rate 2, Kbit/s	Optimized rate 3, Kbit/s
250	250	250	233.12	240.04	212.46
100	100	100	97.1	98.22	96.15
350	350	25	340.04	346.19	24.7

eliminate the description in absence of the implementation measurement.

The second observation regards balancing the data transmitted over channels according to their available data rates and the frequency content of descriptions. The proposed rate distortion optimization, as explained in Sec. 3, creates three descriptions for the DWT coefficients. These coefficients however do not correspond to the same set of frequency contents. The first description contains the horizontal frequency of the frames where the vertical frequency is not significant. The third description corresponds to the vertical frequency of the frames with a small amount of horizontal frequency content. Finally, the second description contains the coefficients of the frames with both horizontal and vertical content. The proposed optimization method has the characteristic that the description carrying the least frequency content is clipped first during the optimization process. This feature causes unbalanced distribution of data over the available channels. An optimization in the implemented code solves the problem by labeling the descriptions as 1, 2, or 3 and sending the largest description (containing the most significant frequency content of the frame) using the channel with highest data rate. The required quality threshold values (Q^T) used for this experiment are 32, 34, and 36 dB and the indicated results are the maximum improvement values obtained. The impact of optimizing the description content with respect to the channel rates is depicted in Fig. 4. The results as depicted in Fig. 4 indicate that the implementation optimization are more effective in low data rates, as expected. In comparison to the method presented in Ref. 20, the optimization method proposed in this paper considers each description as a single unit which can either be transmitted or skipped, while our proposed method reduces the bit rate of the descriptions by truncating the stream of the DWT coefficients. Meanwhile, we have assumed the channel bit rates may change dynamically and hence the proposed method switches the channels used for the transmission of the description to minimize the rate distortion. Meanwhile, a comparison with the method proposed in Ref. 23 reveals that it is not capable of varying the bit-rate of each description to adapt it to the network conditions, a feature provided by our proposed method. Meanwhile, our proposed method has the property of minimizing rate distortion by combining the channel bit-rate and information content of each description. This feature makes it possible for the proposed method to relate the information content of the description to the channel capacity for optimization.

Table 2 Video sequences used for performance evaluation.

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

4.2 *Performance in Presence of Data Loss*

To verify the performance of the proposed method in presence of data loss we have conducted a series of experiments. Some implementation considerations for our experiments are as follows:

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

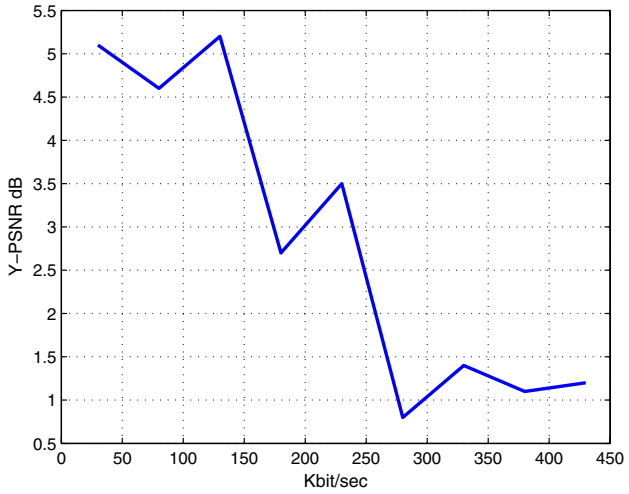

Fig. 4 The effect of optimizing descriptions with respect to channel rates.

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 PSNR criteria. The computed PSNR values for different thresholding levels are then utilized to draw the performance with respect to the bit per pixel values 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 considered three video sequences with the specifications given in Table 2. We included videos with high and low frequency contents for a better evaluation.

Hierarchical coding of the frames with 32 frames per GOP was applied to each video sequence. The hierarchical structure was wavelet transformed and split into three descriptions. The coefficients in each description were then coded using the EZW method. The test cases are devised in a way that both error resilience and scalability of the 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 reconstructed the video using two out of three descriptions. The reconstructed frames were compared with the original frames using PSNR values. The computation was carried out by putting aside one of the descriptions each time and the average of the resulted PSNR values was computed. In all experiments, the lost coefficients are replaced by zeros when we perform the inverse DWT. Table 3 shows the PSNR values for different cases of description losses for each of the test video sequences. The notation $PSNR_{ij}$ indicates that descriptions i and j have been received. The first experiment assumes all 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, the base layer, at each descriptions has also the impact of reducing the quality reduction effect. Our second experiment measured 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 description were computed and averaged for all GOPs in each video sequence. Table 4 shows the computed averages for each sequence separately.

Table 3 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

Table 4 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 reliability of the system and its capability to reconstruct the video in presence of network problems come with a price. The redundancy present in the proposed method reduces the coding efficiency as shown in Table 5. This redundancy is partially due to repeated low frequency coefficients of the wavelet transform and partially due to the coding method. The bit rates provided are obtained by coding the video using joint scalable video model (JSVM) of the joint video team (JVT) of the ISO/IEC moving pictures experts group (MPEG) and the ITU-T video coding experts group (VCEG) which is a publicly accessible tool.[37] As indicated in Table 5, an extra redundancy of 25% to 33.4% has been added to the sample sequences. The frequency content of the sequence is an affecting factor in the redundancy rate. This is on one hand due to better representation of high frequency data by the wavelet transform; however, on the other hand, as only the lower frequency coefficients are repeated in the descriptions of the proposed method, sequences with lower frequency contents have higher redundancy rates.

The scalability capability of the proposed method was 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. We obtained slightly smaller PSNR values compared to the case when the video is coded as a single description,[38] 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

Table 5 The redundancy rate of the proposed method in each of the video sequences used for performance evaluation.

Name	Rows × columns	Frame rate	Redundancy percentage
Foreman	352 × 288	30	25.0%
Stefan	768 × 576	30	33.4%
Suzie	144 × 176	30	31.2%

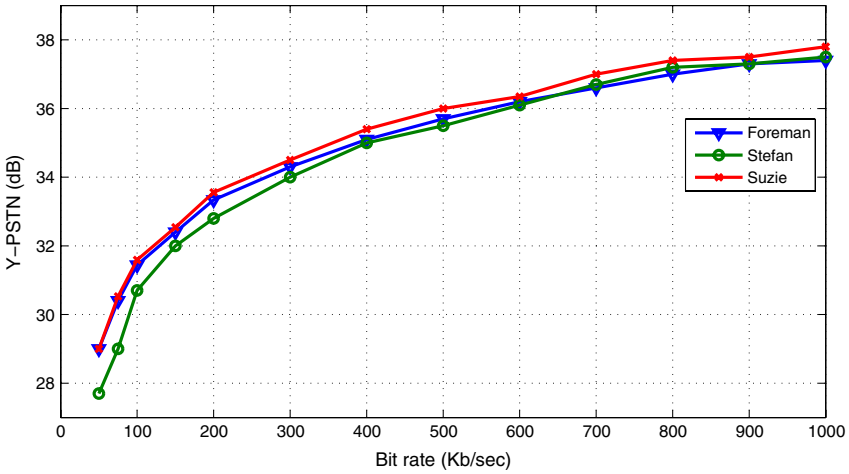

Fig. 5 Video reconstruction quality with respect to bit-rate using all descriptions.

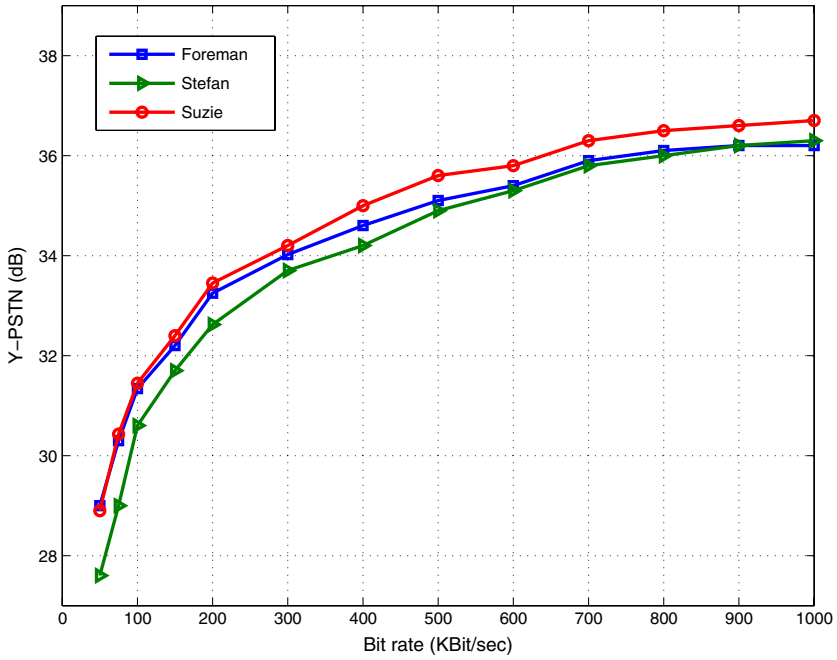

Fig. 6 Video reconstruction quality with respect to bit-rate in presence of a description loss.

levels decreases the size of the base layer and hence a smaller redundancy is imposed. However, the reconstruction error in presence of a description loss increases if the base layer is small. Figure 5 depicts the results of the reconstructed video as PSNR with respect to the bit rate. Here, we have assumed all descriptions are received without error. In our last experiment, we combined the scalability with description loss. The experiment was conducted by changing the threshold value of the EZW encoder to obtain different bit rates. Then we reconstructed the video using only two out of three descriptions. We computed the PSNR for the reconstructed video and considered the average of the PSNR values of reconstructed videos with one of the descriptions omitted. Figure 6 depicts the obtained result from three video sequences. It should be noted that the bit rate in the last experiment is obtained from the total number of bits contained in two delivered descriptions.

4.3 *Comparative Results on Bit-Rate Performance*

The proposed method was compared to another multiple description wavelet coder for some fixed rates[6] where a wavelet coder based on jpeg2000 with 3-level spatial and 4-level temporal decompositions has been used. The authors have called their method a flexible multiple description coder; hence, we will use F-MDC to refer to their method in the following part. The comparative results of the proposed method and F-MDC are provided in Table 6 for descriptions at three different rates and four redundancy levels when one description is lost. Our proposed coder

Table 6 Comparative performance of the proposed method/F-MDC in terms of Y-PSNR.

Sequence	100 kbps	200 kbps	300 kbps
Foreman	31.5/31.6	33.4/32.8	34.0/33.5
Stefan	30.4/30.7	31.6/31.5	33.8/33.4
Suzie	31.6/31.2	33.7/32.7	34.2/33.67

distributes the wavelet coefficients of each frame into three descriptions. However, F-MDC creates two base layer descriptions and two enhancement layer descriptions corresponding to the base layers. In our experiments we assumed a base layer description and its corresponding enhancement layer description are lost. To provide similar testing conditions, we assumed one description in a GOP and two descriptions in the following GOP are lost. This amounts to 50% frame loss in both methods. Our proposed method outperforms F-MDC in low redundancy high data rates. In low data rates most of the data in the descriptions are from the low frequency content which is repeated in the descriptions. However, in high date rate, high frequency data is dominant. Since our proposed method preserves the self similarity between the DWT coefficients, it provides higher performance compared to F-MDC in high data rates.

5 Conclusion

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 resilient transmission method, and a rate distortion method is proposed. In our proposed method which is based on DWT, we organized the coefficients in three different descriptions with a minimized common redundant part. Meanwhile, our proposed method provides the feasibility of having scalable transmission over each channel through encoding the coefficient is each description using a modified EZW zig-zag scanning. In addition, assuming that the frequency content of a video is not equally distributed in different directions (horizontal, vertical, and diagonal), our proposed method dynamically allocates channels to descriptions in order to minimize the rate distortion of the video. The experimental results in presence of one description loss indicates that on average the videos at the rate of 1000 Kbit/s are reconstructed with Y-PSNR value of 36.2 dB. The dynamic allocation of descriptions to the network channels is optimized for rate distortion minimization. The improvement in term of Y-PSNR achieved by rate distortion optimization has been between 0.7 dB to 5.3 dB at different bit rates. A possible extension of the method is combining a feedback system for providing channel rates to the transmitter which makes the algorithm applicable in video streaming over peer-to-peer networks.

References

1. G. Conklin et al., "Video coding for streaming media delivery on the internet," *IEEE Trans. Circuits Syst. Video Technol.* **11**(3), 269–281 (2001).
2. J. Ohm, "Advances in scalable video coding," *Proc. IEEE* **93**(1), 42–56 (2005).
3. H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Sys. Video* **17**(9), 1103–1120 (2007).
4. C. Hewage et al., "Comparison of stereo video coding support in MPEG-4 MAC, H.264/AVC and H.264/SVC," in *Proc. 4th Visual Inf. Eng. Conf.*, IET, London (2007).
5. H. Jiang, "Experiment on post-clip FGS enhancement," ISO/IEC JTC1/SC29/WG11, MPEG00/M5826 (March 2000).
6. E. Akyol, A. M. Tekalp, and M. R. Civanlar, "A flexible multiple description coding framework for adaptive peer-to-peer video streaming," *IEEE J. Sel. Topics Signal Process.* **1**(2), 231–245 (2007).
7. R. Choupani, S. Wong, and M. Tolun, "Scalable video transmission over unreliable networks using multiple description wavelet coding," in *Proc. 7th Int. Conf. Digital Content, Multimedia Technol. Appl. (IDCTA2011)*, pp. 5–10, IEEE, Busan, Korea (2011).
8. Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE* **93**(1), 57–70 (2005).
9. R. Venkataramani, G. Kramer, and V. K. Goyal, "Multiple description coding with many channels," *IEEE Trans. Inf. Theory* **49**(9), 2106–2114 (2003).
10. V. K. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Process. Mag.* **18**(5), 74–94 (2001).
11. N. Franchi et al., "A space domain approach for multiple description video coding," in *ICIP*, Vol. 2, pp. 253–256, IEEE, Barcelona, Spain (2003).
12. R. Choupani, S. Wong, and M. Tolun, "Multiple description scalable coding for video transmission over unreliable networks," in *Embedded Computer Systems: Architectures, Modeling, and Simulation, 9th International Workshop*, pp. 58–67, LNCS, Samos, Greece (2009).
13. H. Kirchhoffer, H. Schwarz, and T. Wiegand, "CE1: simplified FGS," Joint Video Team, Doc. JVT-W090 (April 2007).
14. A. Segall, "CE 8: SVC-to-AVC bit-stream rewriting for coarse grain scalability," Joint Video Team, Doc. JVT-V035, Marrakech, Morocco (January 2007).
15. V. N. Padmanabhan, H. J. Wang, and P. A. Chou, "Resilient peer-to-peer streaming," in *Proc. 11th IEEE Int. Conf. Network Protocols*, pp. 16–27, IEEE, Atlanta, Georgia (2003).
16. T. B. Abanoz and A. M. Tekalp, "SVC-based scalable multiple description video coding and optimization of encoding configuration," *Signal Process. Image Commun.* **24**(9), 691–701 (2009).
17. M. Folli and L. Favalli, *Scalable Multiple Description Coding of Video Sequences*, Riunione annuale Gruppo Nazionale Telecomunicazioni e Teoria dell'informazione (GTTI08), Florence, Italy (2008).
18. V. D. Reguant et al., "Delivery of H264 SVC/MDC streams over Wimax and DVB-T networks," in *Proc. IEEE Int. Symposium Consumer Electronics, ISCE*, pp. 1–4, IEEE, Vilamoura, Portugal (2008).
19. A. R. Reibman et al., "Multiple description video coding using motion compensated temporal prediction," *IEEE Trans. Circuits Syst. Video Technol.* **12**(3), 193–204 (2002).
20. Y. Xu et al., "Multiple description coded video streaming in peer-to-peer networks," *Signal Process. Image Commun.* **27**(5), 412–429 (2012).
21. A. E. Mohr, E. A. Riskin, and R. E. Ladner, "Unequal loss protection: graceful degradation of image quality over packet erasure channels through forward error correction," *IEEE J. Sel. Areas Commun.* **18**(6), 819–828 (2000).
22. M. Wien, H. Schwarz, and T. Oelbaum, "Performance analysis of SVC," *IEEE Trans. Circuits Syst. Video Technol.* **17**(9), 1194–1203 (2007).
23. W.-J. Tsai and H.-Y. You, "Multiple description video coding based on hierarchical b pictures using unequal redundancy," *IEEE Trans. Circuits Sys. Video Technol.* **22**(2), 309–320 (2012).
24. Y. Andreopoulos et al., "Fully scalable wavelet video coding using inband motion ccompensated temporal filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 417–420, IEEE, Hong Kong (2003).
25. R. Choupani, S. Wong, and M. Tolun, "Adaptive embedded zero tree for scalable video coding," in *World Congress on Engineering (WCE2011)*, IAENG, London (2011).
26. J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Process.* **41**(12), 3445–3462 (1993).
27. Y. Xuguang and K. Ramchandran, "Optimal subband filter banks for multiple description coding," *IEEE Trans. Inf. Theor.* **46**(7), 2477–2490 (2000).
28. T. Tillo and G. Olmo, "A novel multiple description coding scheme compatible with the JPEG2000 decoder," *IEEE Signal Process. Lett.* **11**(11), 908–911 (2004).
29. T. Tillo, M. Grangetto, and G. Olmo, "Multiple description image coding based on lagrangian rate allocation," *IEEE Trans. Image Process.* **16**(3), 673–683 (2007).

30. M. R. Ardestani, A. A. B. Shirazi, and M. R. Hashemi, "Low-complexity unbalanced multiple description coding based on balanced clusters for adaptive peer-to-peer video streaming," *Signal Process. Image Commun.* **26**(3), 143–161 (2011).
31. I. V. Bajic and J. W. Woods, "Domain-based multiple description coding of images and video," *IEEE Trans. Image Process.* **12**(10), 1211–1225 (2003).
32. M. Weeks and M. A. Bayoumi, "Three-dimensional discrete wavelet transform architectures," *IEEE Trans. Signal Process.* **50**(8), 2050–2063 (2002).
33. Q. Chen et al., "Multivariate filter banks having matrix factorizations," *SIAM J. Matrix Anal. Appl.* **25**(2), 517–531 (2003).
34. X. You et al., "A blind watermarking scheme using new nontensor product Wavelet filter banks," *IEEE Trans. Image Process.* **19**(12), 3271–3284 (2010).
35. A. Jensen and A. la Cour-Harbo, *Ripples in Mathematics: The Discrete Wavelet Transform*, Springer Verlag, Germany (2001).
36. P. Yin et al., "Drift compensation for reduced spatial resolution transcoding," *IEEE Trans. Circuits Syst. Video Technol.* **12**(11), 1009–1020 (2002).
37. J. Reichel, H. Schwarz, and M. Wien, "Joint Scalable Video Model 11 (JSVM 11)," Joint Video Team, Doc. JVT-X202 Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Antalya, Turkey (January 2008).
38. R. Choupani, S. Wong, and M. Tolun, "A drift-reduced hierarchical wavelet coding scheme for scalable video transmissions," in *Proc. First Int. Conf. Adv. Multimedia (MMEDIA)*, pp. 68–73, IEEE, Colmar, France (2009).

Biographies and photographs of all authors are not available.

Chapter 5

Conclusions

This chapter summarizes the work presented in the previous chapters of this dissertation, and presents the future research directions. In Section 5.1 the summary of the main conclusions is presented. Section 5.2 includes the future research directions of the present work.

5.1 Summary

Chapter 1 "Introduction", introduces the video coding and its adaptation techniques. The chapter describes two challenges in video coding namely, video adaptation with varying network conditions while minimizing its quality degradation, and improving the robustness of videos against packet losses. The chapter briefly introduces the state-of-the-art methods used for solving these problems and presents our motivation in addressing these challenges. In this dissertation we aimed at addressing both challenges simultaneously as we believe that packet loss during video transmission impacts the video adaptation process drastically and degrades the quality of the video. The chapter clarifies the scope of the thesis and summarizes the contributions achieved.

Chapter 2, "Minimizing Drift Error in Scalable Video Coding", addressed the drift error which is an important source of quality degradation in video coding. The drift error is the result of error propagation during video reconstruction. In this chapter a new multi-layer scalable video coding method for optimizing the bit rate per pixel of the video was proposed which is robust against frame losses. The method reduces the drift error by re-organizing the frames in a hierarchical structure to restrict its propagation. In addition, the proposed method optimizes the amount of data stored in the base layer considering the location of the frame in a GOP. This optimization is based on the fact that the drift error and the rate of degradation due to frame data changes depend on the number of frames in a frame dependency chain. Hence, the quantization steps in base and enhancement layers are optimized to minimize this error. Besides the GOP length is also considered as an optimization parameter for the best overall performance in presence of frame loss/change.
Moreover, this chapter presented a solution for the drift error when the video is scrambled for privacy protection. The challenge in this case is the motion compensation of a block using an area from a preceding frame which can be scrambled for privacy. During the decoding process, if the private key for unscrambling the privacy protected area is not available, the motion compensated block will not be reconstructed correctly and hence, suffer from quality degradation which accumulates to the drift error. In order to avoid the drift problem, generally the privacy protected area is not used as reference area for the blocks of succeeding frames. However, representing the privacy protected area and

including its avoiding in the motion compensated video coding loop is an expensive and inefficient process. We propose a method to convert the scrambling operation into a matrix multiplication and let the motion compensation be performed using scrambled data. The proposed method eliminates the drift error completely and solves the privacy protected area representation problem.

Chapter 3, "Optimizing Multiple Description Video Coding in Spatial Domain", presents new method proposed for handling the data loss during the transmission of video streams. The proposed methods are based on combining SVC with MDC where the video is decomposed into spatial sub-streams in the first method, and SNR sub-streams in the second method. In both proposed methods, the error resilience of the video is increased. The proposed methods have the capability of being used as scalable coding methods in which any data loss or corruption is reflected as reduction in the quality of the video. However, except for the case when all descriptions are lost, the video streams do not experience jitter at play back. The method based on combining SNR with MDC defines a multi-layer structure for data in each description and provides the feasibility of reducing data rate by scaling down the video whenever the connection suffers from a low bandwidth problem. In order to measure the performance of the proposed coding methods, distortion rate imposed by data loss and scaling down for rate efficiency, have been utilized. An improvement was proposed for the well-known method of temporal decomposition of video into multiple description by putting odd and even numbered frames in different descriptions. This improvement increases the coding efficiency by grouping the frames in pairs where each frame in a pair is assigned to the description which results in smaller residue values and hence better compression rates. The proposed improvement preserves the balance in the descriptions and avoids creating large timing gaps between adjacent frames in a description.

Chapter 4, "Using Discrete Wavelet Transform for Optimizing Multiple Description Video Coding", includes optimizes the MDC methods using the properties of the DWT. Besides, the proposed methods includes adaptability features which enables error robust and network capacity adaptive video coding. The proposed method considered minimizing video quality degradation. In the first method we organized the coefficients in three different descriptions with a minimized common redundant part. Meanwhile, our proposed method provided the feasibility of having scalable transmission over each channel through encoding the coefficient is each description using a modified embedded zero-tree wavelet (EZW) zig-zag scanning. The arrangement of the descriptions requires a-priori information about the video content, and the channel conditions. For an optimized transmission and assuming that the frequency content of a video is not equally distributed in different directions (horizontal, vertical, and diagonal), our proposed method dynamically allocates channels to descriptions in order to minimize the rate distortion of the video.
The second important contribution presented in this chapter is a method for re-arranging the DWT coefficients in a way that the self-similarity between coefficients at different frequency levels can be utilized for estimating the values of the lost coefficients. The main idea here is that the coefficients at different frequency levels show similarity which can be utilize for estimating the missing data. Therefore, the proposed method groups the coefficients of the DWT transform in a way that the different level coefficients of each sub-band are encoded at different descriptions. In this way, whenever a description is lost the coefficients at the delivered descriptions are available for estimating the missing data using self-similarity property. The experimental results indicate that applying the method to still images provides better improvements that applying it to video sequences. This difference in the improvement rates stems from the fact that the inter-frame coding of the frames results in having a large number of zeros in DWT coefficients. This large number of zeros reduces the impact of self-similarity.

5.2 Future Research Directions

Video streaming over unreliable networks is subject to quality degradation in presence of packet loss. With the rapid growth in multimedia applications in hand-held devices such as smart phones and tablets which are communicating through wireless networks, the importance of developing an error resilient scalable coding solution for these applications has further increased. Interactive multimedia applications such as computer games, video on demand, tele-presence, and broadcast TV programs are some of the applications requiring special coding design. In many of these applications there exist repeating and redundant data. However, since the current video coding schemes utilize pixel based redundancy removal, coding efficiency is not very high. In this dissertation, different video coding methods have been proposed for providing scalability for video in presence of data loss or corruption. The proposed method can be extended in many different ways. Some possible future directions are:

- Different video coding and streaming application have different requirements. Visual surveillance for instance, requires low spatial but high temporal resolution for fast and accurate motion estimation. In a tele-conference application on the other hand, high spatial but low temporal resolution is desirable for texture analysis of the mostly uniform areas. An extension of the proposed methods can be applying them in a content-aware coding. For instance, applying scalable video coding techniques in a content-aware coding will utilize adaptive prediction error coding, adaptive quantization, etc. to cope with the video content for a better bit-per-pixel efficiency. Content-aware scalability can be implemented in frame level as well. For instance, the region of interest in a frame can be encoded with a higher resolution. In a tele-conference application the region defining a person can be assumed as the region of interest. In this respect, the content analysis and fast content segmentation methods will gain more importance. Some of the facts supporting the significance of the frame content for video encoding are as below:[1]
  - Human visual system cannot distinguish minor quality changes in videos. This fact indicates that video quality improvements in a given range, despite affecting the coding bit-rate, will not be comprehended by the end users.
  - The quality experienced by the human visual system is dependent on the object being viewed. For instance, a degradation in the quality of faces is noticed sooner than the same degradation in other objects. In addition, the noticeability of a quality change in human faces varies when the face is part the background in comparison with having the same face in the foreground.
  - The quality of experience (QoE) varies from a person to another.
- Different video contents are of different importance and therefore, require different quality and/or error protection levels. Adapting video and improving its error resilience can be carried out considering the video content importance from and end-user perspective such as:
  - Content aware coding can be combined with scalable coding and multiple description coding methods. In scalable coding of video, the decision about delivered video adaptation is made by the receiver. The content aware coding schemes can be developed in a way that the receiver can decide about the quality of each object present in the video. This means that the video content will be coded in scalable and with different error resilience levels.

[1] Keynote speech by Prof. Dr. C.-C. Jay Kuo at ICICS 2015 conference entitled "Reflection on Image/Video Coding: Where Do We Go from Here?"

The end-user decides which part should be received at which quality, and with what level of fidelity. These types of video coding methods should consider providing scalability and error resilience in texture, shape, depth, etc.

  - Content-aware video coding can be extended to multiple description coding schemes. In the present work, some improvements and novelties to MDC methods have been proposed for decomposing the video in spatial, temporal, and bit-per-pixel (signal-to-noise ratio) domain. However, these decompositions do not consider the frame or video contents from an application perspective. In a tele-conference application for instance, protecting the more important parts of video such as human faces against packet loss errors is more important than protecting the background. This requires the awareness of the video encoder of the contents of the frames being encoded. However, content-aware encoding of the video requires challenging and expensive image processing operations to locate and segment the regions of interest, score them considering to the requirements of the application, and encode them accordingly.
  - In order to obtain the best quality of service and/or quality of experience (QoS/QoE) in terms of low end-to-end delay, and minimum quality degradation, content-aware adaptation of video is necessary. However, in broadcasting or multicasting video and considering the dynamic nature of the Internet, the adaptation cannot be performed at a single point and distributed adaptation is required. Although, a number of projects are going on in this regard [12], the problem still poses a challenge.

- Multiple video data sources in the form of multi-camera systems are becoming very common. These systems provide visual data through cameras which are not widely apart. This means that the videos captured by these cameras contain a large amount of overlapping and hence redundancy. This redundancy can be utilized for reconstructing video when a packet loss happens in a transmission. The similarity between the methods developed in this research for decomposing video into multiple descriptions and the multiple views captured by the multi-camera systems, can be utilized for extending our proposed methods to multi-camera systems. However, the videos captured by a multi-camera system are taken from different view angles and the reconstruction of one them using the other videos will be quite challenging if the view angle differences are large.

Bibliography

[1] Video coding for low bitrate communication; DRAFT ITU-T recommendation H.263.

[2] E. Akyol, A. M. Tekalp, and M. R. Civanlar. A flexible multiple description coding framework for adaptive peer-to-peer video streaming. *IEEE Journal of Selected Topics in Signal Processing*, 1:231–245, 2007.

[3] S. F. Chang and D.G. Messerschmidt. Manipulation and compositing of MC-DCT compressed video. *IEEE Journal of Selective Areas Communications*, 13:1–11, 1995.

[4] M. L. Comer, K. Shen, and E. J. Delp. Rate-scalable video coding using a zerotree wavelet approach. *Proceedings of the Ninth Image and Multidimensional Digital Signal Processing Workshop*, pages 162–163, 1996.

[5] E. J. Delp, P. Salama, E. Asbun, M. Saenz, and K. Shen. Rate scalable image and video compression techniques. *Proceedings of the* 42^{nd} *Midwest Symposium on Circuits and Systems*, 1:635–638, 1999.

[6] M. Domanski, A. Luczak, and S. Mackowiak. Spatial-temporal scalability for MPEG video coding. *IEEE Transaction on Circuits and Systems for Video Technology*, 10:1088–1093, 2000.

[7] I. Unanue et.al. A tutorial on H.264/SVC scalable video coding and its tradeoff between quality, coding efficiency and performance. *Recent Advances on Video Coding, INTECH open access publisher*, 2011.

[8] M. Ghanbari. Two-layer coding of video signals for VBR networks. *IEEE Journal of Selected Areas Communications*, 7:771–781, 1989.

[9] H. Gharavi and M. H. Partovi. Multilevel video coding and distribution architectures for emerging broadband digital networks. *IEEE Transaction on Circuits and Systems for Video Technology*, 6:459–469, 1996.

[10] V. K. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18:74–94, 2001.

[11] V.K. Goyal. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18:9–21, 2001.

[12] M. Grafl, C. Timmerer, H. Hellwagner, D. Negru, E. Borcoci, D. Renzi, A.-L. Mevel, and A. Chernilov. Scalable video coding in content-aware networks: Research challenges and open issues. *Trustworthy Internet*, pages 349–358, 2011.

[13] A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.

[14] International Telecommunication Union (ITU-T). *ITU-T Recommendation H.263 Version 2: Video Coding for Low Bit Rate Communication (H.263+)*. 1998.

[15] M. Karczewicz J. Ridge, Y. Bao and X. Wang. Fine-grained scalability for H.264/AVC. *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications*, pages 247–250, 2005.

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

[17] H. Katata, N. Ito, and H. Kusao. Temporal-scalable coding based on image content. *IEEE Transaction on Circuits and Systems for Video Technology*, 7:52–59, 1997.

[18] F. Ishtiaq L. P. Kondi and A. K. Katsaggelos. On video SNR scalability. *International Conference on Image Processing (ICIP)*, 3:934–938, 1998.

[19] W. Li. *Fine granularity scalability using bit-plane coding of DCT coefficients*. ISO/IEC JTC1/SC29/WG11, MPEG98/M4204, 1998.

[20] W. Li. Overview of fine granularity scalability in MPEG-4 video standard. *IEEE Transaction on Circuits and Systems for Video Technology*, 11(3):301–317, 2001.

[21] W. Li and Y. Chen. *Experiment result on fine granularity scalability*. ISO/IEC JTC1/SC29/WG11, MPEG99/M4473, 1999.

[22] W. Li, F. Ling, and H. Sun. *Bitplane coding of DCT coefficients*. ISO/IEC JTC1/SC29/WG11, MPEG97/M2691, 1997.

[23] R. Mathew and J. F. Arnold. Layered coding using bitstream decomposition with drift correction. *IEEE Transaction on Circuits and Systems for Video Technology*, 7:882–891, 1997.

[24] S. O. Mietens. *Complexity scalable MPEG encodings*. PhD thesis, Technische Universiteit Eindhoven, 2004.

[25] R. Mohan, J.R. Smith, and S. Li. Adapting multimedia internet content for universal access. *IEEE Transaction on Multimedia*, 1(1):104–114, 1999.

[26] F. Pereira and I. Burnett. Universal multimedia experiences for tomorrow. *IEEE Signal Processing Magazine*, 20:63–73, 2003.

[27] E.C. Reed and J.S. Lim. Optimal multidimensional bit-rate control for video communications. *IEEE Transaction on Image Processing*, 11(8):873–885, 2002.

[28] I. E. G. Richardson. H.264 and MPEG-4 video compression. *Wiley*, 2003.

[29] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):243–250, 1996.

[30] J. M. Shapiro. Embedded image coding using zerotrees of wavelets coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

[31] K. Shen and E. J. Delp. Wavelet based rate scalable video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1):109–122, 1999.

[32] W. Tan and A. Zakhor. Real-time internet video using error resilient scalable compression and TCP-friendly transport protocol. *IEEE Transaction on Multimedia*, 1(2):172–186, 1999.

[33] D. Taubman and A. Zakhor. Multirate 3-D subband coding of video. *IEEE Transactions on Image Processing*, 3(5):572–588, 1994.

[34] R. Venkataramani, G. Kramer, and V.K. Goyal. Multiple description coding with many channels. *IEEE Transaction on Information Theory*, 49:2106–2114, 2003.

[35] Q. Wang and M. Ghanbari. Scalable coding of very high resolution video using the virtual zerotree. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(5):719–727, 1997.

[36] Y. Wang, M.T. Orchard, and A.R. Reibman. Multiple description image coding for noisy channels by pairing transform coefficients. *IEEE Workshop on Multimedia Signal Processing*, pages 419–424, 1997.

[37] Y. Wang, A. R. Reibman, and L. Shunan. Multiple description coding for video delivery. *Proceedings of IEEE*, 93:57–70, 2005.

[38] L. Weiping. Overview of fine granularity scalability in MPEG-4 video standard. *IEEE Transaction on Circuits and Systems for Video Technology*, 11(3):301–317, 2001.

[39] D. Wilson and M. Ghanbari. Exploiting interlayer correlation of SNR scalable video. *IEEE Transaction on Circuits and Systems for Video Technology*, 9:783–797, 1999.

[40] D. Wu, Y. Hou, W. Zhu, Y.Q. Zhang, and J. Peha. Streaming video over the internet: Approaches and directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3):282–300, 2001.

List of Publications

1. R Choupani, S Wong, MR Tolun, "Hierarchical SNR Scalable Video Coding with Adaptive Quantization for Reduced Drift Error", VISAPP (1), 117-123, 2015.

2. R Choupani, S Wong, MR Tolun, "Using wavelet transform self-similarity for effective multiple description video coding",Information, Communications and Signal Processing (ICICS), 1-5, 2015.

3. R Choupani, S Wong, MR Tolun, "Drift-free video coding for privacy protected video scrambling", Information, Communications and Signal Processing (ICICS), 1-5, 2015.

4. R Choupani, S Wong, MR Tolun, "Multiple description coding for SNR scalable video transmission over unreliable networks", Multimedia Tools and Applications 69(3), 843-858, 2014.

5. R Choupani, S Wong, MR Tolun, "Spatial multiple description coding for scalable video streams", International Journal of Digital Multimedia Broadcasting, 2014.

6. R Choupani, S Wong, MR Tolun, "Optimized Multiple Description Coding for Temporal Video Scalability", Advances in Computational Science, Engineering and Information Technology, Springer International Publishing, 167-176, 2013.

7. R Choupani, S Wong, MR Tolun, "Unbalanced multiple description wavelet coding for scalable video transmission", Journal of Electronic Imaging 21(4), 2012.

8. R Choupani, S Wong, MR Tolun, "Multiple Description Wavelet Coding for Scalable Video Transmission", 7th International Conference on Digital Content, Multimedia Technology and its Applications (IDCTA 2011), 16-18, 2011.

9. R Choupani, S Wong, MR Tolun, "Scalable video transmission over unreliable networks using multiple description wavelet coding", Digital Content, Multimedia Technology and its Applications (IDCTA), 5-10, 2011.

10. R Choupani, S Wong, MR Tolun, "A drift-reduced hierarchical wavelet coding scheme for scalable video transmissions", IEEE Conference on Advances in Multimedia, MMEDIA'09, 68-73, 2009.

11. R Choupani, S Wong, MR Tolun, "Multiple description scalable coding for video transmission over unreliable networks", International Workshop on Embedded Computer Systems, 58-67, 2009.

12. R Choupani, S Wong, MR Tolun, "Weighted Embedded Zero Tree for Scalable Video Compression", IPCV, 681-684, 2008.

13. R Choupani, S Wong, MR Tolun, ”Main Issues in Scalable Video Coding: A Review”, IPCV, 497-505, 2007.

Sammenvatting

Met de snelle verbeteringen in de digitale communicatie-technologien, heeft de distributie van hoge kwaliteit visuele informatie die sneller verspreid raken. Maar de beschikbare technologien niet voldoende waren om de stijgende vraag naar hoge kwaliteit video te ondersteunen. Deze situatie wordt verder gecompliceerd wanneer het netwerk hulpbronnen zoals beschikbare bandbreedte fluctueert of pakketverlies plaatsvinden tijdens transmissie. In dit proefschrift presenteren wij een aantal video compressietechnieken die in staat zijn aan te passen aan de wisselende netwerk omstandigheden. Wij richten ons zowel uitdagingen namelijk de fluctuaties in de beschikbare middelen, zoals de bandbreedte, rekenkracht en pakket verliezen. Deze problemen op zijn beurt vertaalt zich in degradatie van de waargenomen afspelen van video als jitter en vertraging voordat het afspelen van video begint. Daarom concentreren wij ons op het ontwikkelen van robuuste en snelle adaptieve videocodering regelingen die nodig zijn voor het omgaan met de veranderingen in de fysieke eigenschappen van de communicatienetwerken. Wij presenteren een nieuwe meerlaags werkwijze voor scalaire videocodering (SVC) voor het optimaliseren van de bit per beeld element van de videocomponent die robuust tegen pakketverlies. De werkwijze vermindert de kwaliteit afbraak in aanwezigheid van gegevensverlies door reorganisatie van de frames in een hirarchische structuur en verbeteren de videokwaliteit door middel van ontleden elk frame correct aan de foutvoortplanting beperkt. Bovendien presenteren wij een oplossing voor de kwaliteit degradatie in video reconstructie wanneer de video wordt gecodeerd voor bescherming van de privacy. Wij hebben ook twee werkwijzen op basis van meerdere description videocodering (MDC) om pakketverlies in netwerken handvat met een hoge mate van transmissiefout. De voorgestelde werkwijzen zijn gebaseerd op het combineren SVC met MDC tot ontleden van het video in ruimtelijke deelstromen in de eerste werkwijze en SNR deelstromen in de tweede werkwijze. In beide voorgestelde methoden, wordt de fout veerkracht van de video toegenomen. De voorgestelde methoden hebben het vermogen om te worden gebruikt als SVC methoden waarbij enig verlies van gegevens of corruptie vermindert de kwaliteit van de video in een geminimaliseerde wijze en met uitzondering van het geval wanneer alle beschrijvingen zijn verloren, niet de video streams niet jitter ervaren in het afspelen. De voorgestelde methoden zorgen voor de haalbaarheid van een verlaging datasnelheid door de afbouw van de video, wanneer de verbinding kampt met een lage bandbreedte probleem. Wij stellen ook voor discrete wavelettransformatie (DWT) op basis optimalisaties voor MDC. Een belangrijk nadeel bij MDC werkwijze is hun inefficintie in termen van bits per pixel dat een gevolg is van het behoud correlatie tussen ontleed videosegmenten. Wij stellen een methode op basis van de zelf-gelijkenis tussen DWT cofficinten op verschillende frequentieniveaus de codeerefficintie van DWT MDC gebaseerde te verbeteren. Bij de voorgestelde werkwijze, wanneer een beschrijving verloor de cofficinten van het geleverde beschrijvingen worden gebruikt voor het schatten van de ontbrekende data via zelfgelijkvormigheid eigenaardig.

Curriculum Vitae

Roya Choupani was born in 1969 in Iran. After receiving her BS degree in Computer Engineering she moved to Ankara, Turkey in 1998. She received her MS degree in Computer Engineering from Cankaya University in 2002. Since 2002 she has been working as an instructor in the computer engineering department of Cankaya University in Ankara, Turkey.
In 2007 she joined the computer engineering laboratory of the Technical University of Delft where she started working on her thesis under the supervision of Professor Stephan Wong. Her main fields of interests are visualization, multimedia, and video coding.