ADAPTIVE SCALABLE VIDEO CODING

By

Emrah Akyol

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
The Degree of

Master of Science

in

Electrical-Computer Engineering

Koç University

September 2005

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Emrah Akyol

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. A. Murat Tekalp (Advisor)

_____

Prof. M. Reha Civanlar (Co-advisor)

_____

Assist. Prof. Uluğ Bayazıt

_____

Assist. Prof. Oğuz Sunay

_____

Assist. Prof. Yücel Yemez

Date: _____

To my parents

ABSTRACT

This thesis is composed of three main parts which include three contributions in slightly different fields, all lying on the same framework: Adaptive Scalable Video Coding. First part is about integration of motion compensated temporal filtering (MCTF), the basis for temporal scalability in scalable video coding methods, to the latest non-scalable video compression standard, i.e., H.264/AVC. We propose a GOP structure to implement block-based adaptive MCTF within the H.264/AVC syntax using stored B-pictures, similar to the motion-compensated 5/3 wavelet filtering. We provide experimental results to compare the results of our proposed codec with those of other scalable wavelet video coders which use MCTF. The proposed scheme is also integrated into H.264/AVC reference software as 'Hierarchical B pictures' or 'Temporal Pyramid' and it is currently under investigation of MPEG Core Experiments for the upcoming Scalable Video Coding standard (SVC).

Secondly, we worked on content adaptive scalability type selection problem. State of the art scalable video coders provide different options, such as temporal, spatial and SNR scalability, where bitrate reduction using each scalability type results in different kinds and/or levels of visual distortion depending on the content and the bitrate. In most cases, a single scalability type does not fit the whole video well, and scaling option selection can be optimized for each temporal segment depending on the content of that segment and the target bitrate. This dependency between selection of scalability type, video content, and bitrate is not well investigated in the literature. In this work, assuming that the video is temporally segmented by some content analysis scheme, we propose a method to choose the best scaling type for each temporal segment that results in minimum visual distortion among temporal, spatial and SNR scalability for fully embedded scalable video coders. We employ an objective distortion measure that consists of a linear combination of four component measures, which are a flatness measure, a blockiness measure, a blurriness measure, and a temporal jerkiness measure, to quantify artifacts caused by bitrate reduction by spatial size reduction, frame rate reduction, and quantization parameter scaling. Two subjective tests have been performed to validate the proposed procedure for shot-based selection of optimal scalability type on soccer videos. Soccer videos whose bitrate are reduced from 600 kbps to 100-300 kbps by the proposed content-adaptive selection of scalability

type have been deemed visually superior to those whose bitrates are reduced by a single scalability option for the entire test sequence.

Finally, we worked on adaptive peer-to-peer (P2P) streaming using scalable multiple description coding. Efficient P2P video streaming is a challenging task due to time-varying nature of both the number of available peers and network/channel conditions. To this effect, we propose i) a new flexible scalable multiple description coding (MDC) method, where the number of descriptions, and the rate and redundancy level of each description can be adapted on the fly (by post-processing of a fully-embedded scalable coded bitstream), and ii) a new adaptive TCP Friendly Rate Controlled (TFRC) P2P streaming system based on this new MDC scheme. The optimization of the design parameters of the proposed MDC scheme according to network conditions is discussed within the context of the proposed adaptive P2P streaming framework, where the number and quality of available streaming peers/paths are a priori unknown and vary in time. Experimental results, by means of NS-2 network simulation of a P2P video streaming system, show that adaptation of the number and rate of descriptions/layers and the redundancy level of each description according to network conditions yields significantly superior performance when compared to other scalable MDC schemes using a fixed number of descriptions/layers with fixed rate and redundancy level.

ÖZETÇE

Bu tez ‘Uyarlanır Ölçeklenebilir Video Kodlama’ genel konusu altında üç ana başlıkta hazırlanmıştır. Birinci çalışmada etkili zamansal ölçeklenebilirlik sağlanması amacıyla H.264 standardı içerisinde devinim dengeli zamansal filtreleme(DDZF) öneriyoruz. DDZF geleneksel olarak dalgacık dönüşümüyle yapılan tam ölçeklenebilir video kodlamasında kullanılır. Ancak devinim dengeli 5-3 dalgacıkları kaldırma işlemi yapılarak filtreleme görüntü değişimi olan yerlerde ve video çerçevelerinde yeni çıkan bölgelerin kodlanmasında başarısız olmaktadır. H.264 standardı iki yönlü devinim dengeleme için uyarlanabilir blok büyüklüğü, ileri-geri ve iki yönlü modlar arasında uyarlanabilir mod seçimi,bloksuzlaştırma filtresi ve örtüşmeli devinim dengeleme gibi gelişmiş tekniklere sahiptir. Bu nedenle devinim dengeli 5-3 dalgacık filtresine benzer şekilde H.264 standardı içerisinde blok tabanlı uyarlanabilir DDZF uygulamak için bir görüntü grubu yapısı öneriyoruz. Diğer DDZF tabanlı dalgacık dönüşümü video kodlayıcıların sonuçlarıyla birlikte karşılaştırmak üzere sonuçlarımızı sunuyoruz. Önerdiğimiz DDZF yapısı ‘Sıralamalı B-Resimleri’ ya da ‘ Zamansal Piramit’ ismiyle H.264/AVC referans yazılımına da dahil edilmiştir.

İkinci çalışmada ölçeklenebilir video kodlamada içeriğe bağlı en iyi ölçekleme operatörü seçimi üzerinde çalışılmıştır. Ölçeklenebilir video kodlayıcıları, her biri içeriğe ve bit-hızına bağlı olarak değişik tipte ve miktarda bozuluma neden olan zamansal, uzaysal ve kalitesel olmak üzere üç çeşit ölçeklenebilirlik olanağı sağlamaktadır. Genelde bir tek ölçekleme operatörü videonun bütün kısımları için uygun olmamaktadır; bu nedenle videonun değişik içerikteki her bir parçası için ölçekleme operatörü o parçanın içeriğine bağlı olacak şekilde değiştirilmelidir. Bu çalışmada, video bir içerik inceleme metoduyla içeriğine bağlı olarak değişik kısımlara ayrılmış kabul edilmiş ve her bir zamansal video parçası en düşük bozunuma sonuç veren en-iyi ölçekleme operatörüyle ölçeklenmiştir. Bit-hızı azalımı, uzaysal genişlik değişimi ve zamansal ölçeklemenin yarattığı bozulum, düzlük, blokluluk, zamansal atlama ve bozunukluluk metrikleriyle ölçülmüştür. En-iyi ölçekleme operatörü ayrı bozulum metriklerinin lineer kombinasyonuyla oluşturulan genel bozulum metriğine göre en düşük bozulumu veren operatör olarak bulunmuştur. Bu lineer kombinasyonun katsayıları içeriğe göre ayarlanarak bulunmuştur. Önerilen bozulum metriği ve en-iyi operatör bulma prosedürü futbol videolarıyla iki öznel test yapılarak gerçeklenmiştir.

Üçüncü kısımda ise içeriğe ve kanal koşullarına uyarlanır çok-tanımlamalı video kodlama yöntemi ile uyarlanır video iletimi üzerinde çalışılmıştır. İletişim kanallardaki sıkışmanın neden olduğu paket kayıpları ve gecikme değişimleri, gecikmeye duyarlı multimedya akışı işlemlerini zorlaştırmaktadır. Çok tanımlı video kodlama yöntemleri ile paket kayıplarının yarattığı bu etki azaltılabilmektedir. Ancak bu zamana kadar geliştirilen çok tanımlı video kodlama teknikleri, kanal koşullarına uyum sağlayamamakta, zaman içerisinde tanım sayısının, tanımların içerisine eklenen gereksiz bit miktarinin ve her tanım için harcanan bit miktarının değiştirilmesine izin vermemektedir. Önerilen çok tanımlı video kodlama tekniği bütün bahsedilen değişikliklere olanak sağlamakta ve birçok çok tanımlı video kodlama tekniğinden daha iyi sıkıştırma performansı sergilemektedir. Bu çalışmada önerilen sıkıştırma tekniği diğer benzer tekniklerle birçok değişken koşulda karşılaştırılmış, önerilen tekniğin diğer tekniklere hem sağladığı çok yönlü kanala uyarlanabilme özelliği açısından hem de sıkıştırma performansı /video görüntü kalitesi (PSNR) açısından üstün olduğu gösterilmiştir.

ACKNOWLEDGEMENT

I am truly indebted to my advisors Prof. A. Murat Tekalp and Prof. M. Reha Civanlar for their great support as well as their insightful and stimulating ideas during my graduate studies. Studying under their supervision was a very enjoyable and productive experience that I learned many lessons I will benefit through my academic and personal life.

I would like to thank to Prof. Uluğ Bayazıt, Prof. Oğuz Sunay and Prof. Yücel Yemez for serving in my thesis committee.

Special thanks to all of my friends in College of Engineering for the supportive and friendly atmosphere that I enjoyed during my stay at Koc University.

Last but not the least, I should thank to my parents who provided every opportunity for my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

Although video compression and streaming have experienced phenomenal growth since the introduction of first video compression methods and commercial streaming products, there still remain many challenges to be addressed to achieve resilient and efficient video delivery over unreliably varying environments like the Internet and wireless channels. The difficulty comes from the fact that both channel characteristics and video content vary in time which requires adaptation of encoding and streaming techniques to network and video content. Recently, adaptive solutions have attracted attention of several researchers. Content-adaptive mode-selection in video encoders [1] and motion adaptive update step of the motion compensated temporal filtering[3] are examples of recently proposed content adaptive solution approaches. Channel adaptive streaming has already flourished as an area in itself with many recently popular research topics like optimal forward error correction (FEC) assignment in lossy environments [4] , rate-distortion optimal channel adapted video streaming[3], optimal redundancy setting in multiple descriptions coding [6], optimal mode switching in lossy networks [5] etc. In this thesis, we propose adaptive video coding using scalable video coding due to the efficient adaptability of scalable coders.

Recently, scalable video coding has gained renewed interest since it has been shown [7] that it can achieve compression efficiency that is comparable to that of H.264[8] One of the early important findings in scalable video coding is the usage of non-recursive motion compensation in subband video coding framework [9]. In [11], MC-3D subband coding framework is advanced by providing optimum rate allocation for entropy coding of subbands which provides superior compression efficiency. The lifting implementation of wavelet transform on motion aligned temporal frames is

first proposed in [12][16]. This implementation allows efficient motion compensation while performing temporal filtering on motion aligned temporal frames.

After motion compensated temporal filtering (MCTF), which provides temporal scalability, a spatial wavelet transform can be applied to the resulting high and low frequency frames to obtain spatial scalability. All the subbands can, then, be encoded using an embedded entropy coder to obtain SNR scalability. General flow of a wavelet video coding system can be seen in Figure 1. MCTF based scalable video coding frame work is advanced by incorporating half pixel motion compensation and using longer wavelet filters [11] to obtain more accurate motion compensation and more efficient decorrelation of frames respectively. With these advances, scalable video coders achieve comparable coding efficiency to the that of state-of art predictive video coder, H.264 standard. Detailed survey of recent enhancements in MCTF based scalable video coding can be found in [17].

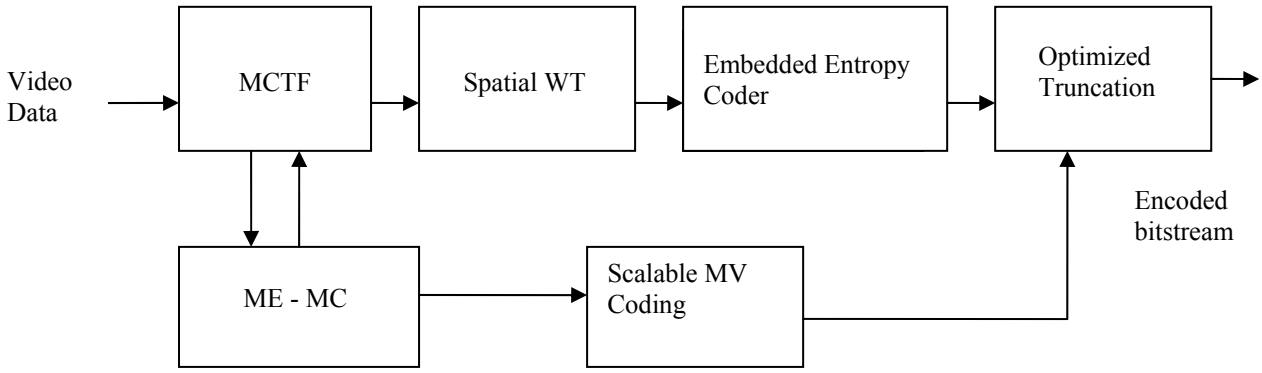

Figure 1: General structure of an MCTF based fully scalable wavelet video coder with scalable motion representation

Since scalable video coders can be considered in the context of adaptive video coding due to their flexible nature that provides efficient adaptation to network conditions by post processing the encoded bitstream, we concentrate on scalable video coding methods to solve the problems arise in adaptive video coding framework. Scalable video coding indeed provides a very flexible framework to adapt its tunable parameters post encoding with its ‘encode once truncate with different parameters’ property.

1.2 Contributions

We have worked on three slightly different subjects in adaptive scalable video coding framework. First, we analyzed motion compensated temporal filtering (MCTF) structure and observed that motion-compensated simple 5/3 lifted temporal wavelet filtering suffers at scene changes, as well as occlusion regions. We note that the bi-directional motion compensation mode in the H.264 standard is best equipped with the state of the art adaptive features such as adaptive block size, mode switching between forward, backward and bidirectional prediction, and in-loop deblocking filter. Hence, we propose a GOP structure to implement block-based adaptive MCTF within the H.264 syntax using stored B-pictures, similar to the motion-compensated 5/3 wavelet filtering [18].

Secondly, we worked on content adaptive scalability type selection problem [19][20]. State of the art scalable video coders provide different options, such as temporal, spatial and SNR scalability, where bitrate reduction using each scalability type results in different kinds and/or levels of visual distortion depending on the content and the bitrate. In most cases, a single scalability type does not fit the whole video well, and scaling option selection can be optimized for each temporal segment depending on the content of that segment and the target bitrate. This dependency between selection of scalability type, video content, and bitrate is not well investigated in the literature. In this work, assuming that the video is temporally segmented by some content analysis scheme, we propose a method to choose the best scaling type for each temporal segment that results in minimum visual distortion among temporal, spatial and SNR scalability for fully embedded scalable video coders. We employ an objective distortion measure that consists of a linear combination of four component measures, which are a flatness measure, a blockiness measure, a blurriness measure, and a temporal jerkiness measure, to quantify artifacts caused by bitrate reduction by spatial size reduction, frame rate reduction, and quantization parameter scaling. Coefficients of the linear combination are adapted to temporal segment (shot) content type by a training procedure. We then define the best scaling option for each shot as the one with the minimum objective distortion score.

Thirdly, we worked on scalable multiple description video coding for adaptive peer-to-peer streaming.[21][22] Multiple description coding (MDC) addresses the problem of encoding source information using more than one independently decodable complementary bitstreams, which, when combined, can provide the highest level of quality and when used independently, can still provide an acceptable level of quality. This is made possible by introducing some redundancy in each description, which will be discarded if all streams are received. It is well known that MDC can provide robust video communication over unreliable networks, such as Internet or wireless networks, by utilizing path/server diversity at the cost of reduced compression efficiency. Providing a variable (flexible) number of descriptions post encoding becomes an important concern in a peer-to-peer (P2P) video streaming, where the number of available "good" source peers is not known a priori. To this effect, we propose a novel scalable multiple description video coding framework, which enables varying

i) the number of descriptions,

ii) the rate of each individual description, and

iii) redundancy level of each description

on the fly **(**i.e., post encoding**)**. These properties of the coder enable efficient adaptation to network conditions. By using the new MDC scheme, a new adaptive TCP Friendly Rate Controlled (TFRC) P2P streaming system is proposed. The optimization of the design parameters of the proposed MDC scheme according to network conditions is discussed within the context of the proposed adaptive P2P streaming framework, where the number and quality of available streaming peers/paths are a priori unknown and vary in time.

Chapter 2

MCTF WITHIN H.264/AVC STANDARD

2.1 Motivation

Motion-compensated lifted temporal wavelet filtering has been reported as a very effective approach for building scalable video codecs in the literature [16]. The basic idea behind this approach is to interpolate frames from their neighboring (past and future) frames in time domain using motion compensation. Recent predictive coders such as H.264 [8]have advanced features for bidirectional prediction like adaptive block sizes, mode switching between forward, backward and bidirectional prediction, deblocking filter and intra-coded macroblocks in inter frames. These features of the predictive coders should prove to be useful in the MCTF structure to obtain better prediction and hence better compression efficiency.

In classical lifting scheme, every predicted frame is computed as the average of forward and backward predictions. This averaging results in worse prediction than only forward or backward prediction especially when a scene change occurs in a group of pictures. Adaptive mode switching between forward, backward and bidirectional prediction can make this scheme to avoid such problems. Deblocking filter decreases the blockiness of the prediction which is an inherent problem of block based motion estimation. Adaptive macroblock size and overlapped motion compensation significantly increase the motion compensation prediction quality. The fact that all of these advanced motion compensation features are part of the H.264 syntax motivates us to implement an MCTF structure within the H.264 standard.

Given the rich prediction feature set of the H.264 standard, we target implementing an MCTF approach within the H.264 standard to obtain an efficient, easy to produce and effective layering

scheme without modifying the standard. Achieving this target is possible by using the stored B pictures which can be used as reference frames for other pictures in H.264

2.1 Motion Compensated Temporal Filtering

Motion compensated lifted temporal wavelet filtering performs temporal biorthogonal wavelet transform on frames using lifting (that is, prediction) and update steps. Although many biorthogonal wavelet kernels can be used, 5/3 wavelet kernel is reported to have the best experimental performance [16]. Implementation of the motion compensated lifting scheme with 5/3 filters for a GOP size of 4 frames is shown in Fig-2. In the prediction step, frames are predicted from their nearest neighbors using motion compensation. In the update step, the reference frames are temporally filtered to prevent aliasing due to subsampling. Motion compensation is also used in the update step, but the direction is reversed. First frame of every GOP, which is intra coded, and the prediction errors are then encoded usually using spatial wavelets. In the encoding part, original frames rather than decoded frames are used.

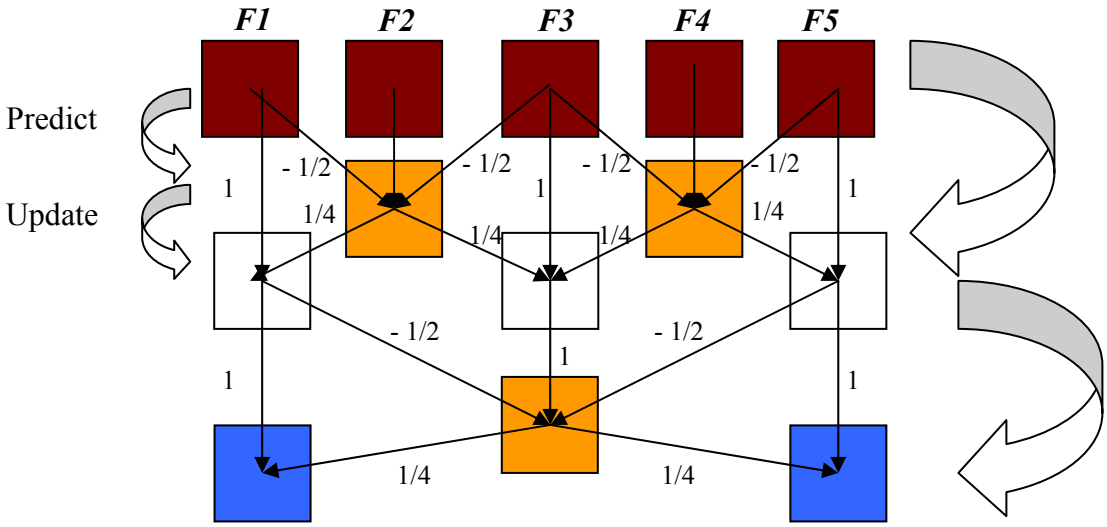

Figure 2: 5-3 lifting scheme (5-3 MCTF) for GOP = 4 .MCTF is applied on frames F1, F2, F3, F4, F5. Resulting high pass frames (H1,H2,H3) , and low pass frames (L1,L2) are obtained by combining the motion compensated sources with the coefficients indicated.

In this approach, existence of significant motion does not affect the compression performance when motion is compensated effectively. If the motion field used in the motion-compensated lifting step is invertible, the update step does not require new motion vectors. Since sending second set of motion vectors would be very costly, the update step can be performed with the inverse of the motion vectors obtained in the prediction step. However, when the motion is not invertible, the motion vectors will not be correct, significantly deteriorating the compression performance.

When mesh based motion estimation is used in the lifting step, the motion vectors for the update step can be obtained by straightforward inversion [16]. However one-to-one prediction fails when uncovered areas appear in the video sequence. Block based motion estimation is not invertible so the update step is either performed with motion vectors inaccurately obtained by inverting the motion vectors in the lifting step [1] or is not performed at all [3]. Bypassing the update step is reported to achieve better compression performance than using non-exact motion vectors [3].

2.2 MCTF within H.264

In our configuration, we encode the first frame of a GOP as an intra frame and all others as B frames as shown in Fig. 3 for a GOP consisting of eight frames.

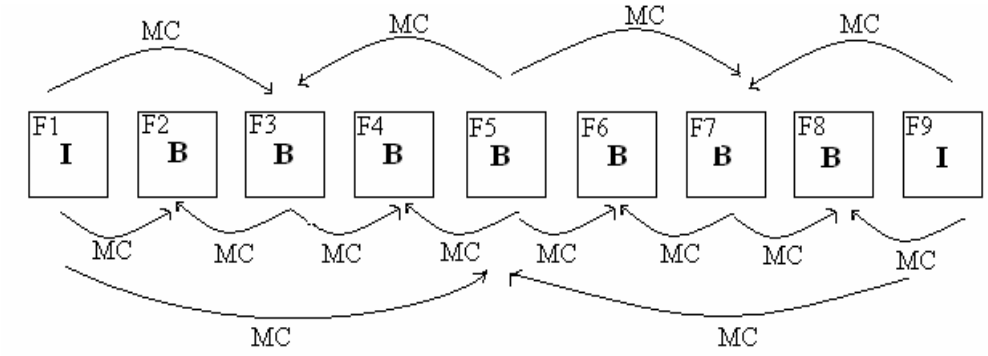

Figure 3: H.264/AVC configuration of lifting scheme with GOP=8

Here, frame **F5** is coded as a B frame estimated from frames **F1** and **F9**. **F5** is used as one of the reference frames for frames **F3** and **F7** which are coded as B pictures also. Frames **F2**, **F4**, **F6**, **F8** are encoded as B frames with reference to neighboring I or B frames. The H.264 syntax permits the use of B frames as reference frames with the feature called *stored B-pictures* [4].

This scheme provides adaptive multilayer temporal scalability to the H.264 compression standard. As an example, for the above configuration, **F1**, **F5** and **F9** are members of the base layer. **F3** and **F7** constitute the first enhancement layer and, the rest of the frames constitute the second enhancement layer, thus providing a three layer bitstream. It is known that an H.264 bitstream with missing B pictures should be decodable by a standard decoder. Hence, the described approach produces compliant bitstreams as long as the bits corresponding to the intermediate frames are inserted into their correct place before decoding. The coding efficiency may be less than the ideal case where the best reference frame is used for each prediction but, this loss can be reduced by appropriate selection of the GOPs. Flexibility in choosing the number of layers in every GOP provides better adaptation to varying bitrates.

2.3 Comparative Results

We compared the results of our approach with other MCTF based scalable video coders. In [13], mesh based motion estimation is used in bidirectional motion compensation. In [23], the inverse of forward motion vectors are used as the backward motion vectors required for update step. In [14], the update step is not implemented. MC_EZBC is the motion compensated embedded zero block video coder which was reference coder of MPEG at the time of this work [11].

The results show that our configuration in H.264 outperforms other scalable video coders based on MCTF. Of course, other MCTF implementations may support SNR and spatial scalability, while the proposed H.264 based encoder supports only temporal scalability. This difference is mainly caused by the advanced motion compensated prediction features of the H.264 standard. The PSNR difference gets larger as the bitrate increases in the comparison with other block based 5-3 lifted wavelet video coders. This can be explained by the amount of side information that encoders send. In H.264 standard, the side information costs more because of several encoding modes, flexible macroblock sizes and motion vectors with 1/4 pixel accuracy. Other scalable coders that we used to compare our results, however, have much lighter side information, reducing their overheads.

Table-1 :Block based without update step [14]vs. MCTF in H.264

Foreman QCIF

Block based without update MCTF in H.264

Bitrate(kbps)	105.56	181.59	333.66	Bitrate(kbps)	101.3	155.2	268.7
PSNR	31.52	34.33	37.34	PSNR	33.16	35.98	39.36

Table-2:Block based with update step ,inaccurate motion vectors[23]vs.MCTF in H.264
Mobile QCIF :

Block based with update step with inaccurate motion vectors MCTF in H.264

Bitrate(kbps)	550.0	850.0	1100.0	Bitrate(kbps)	474.9	815.5	1063.7
PSNR	~32.50	~34.50	~36.00	PSNR	33.99	37.25	39.08

(These results are taken from a graph)

Table-3:Mesh based MCTF[13] vs. MCTF in H.264
Football SIF :

Mesh based MCTF MCTF in H.264

Bitrate(kbps)	500	10000	Bitrate(kbps)	486.2	980.11
PSNR	25.32	28.33	PSNR	26.52	29.74

Table-4: MC_EZBC vs. MCTF in H.264
Foreman QCIF :

MC_EZBC MCTF in H.264

Bitrate(kbps)	108.32	181.30	321.31	Bitrate(kbps)	101.3	155.2	268.7
PSNR	30.83	34.31	37.73	PSNR	33.16	35.98	39.36

We also compared our configuration with the latest MPEG SVC group core experiments' reference codec which implements H.264 like features such as adaptive block size, rate distortion optimization and also uses Barbell lifting for temporal decomposition [15]. This codec implements the update step with inexact inverse motion vectors. For a fair comparison we run Barbell codec with only temporal scalability mode with 3 layers. Since this codec utilizes context adaptive binary arithmetic coding in the entropy coding stage, we set the CABAC configuration

in the H.264 coder. Barbell codec forms temporal decomposition 'on the fly' to avoid boundary effects [8]. We set GOP=16 to avoid GOP boundary effects as much as possible. The optimum configuration would be to place I frames only at the scene boundaries but memory constraints will effect the implementation. Also staying in H.264 syntax forces us to use fixed GOP size.

Table-5:Barbell lifting based vs. MCTF in H.264

Foreman CIF :

Barbell lifting based				MCTF in H.264			
Bitrate(kbps)	496.1	577.9	665.2	Bitrate(kbps)	489,8	569.5	652.0
PSNR	36.42	36.99	37.52	PSNR	36.88	37.55	38.09

We also compared corresponding temporal sub layers. We used corresponding original frames in the PSNR calculation.

Table-6: Low temporal layer comparison

Foreman CIF :

Barbell lifting based				MCTF in H.264			
Bitrate(kbps)	394.1	440.3	501.8	Bitrate(kbps)	377.2	431.5	497.0
PSNR	36.13	36.55	37.07	PSNR	36.15	36.82	37.34

Chapter 3

CONTENT ADAPTIVE SCALABILTY TYPE SELECTION

3.1 Motivation and Related Works

Scalable video coders enable flexible adaptation of video bitrate through signal-to-noise ratio (SNR), temporal, and/or spatial scalability. Furthermore, fully embedded coders enjoy the property of encode once, decode with any of temporal-spatial-SNR resolution for computationally efficient adaptation. Examples of fully embedded scalable video coders include MC-EZBC **[9]**, [**[11]** and MSRA **[12]**. Both employ motion compensated temporal filtering (MCTF) with lifting **[16]** to provide temporal scalability, followed by a spatial wavelet transform to provide spatial scalability, as shown in Fig-2.All subbands are then encoded using an embedded entropy coder to obtain SNR scalability. A survey of recent developments in MCTF-based scalable video coding can be found in **[17]**. In this work, we employed the MSRA coder **[12]** with advanced motion compensation (MC) techniques, such as variable block sizes, ¼ pixel accuracy motion vectors, several MC modes as used in the H.264 standard **[24]**, and overlapped block MC. For entropy coding, it employs the 3D Embedded Subband Coder with Optimized Truncation (3D-ESCOT) **[25]**, which provides rate-distortion optimized multiplexing of subbands that are independently coded by bitplane coding.

Different scalability types generally result in different types of visual distortion on the decoded video depending on the bitrate and content (Section 2). Furthermore, we observe that in many cases a single scalability operator does not fit the entire video well, and the scalability operator should be optimized for different temporal segments of the video depending on the content of the segment. There is only limited amount of work that investigates the dependency between selection of scalability operators, video content, and bitrate, and address the optimum scalability option selection problem **[26]-[29]**. In one of the earlier works **[26]**, authors investigate optimal frame rate selection for MPEG-4 Fine Granular Scalability (FGS), where they conducted subjective tests to derive an empirical rule, based on the PSNR. A metric for the optimal ratio of

spatial and temporal information has been defined in **[28]** and compared with a threshold to select between the spatial and temporal operators. Optimal trade-off between SNR and temporal scalability is addressed in using some content based features, where a machine learning algorithm has been employed to match content features with the preferred scaling option. A similar approach is followed in **[29]** where content based features have been used to select one of MPEG-4 FGS modes based on an objective distortion metric defined in **[30]**. Other works on adaptation of video to available bandwidth by spatial and/or temporal resolution adjustment include those using non-scalable video coders **[31]**,**[32]** or transcoding **[33]**,**[34]**. In **[31]**, optimal rate adaptation is studied by varying spatial resolution, frame rate and quantization step size using integer programming. In **[32]**, optimum frame rate and quantization parameter selection to minimize the mean square error (MSE) is presented with rate-distortion modeling and frame skip. In **[33]**, a content based prediction system to automatically select the optimal frame rate for MC-DCT coded video transcoding based on the PSNR is proposed. In **[34]**, the MSE distortion is used for rate distortion modeling of multidimensional transcoding.

It is well known that visual distortions cannot always be measured meaningfully in terms of the mean square error or the PSNR **[35]**. An example confirming this observation is shown in Figure-5, where the SNR scalable video has a higher PSNR, but is visually inferior to the spatial scalable coded video. Frame rate preferences of low bitrate videos are studied by subjective tests in **[36]**. Because subjective tests are time consuming, many objective measures are proposed to match subjective evaluation scores **[37]**. Objective measures can be grouped as: those based on a model of low level visual processing in the retina and those which quantify compression artifacts **[38]**. An early example of the latter type is **[40]**, where visual distortion for MPEG-2 coded videos is measured considering blockiness and a perceptual model. In **[43]** subjective evaluation of videos coded with several coders, including scalable coders, is investigated and significant correlation is found with distortion based objective metrics. Although, several objective metrics for spatial distortions, such as blockiness and blurriness, have been proposed in the literature **[43]-[40]**, the temporal distortion caused by lower frame rate is not well addressed.

In this work, based on the observation that a single scalability choice does not generally fit entire video content well, we study the relationship between the scalability-type, the content-type and the bitrate. We define an objective distortion measure which is a linear combination of flatness, blockiness, blurriness and jerkiness distortion measures to choose the best scalability operator for each temporal segment at a given bitrate. The parameters of this measure can be

adapted to shot type, since the dominant distortion may depend on the content (e.g., flatness may be more objectionable in far shots with low motion, whereas jerkiness may be more objectionable in shots with high motion). This requires video analysis to be performed for shot/segment boundary detection and shot/segment type classification. There is a significant amount of work reported on automatic video analysis **[47]**, which is beyond the scope of this work. In the past few years specific content analysis methods have been developed for sports video **[48]**. Content/shot dependent video coding and streaming techniques have been proposed in **[49]**, where different shots have been assigned different coding parameters depending on the content and user preferences. In **[50]**-**[51]**, a sports video streaming framework is proposed with content adaptivity with respect to shot type relevancy.

This part offers the following novelties compared to the state of the art reviewed above:

a) We propose a procedure for *automatic* selection of the best scalability type, among temporal, spatial and SNR scalability, for each temporal segment of a video according to content, at a *given bitrate*.
b) We propose an objective cost function that is adaptive to video segment content, and present a training procedure to adapt the coefficients of the cost function to video segment content-type.

Potential applications of the proposed method include: 1) Content re-purposing: Video stored at a server using embedded coding at a high enough bitrate can be down-scaled to the target bitrate (CBR) of a user by changing the frame rate and/or spatial picture size (for encoding only) and/or the quantization parameter. 2) Video streaming over time-varying channels: If the throughput of the user is time-varying, then the target bitrate can be specified for each GoP individually, and the process becomes GoP-based rate adaptation by scaling option selection. The scaling option selected at the server side can be sent as side-information so that the receiver (client) performs appropriate spatial/temporal interpolation, when necessary, for display. The proposed adaptation of the scalability according to video segment content can be performed in near real time, thanks to "encode once decode many times with different parameters" property of fully embedded scalable coders.

This part is organized as follows: we discuss distortion measures in Section 3.2. Section 3.3 presents the choice of scaling options (SNR, temporal, spatial and their combinations) and the problem formulation. Two subjective tests and statistical analyses of the results are described in Section 3.4. The goal of Test I is determination of the coefficients of the overall cost function

for individual shot types using a training process. Test II aims evaluation of the performance of the proposed content-adaptive bitrate scaling system for an entire video clip which consists of several temporal segments to demonstrate that video scaled according to the proposed adaptive segment-based variation of the scalability type is visually preferred to videos scaled by using a single scalability type for the whole duration. Examples provided in this work have been selected from the sports domain. In order to apply the proposed procedure to other content domains, the training step (Section 3.3.C), and hence the subjective tests (Section 3.4) need to re-performed. Conclusions are presented in Section 5.

3.2 Selection of Quality Meaasures

It is well-known that different scalability options yield different types of distortions. For example, SNR scalability results in blockiness due to block motion compensation (see Fig-4) and flatness due to large quantization parameter (Fig-5a) at low bitrates. On the other hand, spatial scalability results in blurriness due to spatial low-pass filtering in 2-D wavelet coding Fig-5b), and temporal scalability results in temporal blurring due to temporal low-pass filtering and motion jerkiness. Because the PSNR is inadequate to capture all these distortions or distinguish between them, we need to employ visual quality measures. *It is not the objective of this research to develop new video quality metrics or verify them*. We only employ such metrics to develop a measure for scalability type selection. The following recently published measures (with small modifications) have been used in this work, although the proposed framework does not rely on any specific measures.

Figure 4: An example of blockiness distortion, truncated with SNR scaling at 100 kbps

a) SNR scaled, PSNR=29.19 b) Spatially scaled, PSNR= 27.79

Figure 5: Spatial and SNR scaled videos at 100kbps. Although the SNR scaled video (on the left) is visually poorer (even the ball is not visible), its PSNR is higher than the spatially scaled video (on the right).

A. Blurriness Measure

Blurriness is defined in terms of change in the edge width **[39]**. Major vertical and horizontal edges are found by using the Canny operator, and the width of these edges are computed by finding local minima around them. The blurriness metric is then given by:

$$D_{blur} = \frac{\sum_i \left(Width_d(i) - Width_{org}(i) \right)}{\sum_i Width_{org}(i)} \qquad (1)$$

where $Width_{org}(i)$ and $Width_d(i)$ denote the width of the i^{th} edge on the original (reference) and decoded (distorted) frame, respectively. Edges in the still regions of frames are taken into consideration. The threshold for change detection is selected as 15 **[40]**.

B. Flatness Measure

Although flatness degrades visual quality, it does not affect the PSNR significantly. Hence, a new objective measure for flatness based on local variance of regions other than edges is used. First, major edges using the Canny edge operator **[41]** are found, and the local variance of 4x4 blocks that contain no significant edges are computed. The flatness measure is then defined as:

$$D_{flat} = \begin{cases} \dfrac{\sum_i \left[\sigma_{org}^2(i) - \sigma_d^2(i) \right]}{N} & \text{if } \sigma_{avg}^2 \prec \text{t} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

where $\sigma_{org}^{\ 2}(i)$ and $\sigma_{d}^{\ 2}(i)$ denote the variance of 4x4 blocks on original (reference) and decoded (distorted) frames, respectively, N is the number of 4x4 blocks in a frame, and t is a threshold value which is experimentally determined. The hard-limiting operation serves two purposes: i) measures flatness in low texture areas only, where flatness is the most visible, and ii) provides spatial masking of quantization noise in high texture areas.

C. Blockiness Measure

Several blockiness measures exist to assist PSNR in the evaluation of compression artifacts under the assumption that the block boundaries are known a priori **[40]-[46]**. For example, the blockiness metric proposed in **[46]** is defined as the sum of the differences along predefined straight edges scaled by the texture near that area. When using overlapped block motion compensation and/or variable size blocks, location and size of the blocky edges are no longer fixed. To this effect, first the locations of the blockiness artifacts should be found. Straight edges detected in the decoded frame, which do not exist in the original frame, are treated as blockiness artifacts. Canny edge operator is used to find such edges. Any edge pixels that do not form straight lines are eliminated. A measure of texture near the edge location, which is included to consider spatial masking, is defined as:

$$\mathrm{TM}_{hor}(i) = \sum_{m=1}^{3}\sum_{k=1}^{L}\left|f(i-m,k) - f(i-m+1,k)\right| + \sum_{m=1}^{3}\sum_{k=1}^{L}\left|f(i+m,k) - f(i+m+1,k)\right|$$

where, f denotes the frame of interest, and L is length of the straight edge, where we set L=16. The blockiness of the i^{th} horizontal edge can be defined as:

$$\mathrm{Block}_{hor}(i) = \frac{\sum_{k=1}^{k=L} \left| f(i,k) - f(i-1,k) \right|}{1.5 \cdot \mathrm{TM}_{hor}(i) + \sum_{k=1}^{k=L} \left| f(i,k) - f(i-1,k) \right|}$$

The blockiness measure for all horizontal block borders, Block_{hor}, is defined as:

$$\mathrm{BM}_{hor} = \sum_{\text{i}\in\text{All horizontal block boundaries}} \mathrm{Block}_{hor}(i)$$

Blockiness measure for vertical straight edges BM_{vert} can be defined similarly. Finally, total blockiness metric D_{block} is defined as:

$$D_{block} = BM_{hor} + BM_{vert} \qquad (3)$$

D. Temporal Jerkiness Measure

In order to evaluate the difference between temporal jerkiness of the decoded and original video with full frame rate, we compute the sum of magnitudes of differences of motion vectors over all 16x16 blocks at each frame (without considering the replicated frames)

$$D_{jerk} = \frac{\sum_{i} |MV_d(i) - MV_{org}(i)|}{N} \quad (4)$$

where $MV_{org}(i), MV_d(i)$ and N denote the i[th] element of the motion vector of the original 16x16 block, motion vector of the 16x16 block of interest and the number of 16x16 blocks in one frame respectively.

E. Dependence on Interpolation Techniques

In cases where bitrate reduction is achieved by spatial and temporal scalability, the resulting video must be subject to spatial and/or temporal interpolation before computation of distortion and for proper display. Then, the distortion between the original and decoded video depends on the choice of the interpolation filter. For spatial interpolation, we use the inverse of the Daubechies 9-7 filter, which is reported as the best interpolating filter for signals down sampled using the wavelet filter **[52]**. We verified that, this inverse wavelet filter performed, on the average, 0.2 dB better than the 6 tap filter of the H.264 standard **Error! Reference source not found.**. Temporal interpolation should ideally be performed by MC filters **[53]**. However, when the low frame rate video suffers from compression artifacts such as flatness and blockiness, MC filtering is not very successful. On the other hand, simple temporal filtering, without MC, results in ghost artifacts. Hence, we employ a zero order hold (frame replication) for temporal interpolation, which results in temporal jerkiness distortion.

3.3 Problem Statement and Method

In this section, we first present a list of scalability options for each video segment, assuming that the input video is parsed (divided) into temporal segments and each segment is classified into one of K classes according to content type using a content analysis algorithm. Shot boundary determination and shot type classification, which are beyond the scope of this work, can be done automatically for certain content domains using existing techniques, e.g., for soccer videos **[48]**. Next, we formulate the problem of selecting the best scalability option for each temporal video segment (according to its content type) among the list of available scalability options, such that

the optimal option yields minimum total distortion, which is quantified as a linear combination of the four individual distortion measures presented in Section 2. Finally, the training procedure for determination of the coefficients of the linear combination, which quantifies the total distortion, as a function of the content type of the video segment is addressed.

A. Scalability Options

There are three basic scalability options: temporal, spatial, and SNR scalability. Temporal scalability can be achieved by MCTF in order to obtain high and low temporal frequency frames, and by skipping high frequency frames and their motion vectors. There are two sources of distortion caused by temporal scaling: i) jerkiness introduced by low frame rate; ii) temporal blur from the update step of lifting to obtain low pass frames. If motion compensation works effectively, temporal blur may not be visible. Spatial scalability is achieved by spatial wavelet decomposition of frames after the MCTF. Spatial scaling introduces blur (in the process of interpolation back to original size for display) and ringing. We also observe that spatially scaled videos have lower PSNR (after interpolating back to original size) than their visual quality suggests (see Fig-4). SNR scalability is provided by the embedded entropy coding of subbands after temporal and spatial decomposition.

We also consider combinations of scalability operators to allow for hybrid scalability modes. In this work, we allow six combinations of scaling operators that constitute a reasonable subset of scalability options for the target bitrates (100-300 kbps) tested. These combinations are:

1. SNR only scalability
2. (Spatial) + SNR scalability
3. (Temporal) + SNR scalability
4. (Spatial + temporal) + SNR scalability
5. (2 level temporal) + SNR scalability
6. (2 level temporal + spatial) + SNR scalability

where the parenthesis indicates the spatial and temporal resolution extracted for each scaling option as shown in Figure 4. For example, option four denotes that the extracted bitstream corresponds to one level temporal and one level spatial scaling that produces half the original frame rate and half the original spatial resolution; and option five produces one quarter of the original frame rate and half the original spatial resolution. Of course, for different target bitrate ranges, such as 25-50kbps, other combinations may also be considered.

B. Selection of Optimum Scalability Option for Each Temporal Segment

Most existing methods for adaptation of the video coding rate are based on adaptation of the SNR (quantization parameter) only, because: i) it is not straightforward to employ the conventional rate-distortion framework for adaptation of temporal, spatial and SNR resolutions simultaneously, which requires multidimensional optimization; ii) PSNR is not an appropriate cost function for considering tradeoffs between temporal, spatial and SNR resolutions.

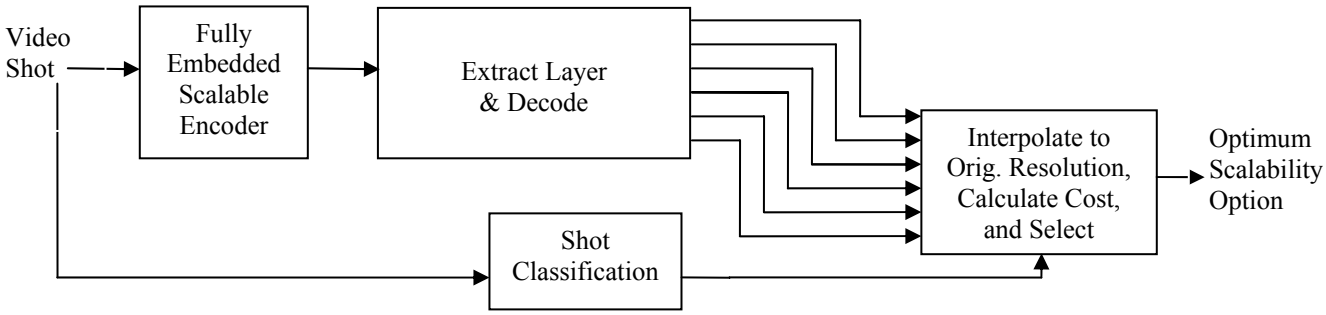

Figure 6: The proposed algorithm of optimal scaling option selection.

Considering the above limitations, we propose a quantitative method to select the best scalability option for each temporal segment by minimizing a visual distortion measure (or cost function). In **[54]**, a distortion metric which is a linear combination of distinct distortion metrics such as edgeness and temporal decorrelation has been proposed. Following a similar approach we define a new objective cost function of the form:

$$D = \alpha_{block} D_{block} + \alpha_{flat} D_{flat} + \alpha_{blur} D_{blur} + \alpha_{jerk} D_{jerk} \qquad (5)$$

where, α_{block}, α_{flat}, α_{blur}, and α_{jerk} are the weighting coefficients for blockiness, flatness, blurriness, and jerkiness measures, respectively. We also propose that the coefficients of this linear combination should be tuned according to content of the shot (shot type). For example, blurriness is more objectionable in close-medium shots; flatness is more disturbing in far shots; and motion jerkiness is more noticeable when there is global camera motion. A procedure for determination of the coefficients of the cost function according to content type is presented in Section 3C.

A block diagram of the proposed system is shown in Fig-6, where a fully embedded scalable video coder is employed. Bitstreams formed according to different combinations of scalability options are then extracted and decoded. Finally, the above objective cost function is evaluated for each combination, and the option that results in the minimum cost function is selected.

C. Determination of the Coefficients of the Total Distortion Measure

In this section, we present a training procedure along with a subjective test (Subjective Test-I) to determine coefficients α_{block}, α_{flat}, α_{blur}, and α_{jerk} of the cost function according to content type. The basic idea is to select the coefficients such that the objective measure (5) is in agreement with the results of the Subjective Test-I as closely as possible. To this effect, a subjective distortion score (6) is defined in Section 2.4.C based on the results of Subjective Test-I conducted on a training set of shots representing each content type class. The coefficients α_{block}, α_{flat}, α_{blur}, and α_{jerk} are computed for each class type separately by least squares fitting the objective cost function (5) to subjective distortion scores (6) for that class type. In particular, the coefficients are found such that the value of the objective cost function for some training shots matches subjective visual evaluation scores of those shots in the least squares sense. The coefficients computed over the training set of shots are then used to select the best scalability option on a non-overlapping test set of shots. The selection of the training shots is an important issue since all probable distortion types for that shot type should exist in the training shots.

a) Far shot with camera pan

b) Far shot without camera pan

c) Close shot with camera pan

d) Close shot without camera pan

Figure 7: Four shot types with respect to distance of shots and type of motion

3.1 Subjective Tests and Results

This section presents two subjective tests, Test-I for training and Test-II for validation of the proposed scalability option selection method. The data-set obtained from Test-I is statistically analyzed to justify our assumptions, e.g., the best scaling option depends on the bitrate and shot type. In our tests, the MSRA coder **[12]** with four-level temporal and three-level spatial decomposition and GoP size of 32 frames is employed. All training/test videos are available from our webpage **[55]**.

A. Subjective Test for Training (Subjective Test-1)

The goal of Test I is determination of the coefficients of the objective cost function (5) for individual shot types using a training process (as discussed in Section 2.3.C). This test is set up with 20 subjects according to ITU-R Recommendation BT.500-10 **[56]**, using a three level evaluation scale instead of ten levels. A *Single-stimulus Comparison Scale* is used in the test, i.e., assessors viewed six videos generated by the scaling options listed in Section 2.B in random order without seeing the originals. For each "bitrate"–"shot-type" combination, each assessor was asked to rank the six videos using the three levels: Good, Fair and Poor; with ties allowed. The video clips used are of 3-5 seconds duration at CIF resolution and contain typical shots from a soccer game. For the soccer video domain, we define 4 shot types according to camera motion and distance as: Type 1: Far shot with camera pan; Type 2: Far shot without camera pan; Type 3: Close shot with camera pan; Type 4: Close shot without camera pan. Examples of these shot types are shown in Fig **Figure** . We tested three different bitrates: 100 kbps, 200 kbps and 300 kbps. At these rates, all shot types other than Shot 3 (close shot with camera pan) are affected by flatness, blurriness and jerkiness distortions; Shot 3 has blockiness instead of flatness as the significant artifact. Each subject evaluated four shot types decoded at three different bitrates with

six different scaling options. For each subject, the evaluation is organized into 12 sessions, where in a single session a subject evaluated one shot type decoded at the same bitrate for six different scaling options. Calculation of coefficients given the results of Test-I is explained in Section 2.4.C.

B. Statistical Analysis of the Subjective Test Results

We performed statistical analysis of the subjective test results to answer the following questions:

1. *Is there a statistically significant difference in the assessors choices created by the scaling operator selection? In other words, does the selection of the scalability operator matter?*
2. *If an optimal scaling operator exists, does it change with respect to the shot-type, i.e., is the shot-type a statistically significant factor in ranking scalable coded videos?*
3. *Is the bitrate a significant factor in addition to the scaling option and the shot-type?*
4. *Are there significant clusters in the choices of assessors? Or, is the optimal operator selection subjective?*

To answer the first three questions we applied the Friedman test **[57]**, which evaluates whether a selected test variable, e.g., bitrate, shot-type, etc., can be used to form test result clusters that contain significantly different results as compared to a random clustering. The output of this test, ρ, is the significance level, which represents the probability that a random clustering would yield the same or better groups. A result with ρ less than 0.05 or 0.01 is assumed to be significant in general.

The results of the Friedman's test are as follows:

-Clustering with respect to the scaling option is significant with ρ being almost zero. With this result, scaling operator selection is indeed significant.

-After scaling option clustering, clustering with respect to shot-type is found to be significant with $\rho=0.004$

-In addition to scaling operator and shot-type, bitrate is a significant factor in clustering with significance $\rho=0.001$.

User dependency of the results seemed to be another factor to analyze. We first calculated the correlation of the user's scores, shown in **Figure** , to see if there is any clustering. We observe two types of user groups: one group prefers higher picture quality over higher frame rate (type A) and the other group prefers higher frame rate (type B). Based on this observation, we clustered subjects into two groups using 2-means clustering. We also determined the significance of the clustering by rank-sum test for each video. The separation of users into two groups is found to be

significant at 5% level for 30 videos out of 72 videos coded with different scaling option, bitrate, and shot type combinations. Most of these 30 videos that users' preferences differ are coded at low bitrates, which leads us to conclude that the difference in the users frame rate preferences increases as the overall video quality decreases. This observation is also confirmed by Subjective Test-II.

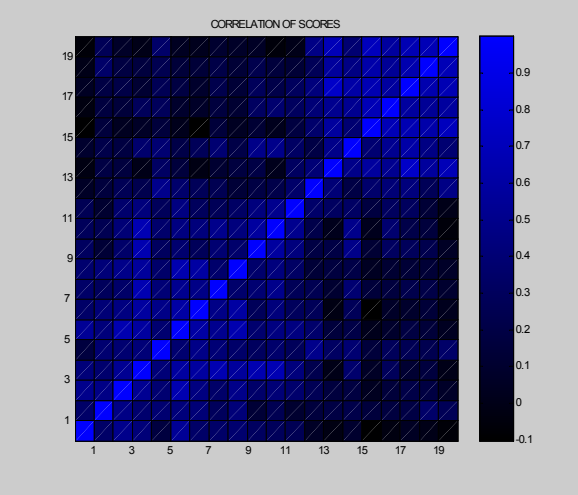

Figure 8: The autocorrelation of subjective scores shows a noticeable difference between two groups of subject

C. Subjective Distortion Score and Coefficient Calculation

In order to quantify the results of the subjective test for least squares fitting, we define the subjective distortion score (SDS) of a video shot (segment) as:

$$SDS = \frac{2}{1 + (2 \times S_1 + S_2)/(2 \times S_{\max})} \qquad (6)$$

where S_1 and S_2 are the numbers of "good" and "fair" grades, respectively, and S_{max} is the number of assessors.

We determine the coefficients of the objective cost function (5) for each shot type by least-squares fitting to corresponding SDS. The coefficient sets for type 2 and type 4 shots (without pan) are calculated only on shots coded at 100 kbps, while coefficients for type 1 and type 3 shots (with pan) are computed on shots coded at 200 kbps since all probable kinds of distortion (blurriness, flatness, blockiness and jerkiness) should exist in the training shots. The coefficient sets computed for all users together, and type A users and type B users separately, are

shown in Table-1. The objective measure (5) with the least squares fitted coefficients and subjective distortion score (6) for shot type 4 are illustrated in Figure 9.

In order to demonstrate that coefficients computed at a given bitrate also perform well at other bitrates for a particular shot type, we computed the Spearman rank correlation between the SDS (6) and the ranking provided by our method as shown in Table 2. It can be seen that our algorithm finds the best or the second best scaling option from the six scaling options for most cases. Furthermore, the results of the Subjective Test-II confirm that coefficients found for a given shot type in a specific video will work for the same shot type in any other video.

We also employed the well-known VQM objective measure, defined in **[30]** and **[58]**, instead of our objective measure (5) in the proposed selection scalability option selection algorithm at several bitrates (see Table 2). Results show that our metric performs better than the VQM, since VQM does not have parameters that can be tuned for different content types.

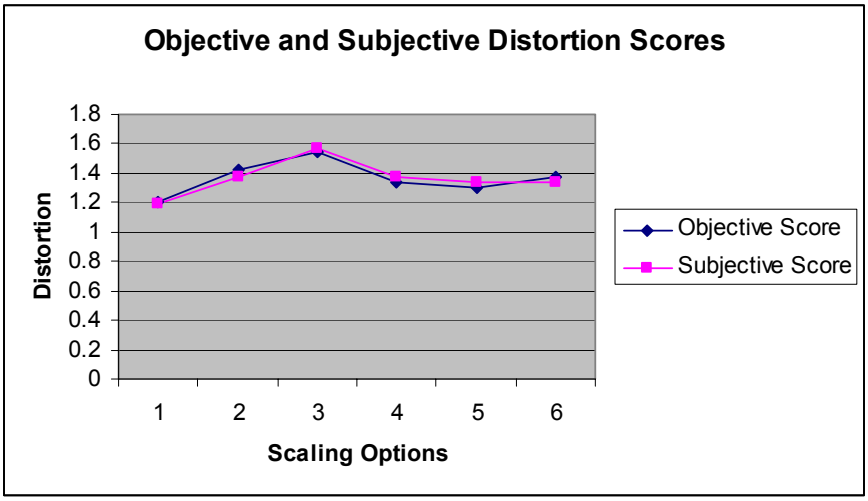

Figure 9: Objective Measures (after least-squares fitting) and Subjective Distortion Scores for different scaling options on the training set of shot type-4 decoded at 100kbps.

D. Performance Evaluation by Subjective Tests (Subjective Test-II)

In this test, a new test video clip is divided into temporal segments according to the shot-types defined above. For each temporal segment, the best scaling option is determined using our proposed method with coefficients determined as described above. The segments extracted with the best scaling option are then cascaded to form the extracted test video bitstream. In this test, two comparisons are performed to answer two questions:

i) Does changing the scalability option with respect to content-type really make significant difference in the visual quality of the scaled video when compared to using the same

scalability option for the whole sequence? To answer this question, adaptively scaled video is compared to videos decoded at the same rate but obtained with all fixed scaling options.

ii) Is it useful to consider subject type (i.e., type-A or type-B as defined in Section 4.B) in determining the best scalability option? Changing the scalability option according to subject type requires knowledge of the subject type beforehand which makes the system rather difficult to implement, so learning the extent of the improvement when subject type is used, will be beneficial for practical application scenarios. To answer this question, subjects are asked to choose from videos which are content-adaptively scaled with coefficient sets tuned to their specific subject types vs. tuned to general type.

The results confirm that content adaptive scaling provides significant improvement over fixed scaling as shown in Table-3. Majority of the subjects prefer dynamically scaled video to any constant scaling option for all bitrates tested. The effect of subjective preferences on the scalability operator selection is observed to be somewhat important at low bitrates and not important at higher rates; a result which was observed in the first subjective test also. This result agrees with the observation that 'information assimilation' (that is, where the lines are, who the players are, which teams are playing) of a video is not affected by the frame rate but 'satisfaction' is **[59]**. At high bitrates, spatial quality is high enough for information assimilation and best scalability operator is selected mainly from satisfaction point of view which leads to similar choices of scaling option for all users. At low bitrates, picture quality may not be good enough for information assimilation. Hence, information assimilation plays a key role on optimal operator selection for type-A subjects; where for type-B subjects satisfaction is still more important in determination of optimal scaling choice, resulting in significant clustering among subjects in the subjective evaluation of videos coded at low bitrates.

Table-7: The scaled coefficients of the cost function for all users / type A users / type B users, respectively.

	Blurriness	Flatness	Blockiness	Jerkiness
Shot-1	0.9290/0.9378/ 0.8828	0.0385/0.0411/0.0451		0.0325/ 0.0212/ 0.072
Shot-2	0.9399/ 0.9589/ 0.9090	0.0079/0.0089/0.0060		0.0521/0.0322 /0.0849
Shot-3	0.9598/ 0.9754/ 0.7631		0.0158/0.0109/0.0794	0.0244 / 0.0137 / 0.157
Shot4	0.9809/ 0.9927/0.9446	0.0066/ 0.0054/0.0097		0.0125/ 0.0019/ 0.0457

Table-8: The performance of our optimal operator selection algorithm: the Spearman rank correlation, the subjective rank of the option that our algorithm finds and the subjective rank of the option that another objective metric finds (applicable for only all users part) respectively.

	All users			Type-A users			Type-B users		
	100kbit	200kbit	300kbit	100kbit	200kbit	300kbit	100kbit	200kbit	300kbit
Shot1	0.74/1/1	0.94/1/4	0.77/1/3	0.6/1	0.83/1	0.54/2	0.84/1	0.9/1	1/1
Shot2	0.31/3/5	0.71/1/1	0.99/1/1	0.17/3	0.37/1	1/1	0.99/1	0.99/1	1/1
Shot3	0.43/4/3	0.77/1/1	0.49/1/1	0.5/4	0.93/1	0.6/1	0.77/3	0.79/1	0.37/1
Shot4	0.86/1/4	0.94/1/4	1/1/1	0.93/1	0.84/2	0.69/2	0.81/2	0.9/1	1/1

Table-9: The first row shows percentage of users who preferred the proposed content adaptive scaling option to every fixed scaling option. The second row shows the percentage of subjects who preferred the adaptive scaling option with respect to subject type rather constant scaling option with respect to subject type.

	100kbit	200kbit	300kbit
Adaptive scaling performance	%95	%75	%75
Bimodal user separation	%20	%5	%5

Chapter 4

SCALABLE MULTIPLE DESCRIPTION CODING FOR ADAPTIVE PEER-TO-PEER STREAMING

4.1 Motivation and Related Works

In the following, we provide a short introduction on multiple description coding, peer-to-peer streaming, and the relation between them. We also discuss the related literature and our contributions.

A. Multiple Description Coding

Multiple description coding (MDC) addresses the problem of encoding source information using more than one independently decodable and complementary bitstreams, which, when combined, can provide the highest level of quality and when used independently, can still provide an acceptable level of quality. This is made possible by introducing some redundancy in each description, which will be discarded if all streams are received. The amount of redundancy introduced can be optimized according to assumed loss rate. It is well known that MDC can provide robust video communication over unreliable networks, such as the Internet, when combined with path/server diversity at the cost of reduced compression efficiency [60].

There has been significant amount of work on multiple description video coding [61]- [65]. Notably, Wang et al. [61] used motion estimation across descriptions, called motion compensated multiple description coding (MC-MDC). Ortega et al. introduced unbalanced multiple descriptions, where the descriptions do not have identical rates, i.e., one description is coded at a lower bitrate than others.[62] It is shown that unbalanced MD is very useful for Internet streaming where paths with different bandwidths are common. Barlaud et al. [63]proposed a MDC framework based on Discrete Wavelet Transform (DWT) which allows redundancy adaptation to varying wireless channel conditions. However, their approach is valid only for N=2 descriptions

and they do not consider adaptation after encoding which is very important for content distribution since peers have limited storage capacity. In order to make post-encoding adaptation possible, MDC should be employed with a scalable video coding scheme.

One example of scalable MDC is based on motion compensated temporal filtering, where high frequency frames are grouped into two descriptions and missing frames are estimated using motion vectors in the two descriptions [64]It is reported to outperform existing non-scalable MD video coders in compression performance while providing flexible rate allocation and redundancy control, although its performance is degraded under significant motion since estimating missing frames then becomes a difficult task. Puri et al. [65]introduced use of forward error correction (FEC) with MDC. FEC-MDC unequally protects a progressive bitstream with erasure channel codes such as Reed-Solomon codes according to the importance of bitstream segments. Every description includes a protected version of the most important layer, then half of the next important layer and so on. However, this requires a significantly high number of descriptions (such as N=16 or N=32) to allocate redundancy effectively; such a high number will deteriorate the compression efficiency especially at low packet loss rates. Also determining the optimal FEC allocation according to the varying channel conditions on the fly is a difficult task. Altunbasak et al. [66]proposed a network adaptive unbalanced MDC method grouping 3-D SPIHT coefficients into two unequal groups and applying unequal error protection to bitplanes. They change the amount of FEC according to packet loss rates and allocate the number of wavelet coefficients to code for each sender according to the estimated TFRC rate. Hence, rate and redundancy is adapted during encoding. We note that, none of the available MDC methods addresses post encoding adaptation of the number of descriptions/layers and the amount of redundancy in descriptions/layers according to the network conditions.

B. Receiver Driven P2P Streaming

In traditional video-on-demand systems, the main server finds the edge server nearest to the user when the user requests a video, and the video is streamed from that edge server. However, this approach of streaming requires many popular videos coded and stored in edge servers, many edge servers to provide scalability and a coordinator server to find the nearest edge servers which may be very costly to the end user. Instead, Peer-to-Peer (P2P) networks can be used to stream on demand media.

P2P video streaming has recently gained interest [68]- [71] since it can provide low-cost streaming of media data in a scalable manner due to possible large deployment of P2P networks. In P2P streaming, there is no need for dedicated edge servers to store and distribute videos, instead peers who store the requested coded video, stream to the requester peer when the request occurs. Most of the earlier works focused on tree structures for multicast streaming which is efficient in dealing with flash crowds. In on demand P2P streaming, the requesting peer can also coordinate the peer selection and streaming from multiple peers to avoid a central coordinator or specific tree structures unlike multicasting. The sending peers can be encouraged to store and stream videos with some kind of privileges like the ones in Kazaa etc, or some other fairness criteria as proposed in [72]However P2P streaming has some challenges that need to be addressed:

Peer Query and Selection: Selecting the optimal peers for streaming is a difficult task because of the heterogeneity of sending peer conditions. Round Trip Times (RTT) and upstream transmission capacities of peers may vary from peer to peer. Any peer may tune out unpredictably, or new peers with good conditions may become available during the transmission. Hence, peers should be monitored and peer query/selection should be performed not only in the beginning of session but also during the transmission. However, peer queries may create significant additional network traffic when performed frequently. Also, some of the peers may actually share a link in their path to the receiver, a situation not easily detectable. To use disjoint paths is important since any loss on the shared link effects both of the streams transmitted over two paths. Therefore, to find disjoint paths, the receiver may need to monitor the correlation of packet losses between all path pairs.

Packet Losses: During video transmission, any sender peer may turn off, a link may be broken or packets may be lost due to competing TCP traffic. Because of the stringent delay constraints coupled with possibly high RTT values, lost packets may not be retransmitted. MDC or layered coding can be a remedy to this problem as the decoder can generate video with graceful degradation from the received packets under packet loss.

High RTT values: Analysis of P2P systems shows that a significant portion of peers suffer from high latency [73]. This high latency condition makes Auto-Repeat-Quest (ARQ) type of error resilience techniques infeasible. It also makes the signaling between receiver and sender a difficult task.

Low and Heterogeneous Upstream Rates: Usually upstream rate of an individual sending peer is much lower than the downstream capacity of the receiving peer. Hence, some kind of distributed streaming is necessary to achieve high quality streaming. Also, peers may be connected to the Internet via different speed connections. So, heterogeneous upstream capacities necessitate a rate allocation algorithm to minimize the overall distortion. For MDC streaming, a flexible unbalanced MDC is needed which should allow any rate partition post encoding. Rate allocation should be performed at the receiver since only receiver knows the statistics of each path and sending peers may not be willing to waste resources on rate-distortion optimization performed for rate allocation. Moreover, there should be a reliable mechanism to send rate and packet partitioning information (control packets) from receiver to senders.

Time Varying Network Conditions: In P2P, it is common that packet loss rates and upstream capacities may change due to external traffic or any peer may tune out unpredictably. According to the analysis of P2P systems, around %60 of the peers keep active less than 10 minutes each time they join the system [73]Hence, coding algorithm should allow post-encoding adaptation according network conditions, especially the number of descriptions/layers should be flexible.

Competing TCP Traffic: It is highly likely that there is competing TCP traffic along with the streaming between the sender and the receiver or on the subsections of the path between them created by other machines. Clearly, streaming traffic should not suppress the competing TCP traffic while allocating the necessary bandwidth for streaming. To this effect, TCP Friendly Rate Control (TFRC) should be used to calculate sender bit-rates to achieve a fair distribution of bandwidth between TCP data and video.

Zakhor et al. [74]proposed a general framework for receiver driven, simultaneous distributed streaming where the receiver coordinates the packet transmissions from each sender. They define a rate allocation algorithm at the receiver for fair distribution of the total receiver bandwidth

among heterogeneous senders with different rate and channel characteristics. Also a packet partition algorithm running at the senders is proposed to ensure that each packet is sent only once. Chou et al[71] proposed a tree management algorithm, CoopNet, to provide path diversity in the P2P multicast of FEC-MDC coded video. P2P multicast is appropriate for live video streaming since identical content is requested by many peers where in on-demand streaming there is only one requester. Ortega et al.[62]proposed an adaptive layered streaming framework, Pals, for P2P on-demand streaming. They proposed a receiver driven coordination framework with congestion control of layered streaming from multiple sender peers. Wang et al. [69] proposed using FEC-MDC in P2P streaming where each peer stores only certain descriptions coded at some fixed rate. If one serving peer fails, system searches for another peer containing the same descriptions. They also analyzed using layered coding in place of multiple descriptions and concluded that layered coding outperforms MDC when replacement time of down peers with new ones is relatively small [70].

C. Contributions

We propose an adaptive receiver driven P2P streaming framework based on a novel scalable multiple description coding scheme. The main contribution of this work is two-fold: First, we propose a novel scalable MD video compression scheme which provides efficient adaptation to the network conditions while providing high compression efficiency. Second, a P2P video streaming system is designed using the proposed flexible MDC scheme, varying:

1) the number of descriptions –layers
2) the redundancy level of each individual description and
3) the rate of each description/layer,
4) Rate allocation among descriptions, i.e., balanced/unbalanced MDC - layered coding on the fly (post encoding).

Providing a variable number of descriptions/layers with varying rates becomes an important concern in P2P video streaming, where the number of available “good” source peers and their channel conditions are not known a priori and change in time. Consider the scenario, where there are n source peers with scalable coded video available to start with, and each send one out of n descriptions toward a common destination. Performance of each source path can be measured at the destination for a given period. As a result of this measurement, the best $k < n$ sources can be selected out of the initial n and $r < k$ of them are now requested to provide one of r base

descriptions and remaining k-r sources can be used for enhancement descriptions. All of these can be produced (post encoding) from the scalable descriptions available at each source. We also note that number of serving peers and channel conditions (bitrate and packet loss rate) may change during transmission. For each Group of Pictures (GOP), the number and the rate of descriptions/layers and the level of redundancy in each description/layer can be optimally found with model based rate-distortion optimization at the receiver peer using packet loss and bit-rate statistics. To avoid excessive peer queries resulting in significant congestion, new queries are performed only when a path that carries base multiple descriptions fails, otherwise system distributes the total load among already found serving peers. Sending rate and layer/description allocation information can be communicated to the sender peers through control packets. The overview of the proposed system is shown in Figure 10.

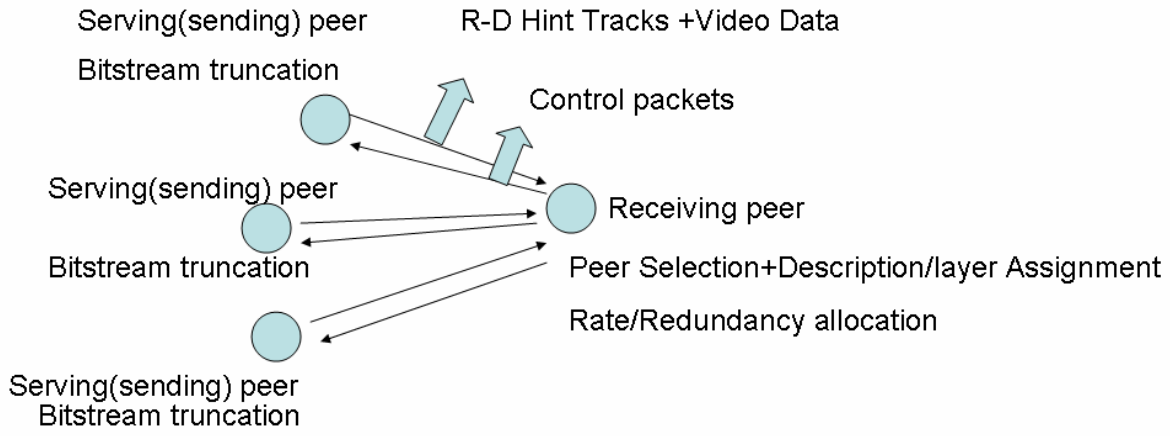

Figure 10: Overview of the proposed P2P streaming system

4.2 Scalable Multiple Description Coding

In this section, we propose a novel flexible multiple description video coding framework that is based on fully scalable (embedded) wavelet video coding. In embedded wavelet video coding, first motion compensated temporal filtering (MCTF) is performed along the temporal direction to efficiently decorrelate frames within a GOP. Then, all filtered frames (i.e., temporal subbands) are coded using JPEG-2000 coder as shown in Fig-11. It is well known that, this scheme provides compression efficiency comparable to H.264/AVC, which is the state of the art non-scalable video coder, while providing embedded SNR, temporal and spatial scalability.

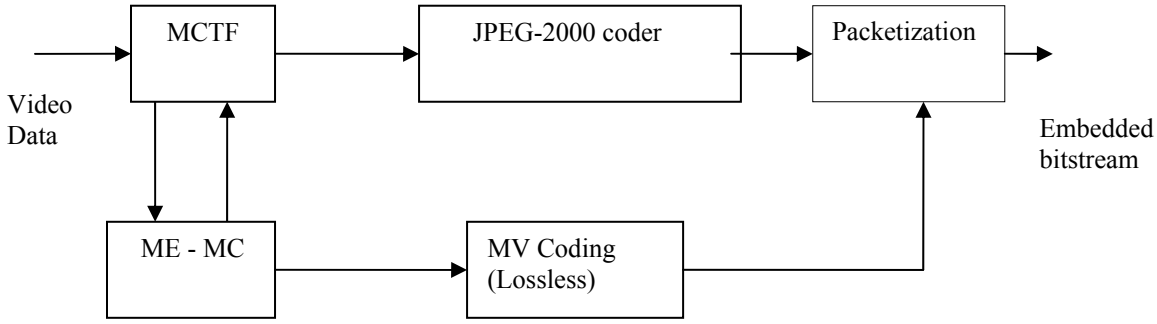

Figure 11: General structure of the used t+2d wavelet video coder

In JPEG-2000, every spatial subband is divided into non-overlapping code-blocks and each code-block is encoded independent of other code-blocks. Since each code-block is coded with bitplane coding, they can be truncated at any rate post encoding. For each layer, contribution of each block to the total distortion is found by Embedded Block Coding with Optimized Truncation (EBCOT) and bits from every code-code block are truncated according to their contribution to the overall distortion. The number of bits contributed from each code-block to overall layer rate and rate-distortion slope is embedded in the packet[1] header for fast post compression rate-distortion optimized truncation. Hence, the JPEG-2000 bitstream already contains rate-distortion information of every code-block for each layer in its packet headers. If we ignore the temporal drift problem, which is already mitigated by the open loop MCTF structure, total distortion can be written as weighted sum of the code-block distortions as

$$D_{total} = \sum_{over\_all\_codeblocks} v_i \times w_i \times D_{code\_block}^{\ i} \qquad (6)$$

where v_i and w_i denote temporal and spatial weights defined as L_2 norms of the spatial and temporal wavelet synthesis filter coefficients [16], under the assumption of orthogonality of the spatial and temporal wavelet filters and motion compensation, since the distortion in spatial domain will then be identical to distortion in the wavelet domain.

A. Generation of Multiple Base and Enhancement Descriptions

[1] In JPEG-2000 terminology, a packet is a collection of coded code-blocks from the same resolution and the same layer

In the proposed framework for generating multiple descriptions, each description is composed of code-blocks extracted at different rates from a single embedded scalable video bitstream. Since every code block can be truncated at any rate, independent of other code-blocks, we truncate every codeblock once at a high bitrate R_{BH}, and once at a low bitrate R_{BL} to generate two so-called base streams. Since both base streams are formed by the most significant bitplanes (MSB) of each codeblock, they can be independently decoded. The remaining bitplanes of each code block can be used for generating enhancement streams. Clearly, the enhancement streams require availability of the base streams for decodability. In the following, we describe how to generate N base descriptions and M enhancement descriptions from these streams.

Base descriptions are formed by various combinations of low and high rate codeblock base streams. For example, N base description can be generated by including one code-block truncated at the high rate out of every N, and remaining N-1 code blocks at the low rate for each description. The lowest frequency code-blocks in both temporal and spatial domains are coded at the high rate in all descriptions, since they affect the visual quality more than the other code-blocks. The case N=2 is depicted in Figure 12, where the ordering of code-blocks follows a zig-zag scan order. Note that description 1 has high-low-high-… rate ordered codeblocks along the zigzag scan order, whereas description 2 has low-high-low-… rate ordered codeblocks after the lowest frequency codeblock which is coded at the high rate in both descriptions. The codeblock truncation rates R_{BH} and R_{BL} can be determined by rate distortion optimization for each block as in EBCOT [75]. All overhead including motion vectors is lossless coded for every description. The proposed framework allows generation of both balanced and unbalanced base descriptions. Unbalanced descriptions can be generated by truncating unequal amount of code-blocks at high and low rates in different descriptions. For example, Figure 13 shows the case of N=2, where two out of every three code-blocks is truncated at the high rate in description 1, and one out of three is truncated at the high rate in description 2. In the decoder side, if all descriptions are received, we use only code-blocks which are coded at the high bitrate. On the other hand, if only one description is received we still have an acceptable video quality with some code-blocks decoded at low-bitrates. Since the code-blocks at different rates are extracted from a single fully embedded video bitstream, description rates as well as the number of descriptions generated by various combinations of them for each code-block are totally flexible and can be varied post encoding.

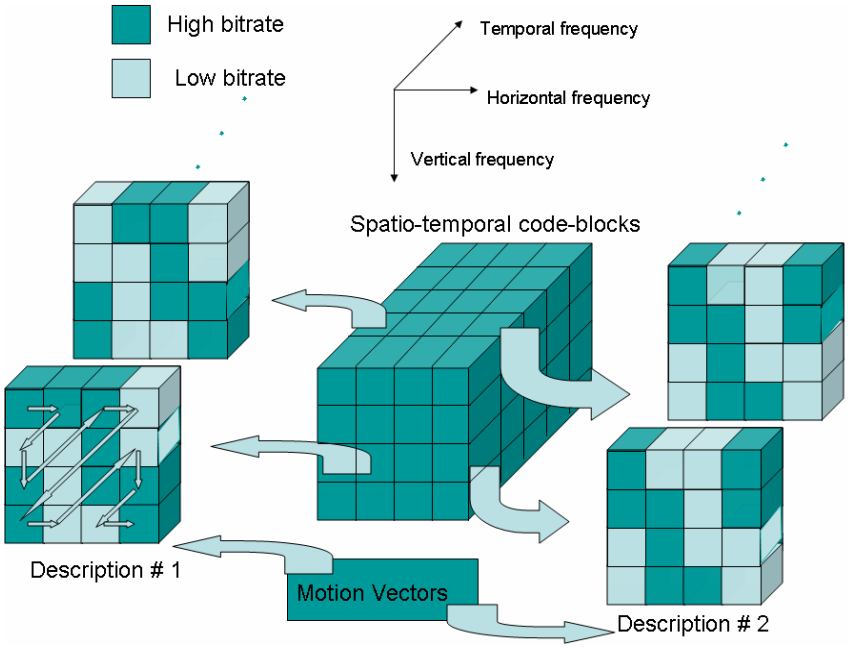

Figure 12 : Proposed MDC method for N=2 descriptions derived from two streams decoded at the high and low rates. Code-blocks follow the zig-zag scan order.

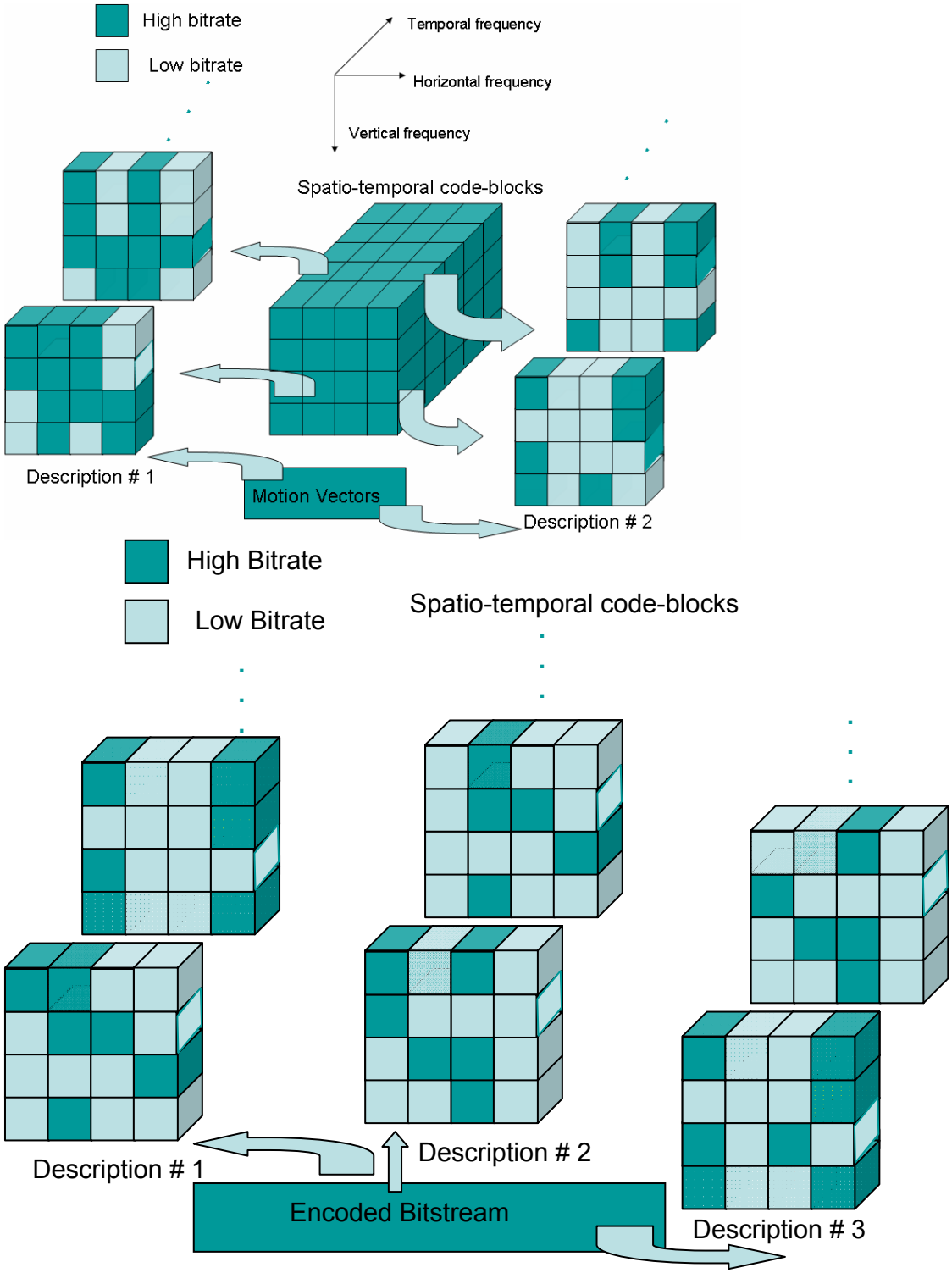

Figure 13: Proposed MDC method for N=2 unbalanced descriptions coded at two rates. Ratio of the rates of Description-1 and Description-2 depends on the high/low rates, individual code-block rates and the amount of overhead.

In order to make every enhancement layer decodable without depending on any other enhancement layer, we also generate multiple enhancement descriptions from the remaining bitplanes (not used in the base descriptions) by specifying a starting rate, R_S, which can be either R_{BH} or R_{BL}, and a low R_{EL} and a high rate R_{EH} using an approach similar to the one used for generating base descriptions.

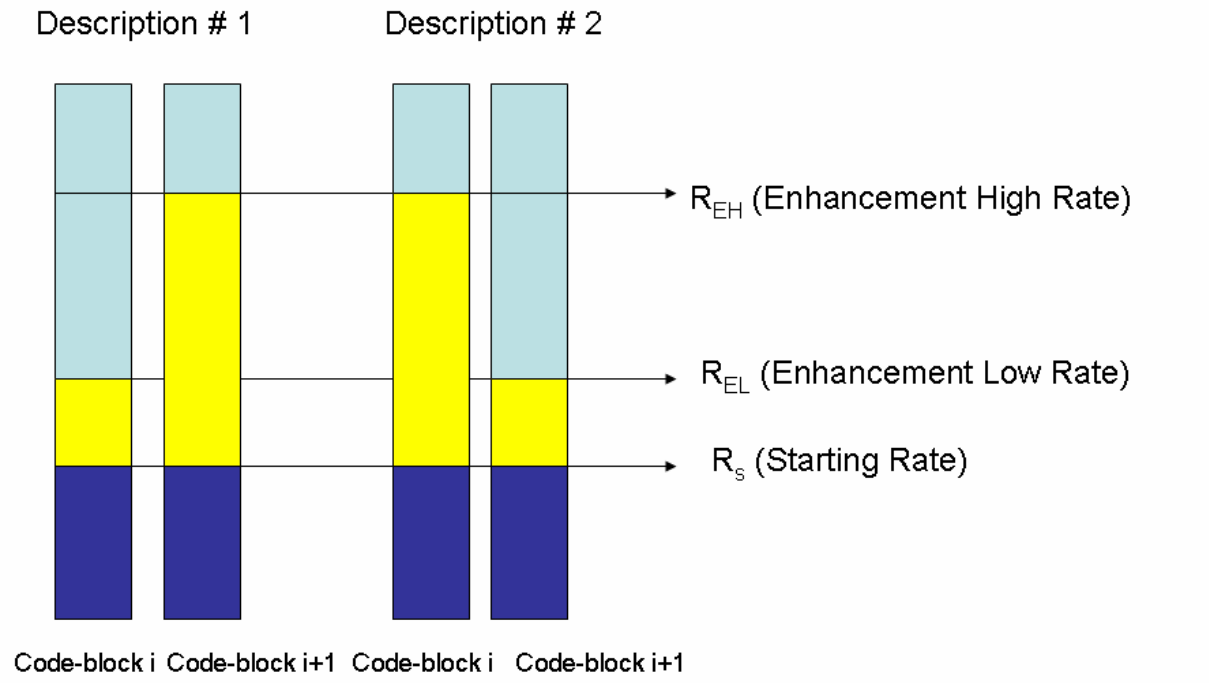

Figure 14 :Enhancement Descriptions for N=2 descriptions. Yellow parts show the descriptions truncated from starting rate (R_s) upto high/low rates(R_{EH},R_{EL}).

In summary, the process of generation of multiple base and enhancement descriptions is completely specified in terms of the following design parameters:

i) Number of base and enhancement descriptions, N and M: Every based description has some code-blocks extracted at the high rate and others at the low rate, where the number of the code-blocks extracted at the high rate decreases as the number of descriptions increase. The number of enhancement layers can also be adjusted post encoding thanks to embedded coding of the code-blocks.

ii) High and low rates to generate base descriptions (R_{BH}, R_{BL}): As the high rate increases; the low rate should decrease in order to maintain a fixed average rate for the description.

iii) Assignment of high and low rate code blocks to descriptions, $C_i = [c_{i1}, c_{i2},\ldots,c_{iL}]$, i=1,…,N: The vector C_i specifies which code-blocks will be truncated at the high and low rates in description i, where L is the number of code-blocks in one GOP. The jth element $c_{ij}=1$ of C_i indicates that jth codeblock will be coded at the high rate in description i, and $c_{ij}=0$ indicates the low rate. Since C_i needs to be sent from the receiver to all senders for every GOP, it should be expressed in the minimum form possible. To this effect, we assume that the codeblock pattern repeats periodically with a period of K codeblocks, and a new vector $C_i^{'}$ of length K (K<<L) is defined such that $c_{ij} = c_{ij}^{'}(j(\mathrm{mod}\, K))$. At the receiver side, the vector C_i can be easily obtained from $C_i^{'}$, which is more compact to send. We note that, for base descriptions, the lowest spatio-temporal frequency is represented at the high rate without any consideration to the value of the element of the vector C_i for that code-block.

iv) Specification of enhancement descriptions (R_S, R_{EH}, R_{EL}): We note that, enhancement layers can also be sent with redundancy. The starting rate (R_S) shall be either R_{BH} or R_{BL} depending on whether the base is coded at the high or low rate, respectively. Since enhancement layers are also sent as multiple descriptions, we also need to specify high/low rates (R_{EH}, R_{EL}) for enhancement descriptions. Figure 14 illustrates R_S, R_{EH}, R_{EL} for the case of two code-blocks and two enhancement descriptions.

The proposed scalable multiple description video coding framework enables computation and post encoding adaptation of the rate and redundancy of multiple descriptions in terms of the above design parameters:

- rate of each individual description/layer: Since embedded bitplane coding is used for encoding every code-block, each description is inherently rate scalable.
- redundancy in each description: The redundancy r for each description is defined as the ratio of the amount of redundant bits that are not used and the amount of bits that are used when all descriptions are received, i.e.,

$$r = \frac{R\_not\_used}{R\_used} \tag{7}$$

where R_not_used stands for the number of redundant bits when all descriptions are available and R_used is the number of bits used when all descriptions are available

In the proposed scheme, redundancy is determined by the number of descriptions, high/low rates of code-blocks and overhead including motion vectors. Since motion overhead is lossless coded, high/low rates should be tuned to vary redundancy in one description. As low and high rates becomes close to each other, redundancy inserted in descriptions increases and vice versa. The extreme case of equality of high and low rates corresponds to identical descriptions with %100 redundancy level. Hence, it is possible to control the redundancy level by changing high and low rates of code-blocks.

We note that the design parameters can be adjusted on the fly, post encoding. Hence, the rate and redundancy of each description can be adapted according to transmission/network conditions on the fly without re-encoding the video. The compression efficiency of the proposed MDC method is compared to other MDC methods such as MD-MCTF [64] and MD-MDC. The comparative results are presented in Section.4.4-A.

B. Determination of Design Parameters-An Example

Here we provide an example in order to demonstrate the determination of design parameters for a special case of two peers available with identical bandwidth (R) and packet loss rates (p). Since there are only two paths, we set N=2 and M=0; hence we do not send any enhancement descriptions. Identical packet loss rates necessitate use of balanced descriptions, which can be achieved by setting assignment vectors as C_1=[0,1] and C_2=[1,0].

R_{BH} and R_{BL} can be found by a Lagrangian rate-distortion optimization procedure using packet loss rates and distortion expression. Distortion estimate (6) for a code-block i to minimize can be written as

$$D_{est}^{i} = (1-p)^2 D_1^{i} + (1-p)p D_2^{i} + p(1-p) D_3^{i} + p^2 D_4^{i} \quad (7)$$

where p is the packet loss probability and D_1, D_2, D_3, and D_4 are distortions respectively when i) both versions (high rate-low rate) of the code-block i arrive, ii) only low rate version arrives, iii) only high rate version arrives, iv) none of the code-blocks arrive. The decoder uses the high rate coded code-block, if it exists, otherwise, it uses the low rate code-block. If none of the code-blocks are available, no concealment is performed. Hence, distortion estimate for code-block i becomes

$$D_{est}^{\ i} = (1-p)\times D_1^i + p\times(1-p)\times D_2^i + p^2\times D_4^i \quad (8)$$

Since bit allocation for both high and low bitrate codeblocks are performed using EBCOT, we can safely assume that the total rate for one codeblock is constant, i.e.,

$$R_L^i + R_H^i = R_{total}^i \quad (9)$$

where R_{total}^i is the rate of the code-block when all bits are spent on high rate description. Bits spent on motion vectors, other overhead and the distortion where no code-block is available can not be minimized, therefore, the problem reduces to investigation of low and high extracting rates for i[th] code-block (i.e., R_L^i, R_H^i) which minimizes

$$D^i = (1-p)D_1^i + p(1-p)D_2^i \quad (10)$$

Hence, the Lagrangian cost to minimize for every code-block is

$$J = D^i + \lambda \times R_{total}^i \quad (11)$$

From the minimization of the expression in (11), we get

$$\lambda_{high} = p \times \lambda_{low} \quad (12)$$

where λ_{high} and λ_{low} respectively correspond to rate-distortion slopes of low and high bitrate coded code-blocks.

Hence, optimum rates for high and low rate code-blocks can be found by jointly iterating high/ low rates (R_L^i, R_H^i) and slopes (λ_{low} , λ_{high}) to satisfy both Eq.9 and Eq.12 using the embedded rate and slope information in the code-block.

NS-2 simulation results of the proposed derivation of high/low rates for this special case can be found in Sec-4.4-B. In the following section we explain how to optimize the parameters to adapt general network conditions with arbitrary number of paths with different conditions.

4.3 Adaptive Peer-to-Peer Streaming System

A. Overview of the Proposed Streaming System

We propose a receiver-driven many-to-one (unicast) P2P streaming system which dynamically adapts the number of base and enhancement descriptions, redundancy level of each description, and rate allocations between descriptions sent to the receiver by more than one sending peers.

A streaming session is initiated by the receiver with a *peer query process*, where the receiver finds all available peers that can serve the requested video. Peer query techniques that can be used are reported in [70]. The receiver then sends the total number of available sending peers and which description to send to each available peer found. Upon the receipt of this information, all sending peers start sending their assigned descriptions and the receiver starts the playback after the usual pre-roll delay. The receiver continuously monitors the quality of all paths from the sending peers through a *path measurement process*. The path measurement process, explained in detail in the next subsection, is used to estimate the TCP friendly bandwidths of all paths from the sources to the receiver and packet loss correlation between these paths. The receiver performs rate-distortion optimization using packet loss rates and estimated TCP friendly bandwidths obtained during the streaming of the previous GOP, to determine for the current GOP:

i) Number of base descriptions to be used: Most of the time, two descriptions give the best compression efficiency - loss resilience trade-off. However, for high packet loss rates, three or more descriptions may be preferable.

ii) Rate allocation among descriptions: Receiver determines which peers will send the base and which ones will carry the enhancement descriptions. Also, rate allocation between descriptions is performed by utilizing the flexible number of balanced/unbalanced descriptions.

iii) High/low rates: the redundancy level determined by both the number of descriptions and high/low rates of code-blocks.

Peer_query;

Send the total number of participating peers and base description assignment to each peer;

Receive base descriptions from participating peers with initial high/low rates ($R_{high\text{-}init}$, $R_{how\text{-}init}$);

Measure_path_performance;

Determine design parameters through RD optimization;

Send signaling information (control packets) to each peer;

while number_of_paths_eligible ≥ 2 {
Measure_path_performance
if any change in network conditions
new rate_allocation }

while number_of_paths_eligible < 2 {
new rate allocation
Measure_path_performance
Peer query;}

$R_{high\text{-}init}$, $R_{how\text{-}init}$ denote the initial low/high rates of the multiple descriptions to be transmitted at the beginning of the streaming session, which depend on the video content and resolution.

B. Path Measurement and Peer Selection

During a streaming session, the receiver measures the following parameters for each path *i*:

1. *average packet loss rate (p_i)*
2. *average receiving bandwidth (R_i)*
3. *packet loss correlation between path i and path j (η_{ij})*

The average packet loss rates p_i are computed using the techniques described in the Real-time Transport Protocol (RTP)[76]. Given the packet loss rate p_i, the average TFRC bandwidth R_i for ith sending peer is estimated using the TFRC algorithm [79] as:

$$R_i = \frac{S}{RTT_i\sqrt{\frac{2p_i}{3}} + t_{RTO_i}\left(3\sqrt{\frac{3p_i}{8}}\right)p_i\left(1+32p_i^{\,2}\right)} \tag{13}$$

where t_{RTO_i} denotes retransmit time-out value, RTT_i denotes the round trip time, and S denotes the packet size. It can be assumed that $t_{RTO_i} = 4 \times RTT_i$.

We estimate packet loss correlations between multiple paths in order to determine if paths have shared links. We need to avoid receiving multiple base descriptions over paths with shared links, because there is no benefit to using MDC if a shared link is broken. To this effect, let m_{ij} denote the number of lost packets in paths i and j within the same time interval, M denote the total number of lost packets in paths i and j, and τ be a threshold value. If the ratio

$$\eta_{i,j} = {m_{i,j}}/{M} \geq \tau \quad (14)$$

then paths i and j are decided to share a common link; otherwise, paths i and j are assumed to be disjoint [77]

The receiver determines paths eligible to carry multiple base descriptions according to three criteria:

1. Base layer MD paths should have available bandwidth R_i above some threshold R_{base}
2. Base layer MD paths should be disjoint
3. Average ON-time of a sending peer should be above some threshold T_{base}

The receiver will not allocate two base descriptions on correlated paths, but it may place a base layer multiple description on one path and enhancement layer on the other to mitigate the effect of a broken shared link.

C. Estimation of Total Distortion at the Receiver

The total distortion at the receiving peer is due to i) truncation of the fully embedded bitstream for generation of base and enhancement descriptions (source coding), and ii) packet losses during streaming. The modeling and estimation of these two kinds of distortion is discussed in the following.

i)Estimation of Distortion Due to Bitstream Truncation using Hint Tracks

Optimal number of descriptions and redundancy levels are obtained through a rate-distortion optimization process performed at the receiver. The distortion function should either be based on an analytical model or some rate-distortion hint tracks should be sent to the receiver along with data.Analytical models lack the accuracy unless they match the video content. Since the video is

not available to the receiver at the time of streaming, receiver can not match the model to the video without receiving some hint tracks from the sender.

In JPEG-2000, length of each code-block to each layer is recorded in the packet header. This auxiliary information in packet headers can be packed into one group (i.e, packed packet header in JPEG-2000), encoded with tag tree coding [75]and can be sent to the receiver. The receiver can use this code-block length and layer rate-distortion slope information to deduce rate-distortion curve instead of matched analytical rate-distortion models. We note that using packet headers in rate-distortion computation requires no extra bandwidth since receiver already needs to know the code-block lengths for each layer to form packets contents and in decoding code-block headers. Receiver uses this pre-fetched rate-distortion information in rate-redundancy allocation before requests from senders and also in decoding the encoded code-blocks. We note that distortion can be estimated using code-block lengths and layer rate-distortion slopes by the expression

$$D_{code-block} = \sum_{i:over\_all\_layers\_upto\_this\_layer} \lambda_i \times \ell_i \quad (15)$$ and Eq.6. where λ_i and ℓ_i denote rate-distortion slope and code-block length change for layer i.

We note that the receiver needs to know only the distortion change between layers instead of the absolute distortion for rate-distortion optimized rate allocation and description-layer assignment. Nevertheless, absolute distortion estimate quite successfully matches to the real absolute distortion as shown in Fig-15.

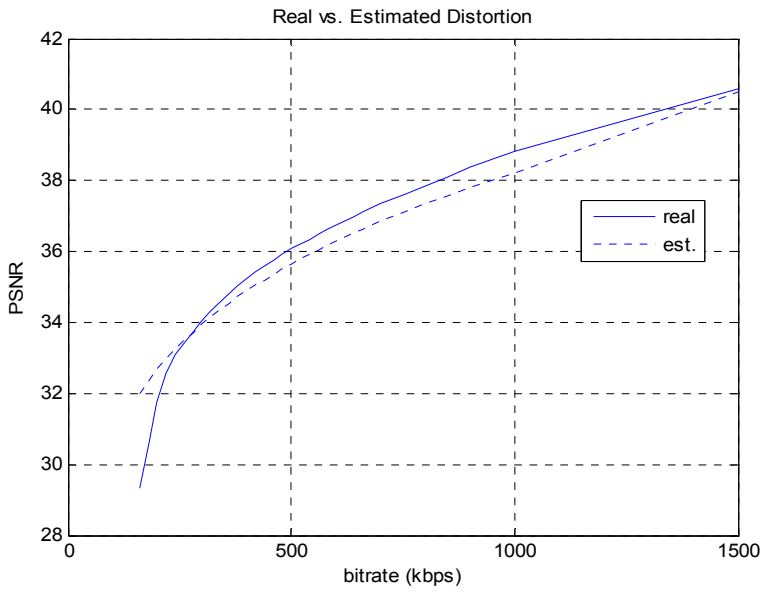

Figure -15 Real vs. Estimated Distortion from rate-distortion slopes and code-block lengths (R-D hint tracks used in streaming) Sequence: Foreman-CIF, 40 layers

Estimation of Total Distortion in the Presence of Packet Losses

Packet losses are typical in P2P networks due to unpredictable peer tune-outs and congestion caused by external traffic. In most P2P networks, average ON time of each individual peer is nearly 15 minutes [73], which is well below the typical length of a streaming session. Hence, estimated ON-time of each individual peer should be incorporated into the distortion expression. We model the ON-time with an exponential distribution $f(t) = \lambda \times e^{-\lambda t}$ with mean $\mu = 1/\lambda$. where the parameter λ depends on the peer behavior specific to the network. Since exponential distribution is memoryless, it can well capture the unpredictable tune outs of sending peers as the probability of a peer being ON for an amount of time T depends only on T, and not the starting time, i.e.,: $P\{x > t+T \mid x > t\} = P\{x > T\}$. (16)

Each sending peer may have a different average ON-time which can be stored at each peer and can be communicated at the beginning of a streaming session. Peers with very low average ON-times will not be preferred as senders.

In the Internet, packet losses due congestion are usually in bursts. Although the burst length is effective on the observed distortion [76],in this work we assume that packet losses are independent from each other for simplicity. To account for the peer tune outs in the distortion expression, we modify the observed packet loss rate as

$p_{est} = 1 - P \times (1 - p_{obs})$ (17) where, p_{est}, p_{obs} and P denote, packet loss estimate used in distortion expression, measured packet loss and the probability that the peer is on during that GOP respectively.

Hence, the overall distortion for each code-block can be written as:

$$D_{code-block} = D_o - \sum_{k=0}^{K} \left[\prod_{l=0}^{k-1} \left((1 - \prod_{n=0}^{N_l - 1} p_{est\_ln}) \right) (1 - \prod_{m=0}^{N_k - 1} p_{est\_km}) \times D_k \right] \quad (18)$$

where N_l is the number of descriptions for layer 'l', P is the probability that a peer will be ON during that GOP, D_o denotes the distortion when that code-block is not available, D_k is the reduction in distortion when that layer is available, and p_{est_ln} stands for packet loss estimate of the path that carries layer 'l', version 'n'

The total distortion is given by

$$D_{total} = \sum_{over\_all\_codeblocks} w_i \times D_{code-blcok}^{\ i} \quad (19)$$

where w_i denote the synthesis filter coefficients of the temporal transformation.

Note that D_k information is calculated during encoding and attached to the packet headers in JPEG-2000. Hence, if packet headers are obtained, distortion expression, required in rate-distortion optimized rate allocation process, can be written for the whole GOP as well as for each code-block.

D. Rate Distortion Optimization for Description Rate Allocation and Description Assignment

The receiver shall determine

a) The number of descriptions to be used

b) rate allocation among distributions (high/low rates)
c) description assignments to peers (peer selection)

based on the rate-distortion information for each code-block in a GOP.

Using the distortion estimate expressed in (9), for each code-block, rate-distortion optimization can be performed to find:

i) high and low rates of that code-block,
ii) the peers who will send that code-block
iii) the number of identical versions of that code-block .

However, this approach will be too costly in terms of not only complexity but also signaling, since for every GOP a list of code-blocks with high/low rate information should be sent to each sending peer. Such frequent and large control packets will induce congestion in the network and high delays which is common in P2P, may delay control packets.

A more practical approach would be sending only high/low rate information and description assignments for a whole GOP instead of each code-block. In that case, GOP based distortion minimization will be used to find high/low rates and peer selection instead of code-block based rate-distortion optimization. Distortion estimated using the expression (9) and best allocation policy is determined as the one with minimum distortion estimate.

Assuming that there are M peers that can send MD out of a total of N peers,

$$S = \sum_{k=2}^{M} \binom{M}{k} = 2^M - M - 1 \quad (20)$$

different description allocation polices exist. All enhancement layers are sent as multiple descriptions to make every layer received usable without depending on other enhancement layers. We note that, sending enhancement layers as multiple descriptions have no compression penalty when all descriptions are received if the high/low rates are properly chosen.

As an example, if there are five peers with the requested video content and three of them (Paths 1-2-3) are eligible to c arry the base MD, there may be may be four different description/ allocation strategies as:

1. Paths 1-2 carry base MD, Paths 3-4-5 carry enhancement layer descriptions
2. Paths 1-3 carry base MD, Paths 2-4-5 carry enhancement layer descriptions
3. Paths 2-3 carry base MD, Paths 1-4-5 carry enhancement layer descriptions
4. Paths 1-2-3 carry base MD, Paths 4-5 carry enhancement layer descriptions

The last option which is all sending peers may transmit MD which may be the preferred when all paths are highly error prone. In the case of asymmetric estimated TFRC rates, unbalanced MD can be used; hence, in that case, number of rate allocation strategies may increase. Every rate allocation policy is optimized for different low/high rate combination before comparison to other policies, hence low/high rates are also found in this step. The proposed system calculates the estimated distortion for every rate allocation policy with optimum high/low rate combination and chooses the policy which achieves minimum distortion. We note that since there is no decoding or re-encoding process as the ones like H.264/AVC standard [8], this full search minimum distortion search algorithm is not that complex. The decoder should perform distortion estimation using only S (number of rate allocation policies) $\times$ K^2 (number of high/low rate combinations where K is the number of total layer information in RD hint tracks) X L (number of vectors)for base descriptions. For enhancement layers, it can not be evaluated in closed form since they depend on base high/low rates, but the complexity order is expected to be close to that of base descriptions.

The pseudo-code of the method can be written as:

Find M paths eligible to carry base MD out of N sending peer paths
Estimate distortion for all possible combinations of k (k<M) base multiple descriptions and N-k layers
Optimize each policy according for different high/low rate combinations
Find the policy with minimum distortion estimate
Form the control packet with the found high/low rate and policy

E. Signalling using Control Packets

Optimum rate and description/layer assignment information is signaled from receiver to all senders. The format of the control packet is defined for both description sending peers and layer sending peers identically as such:

Number of Descriptions/ Description Assignment Low/High Rates (Start) Low/High Rates (End)

For peers sending base multiple descriptions starting low-high rates are zero. We note that for different layers,(i.e., base layer, enhancement layer-1, enhancement layer-2) different number of descriptions can be used to achieve unequal error protection. Example content of control packets may be: 2/1, 0/0, 140kbps/80kbps; 2/2 , 0/0, 140/80 ; …etc.

4.4 Results

A. Comparative Results on Compression Performance

The proposed method is compared to other multiple description coders with redundancy-distortion curves for some fixed bitrates. We used a wavelet coder based on JPEG-2000 [75], which uses EBCOT for rate allocation, however other wavelet based scalable coders can also be used. 3 level spatial and 4 level temporal decompositions are used in the coder.

Comparative results of the proposed coder and MD-MCTF are provided below for case N=2 descriptions at three different bitrates and four redundancy levels when only one description is received. *We note that while obtaining our results, we did not use any layering other than fixed comparison rates. Layering information is not either reported in [64]where MD-MCTF results are taken.*

Table-10: Foreman_QCIF_30fps; PSNR of the proposed method/MD-MCTF [64]

Redundancy	100kbps	200kbps	300kbps
20%	NA/27.4	31.64/28.8	34.75/29.5
30%	29.44/27.7	34.03/29.4	37.02/30.8
40%	31.09/28.0	34.99/30.2	38.01/31.9

50%	32.35/28.1	35.76./30.2	38.59/32.0

Table-11: Akiyo_QCIF_30fps; PSNR of the proposed method/MD-MCTF [64]

Redundancy	50kbps	100kbps	200kbps
20%	29.21/31.9	36.36/36.0	46.21/39.0
30%	31.03/32.0	37.76/36.2.	46.82/40.1
40%	32.15/32.0	38.67/36.3	46.95/41.0
50%	35.11/32.0	39.33/36.3	47.02/41.0

Our coder outperforms MD-MCTF significantly at medium motion sequence (Foreman), however, for sequences with low motion (Akiyo) MD-MCTF performs comparable (at some low rates better) to our coder because MD-MCTF can properly estimate missing frames at sequences with low motion. Since MD-MCTF is reported to outperform MC-MDC [61]and we observed that our coder performs better than MD-MCTF in most rate-redundancy levels, we did not compare our coder to MC-MDC coder.

B. Streaming Results with Comparisons

For streaming part, we simulate packet losses with NS-2 simulator[80]. In Part-1 and Part-2 we simulate the proposed system for special settings to demonstrate that changing the number and redundancy of each description on the fly improves the streaming performance.Part-3 includes a comprehensive simulation of a real P2P streaming system with general settings and TFRC rate allocation.

Part-1:

In this part, we show the use of changing redundancy on the fly according to the derivation of optimum high and low rates in Section 2-C for the example basic scenario. This comparison shows the importance of the selection of high and low rates.

The luminescence component of the Foreman sequence in QCIF format is coded with wavelet coder with 3 spatial and 3 temporal decomposition levels for 296 frames at 30fps. Other than the lowest frequency frame in the temporal decomposition, every frame is packetized into one packet with maximum size of 1000 byte. Ever spatial resolution in the lowest frequency frame is put into

one packet. All motion vectors for a GOP length 8 are put into a total of 2 packets. Traffic trace files generated in the coder are used in the ns-2 simulation to specify the timing and size of each packet.

There are 2 senders who have the encoded video and description generator to generate description with any redundancy level. The last hop link is bottleneck link with 100kbps bandwidth and high error rate. Both senders send multiple descriptions over disjoint paths links. Every path from senders to receiver shares one link with 200kbps bandwidth with external traffic. External cross traffic is randomly specified as %50 of the link capacity with exponentially distributed packet sizes and sending intervals.

The simulation time is the two full play time of the video, *T=20sec***.** The time period between the time at link changes the packet loss rate and sender sides become aware of that event, the recognition time, is assumed to be t_{rec}*=1sec*.. The rate of each description is set to the bottleneck bandwidth *R=100kbps.*

The proposed system starts with medium redundancy. After a recognition time t_{rec}*=1sec*, it adapts the redundancy level with respect to the packet loss rate %5. At time t=10.sec, after a t_{rec}*=1sec* time from packet loss change , it changes the redundancy level according to the loss rate %20.

For comparison purposes, the performance of a test system with fixed level of redundancy is also simulated. Since the packet loss rate alternates between %5 and %20 during the simulation, the level of redundancy is fixed according to %15 loss rate.

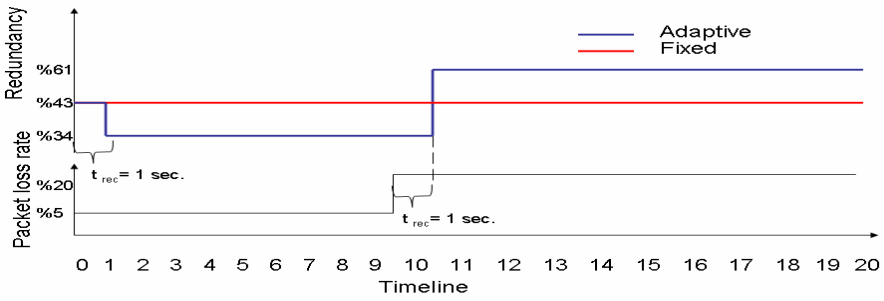

Figure 16 : Packet loss and redundancy level in time

Table-12: Low and high bitrates found by our algorithm.

	Packet Loss Rate=%5		Packet Loss Rate=%20	
	Low BR	High BR	Low BR	High BR
Adaptive	50 kbps	148 kbps	75 kbps	124 kbps
Fixed	60kbps	139 kbps	60kbps	139kbps

We compare the proposed system with the test one which has fixed redundancy level through the simulation. Packet loss rates are found by analyzing the ns-2 output trace files of the paths. The results are found by averaging 15 realizations of the simulation..

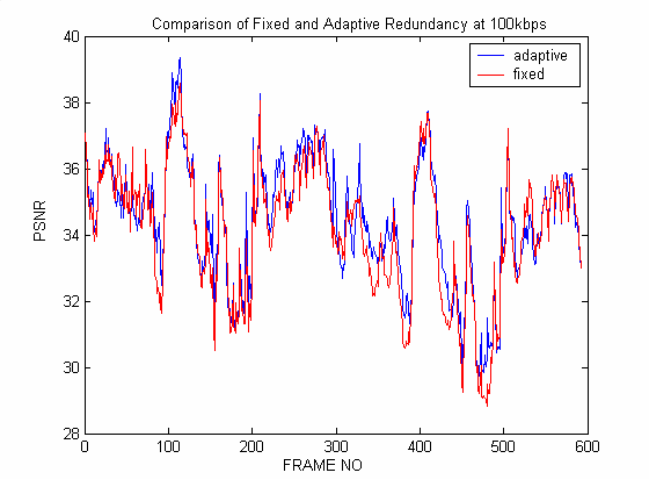

Figure 17: Comparison of adaptive and fixed redundancy

Proposed system with adaptive redundancy level outperforms fixed redundancy by 0.29dB PSNR in the first half and 0.31 dB in the second half of the simulation

Part-2:

To show the use of adapting the number of descriptions to the number of available channels; we describe this example scenario: There are 8 senders who have the encoded video and description generator to generate any number of descriptions. The last hop link is bottleneck link with 100kbps bandwidth and high error rate. All senders send multiple descriptions over disjoint links. Every path from senders to receiver shares one link with 200kbps bandwidth with external traffic.

External cross traffic is randomly specified as %50 of the link capacity with exponentially distributed packet sizes and sending intervals.

The shared link can sometimes be heavily loaded with external 1Mbps constant bitrate traffic, so the path enters loaded state. Two of the eight available paths are in loaded state initially. At 9th sec., randomly four of them also become loaded with external traffic which runs till the end of simulation. The simulation time is the two full play time of the video, *T=20sec***.** The time period between the time at path entering loaded state and sender sides become aware of that event, the recognition time, is assumed to be t_{rec}*=1sec*. The rate of each description is set to the bottleneck bandwidth *R=100kbps.* Encoding and packetization details are identical to the simulation in Part1.

The proposed system starts with N=8 descriptions. After a recognition time t_{rec}*=1sec*, it adapts the number of generated descriptions to N=6. At time t=10.sec, after a t_{rec}*=1sec* time from congestion beginning time, it changes the number of descriptions to N=2.

For comparison purposes, the performance of a test system with fixed number of descriptions is also simulated. Since the number of available good channels alternates between two and six during the simulation, the number of descriptions is fixed to N= 4. The test system begins with sending four descriptions over four channels, and uses remaining four paths to send four identical back-up descriptions. After a time of t_{rec}*=1sec.*, it stops sending two of the back-up descriptions which are on loaded path. After time *t=9.sec*, when congestion begins in four of the channels, it stills continues to send four descriptions over four paths although two of them are heavily loaded.

We compared the proposed system with the test one with two different packet loss rates as %5 and %20. Packet loss rates are found by analyzing the ns-2 output trace files of the paths which do not experience congestion. The high and low bitrates which determine the amount of redundancy is found according to the packet loss rate estimated in the first t_{rec}*=1sec* time period. The results are found by averaging 15 realizations of the simulation.

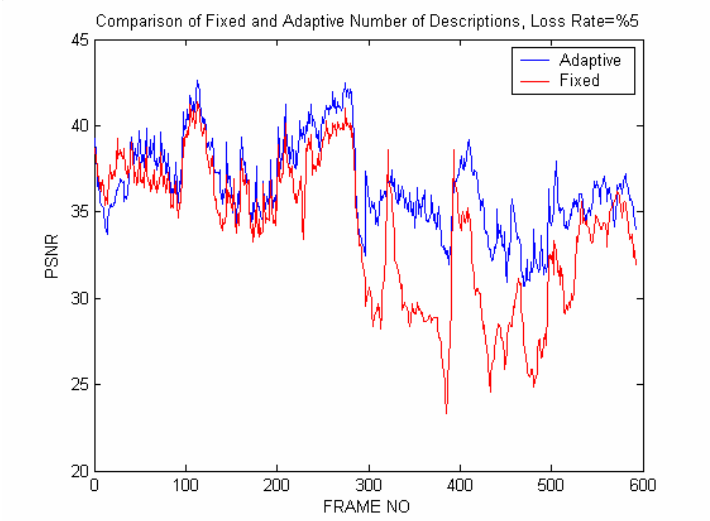

Figure 18 : Comparison of fixed and adaptive number of descriptions at %5 packet loss rate.

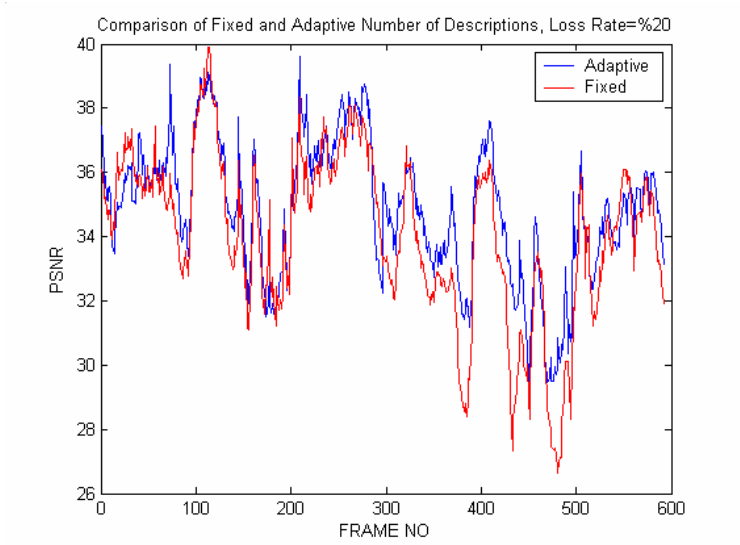

Figure 19: Comparison of fixed and adaptive number of descriptions at %20 packet loss rate.

The proposed system outperforms the fixed number of descriptions by 2.55 dB at %5 loss rate, and 0.77dB at %20 loss rate. The PSNR gap between the adaptive and fixed number of descriptions decreases as loss rate increases. The reason behind this observation is that resending the identical descriptions performs quite close to sending multiple descriptions at very high loss rates.

Part-3

In this part we simulate the adaptive streaming system for a general network setting. The luminance component of the Foreman sequence in CIF format is coded with wavelet coder with 4 spatial and 4 temporal decomposition levels for 256 frames at 30fps. We use fixed packet sizes as 500bytes. We note that although we use fixed packet sizes, we can fully utilize the packet content unlike the packetization of non-scalable bitstreams since we can extract the bitstream at fine granularity. For rate-distortion analysis we formed 20 layers between 150kbps and 1Gbps. Hence there are 20 choices for the high/low rates.

We use TFRC rate control runing at the receiver. For the specific delay parameters used in the simulation, TFRC rates are in slow start in the first 10 GOP times(~5 sec.), hence we use first 10 GOP for the path identification process, ie: whether this path shares a common link with other paths or not. Since any rate can not even exceed motion bitrate in the first 5 GOP, probe packets are sent.Then, from 5 GOP to 10 GOP time, we set the description number irrespective of the rates or packet loss information. However, we use this information for high/low rate assignment and rate allocation among descriptions (unbalanced descriptions).

The simulation set-up is shown in the Figure-20 . All paths share one link with TCP connections, Paths 1..4 have 6 TCP persistent connection where Path 5 has 13 persistent TCP connections. Paths 1…4 have additional transient 9 TCP connections which start at random times after first 10 GOP time and stop at 30 GOP time, and then starts and stops periodically for 20 GOP times. We note that the bottleneck for all paths is the shared link that carries both TCP and TFRC flow. In this simulation we assumed all sending peers have large ON time i.e, we did not use ON-OFF modeling in our distortion computation.

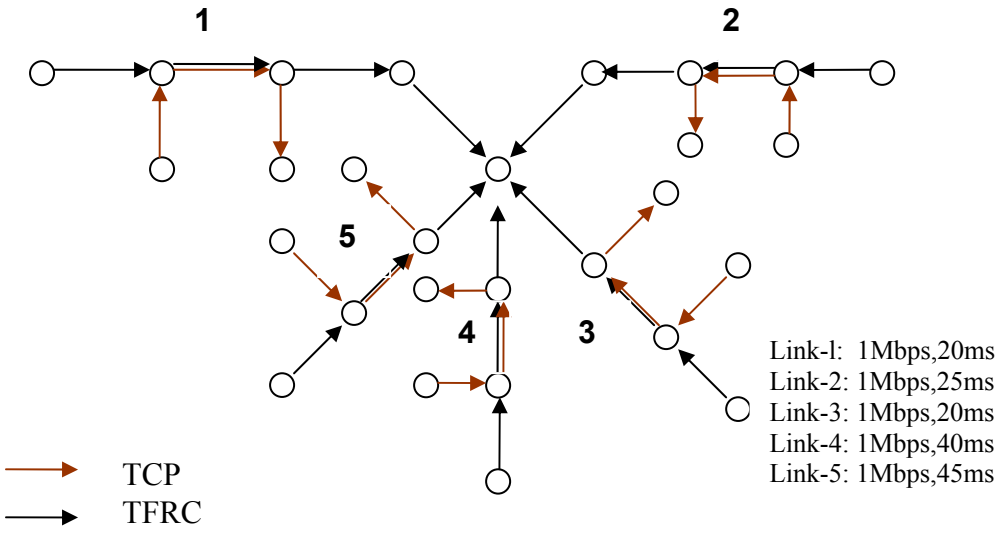

Figure 20: Simulation Set-up

As a comparison we simulated the performance of fixed multiple description streaming system where the number of descriptions is set to the available paths (i.e, with no enhancement descriptions) and the high/low rates are set to achieve minimum redundancy, code-block assignment vectors are set to C_1= [1,0,0,0,0], C_2=[0,1,0,0,0],….We note that from 5 GOP to 10 GOP time, compared and proposed systems send the identical packets, hence have the same performance.

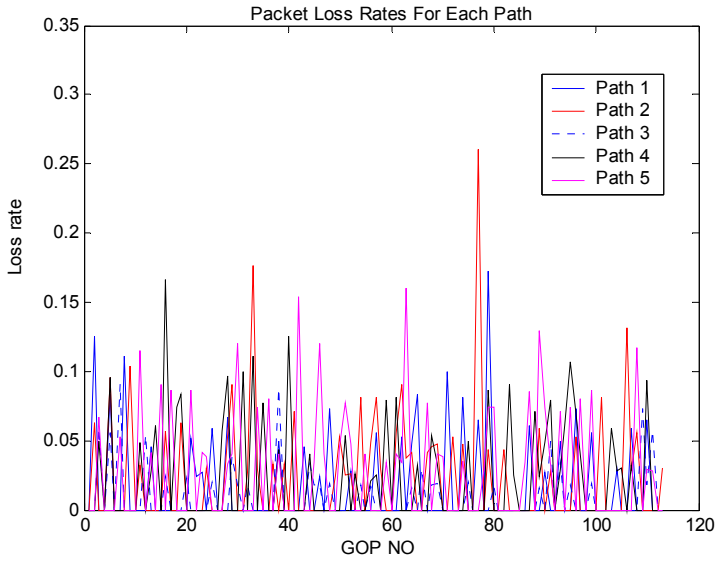

Figure 21: Loss Rates as per GOP

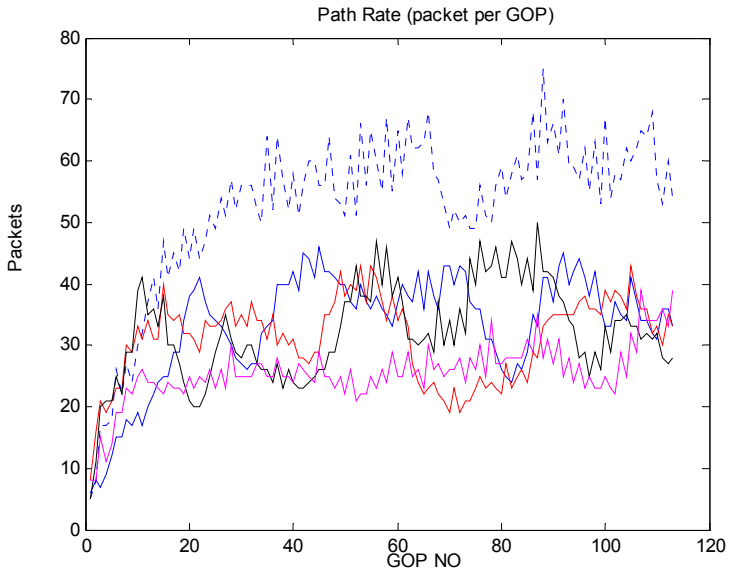

Figure 22: Path Rates as Packets per GOP

Both the proposed and compared systems use TFRC rate control and hence, have the same rate and packet loss patterns. The proposed system however sends additional packets for rate-

distortion hint tracks with along with motion vectors, i.e, the paths that carry base descriptions also carry the rate –distortion hint tracks.

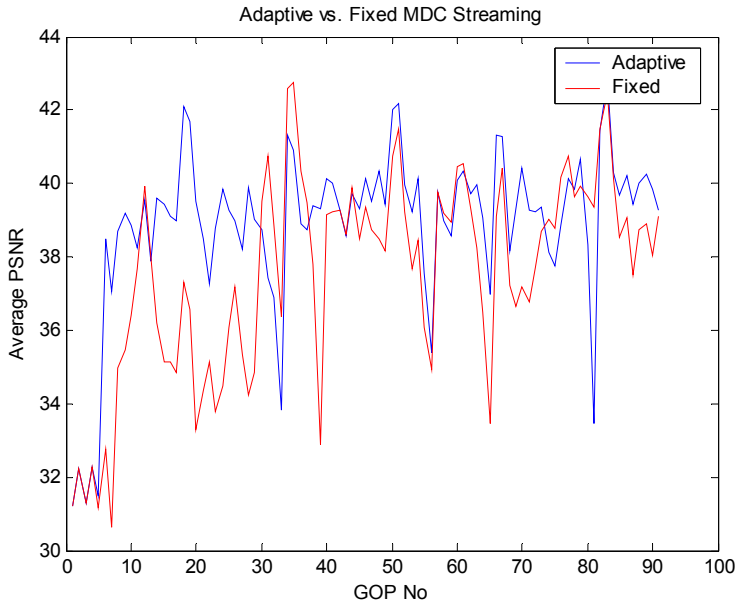

Figure 23: Simulation Results

As Fig- shows the proposed adaptive MDC streaming method outperforms fixed MDC 1.3 dB in the average even when no peer tune outs or no significant throughput change occurs in any of the paths. For the tune out scenario we already showed in the basic setting in Part-2 that adaptive MDC streaming outperforms fixed upto 2-3 dB PSNR.

Chapter 6

CONCLUSION & FUTURE WORK

In this thesis we have worked on three different problems in adaptive scalable coding framework. First, we propose an implementation of the MCTF structure in the H.264 framework to provide adaptive multilayer temporal scalability within the H.264 standard. Our results show that by utilizing H.264 standard's advanced features for motion compensation, we can achieve better compression performance. Our results may be used as a benchmark for better motion compensated prediction in temporal lifting schemes. The proposed scheme is also integrated into H.264/AVC reference software as 'Hierarchical B pictures' or 'Temporal Pyramid' and it is currently under investigation of MPEG Core Experiments for the upcoming Scalable Video Coding standard (SVC).

In the second work, we propose a content adaptive scalable video streaming framework, where each temporal segment is coded with the optimum scaling option. Optimum scaling option is determined by a cost function which is a linear combination of different distortion measure such as blurriness, blockiness, flatness and jerkiness. Two subjective tests are performed to find the coefficients of the cost function and to test the performance of the proposed system. Statistical significances of the test variables are analyzed. Results clearly show that best scaling option changes with the content, and content adaptive coding with optimum scaling option results in better visual quality. Although our results and analysis are provided for soccer videos, the proposed method can be applied to other types of video content as well.

In the last part, a flexible multiple description video coding framework, based on fully scalable wavelet video coding, with high compression efficiency is proposed. Also, a novel receiver driven unicast (many to one) P2P streaming system using the proposed MDC scheme is presented. The main contribution is optimally varying redundancy and rate of each description as well as the

number of total descriptions/ layers according to network conditions. The superiority of the proposed adaptive system to fixed MDC systems is shown with NS-2 streaming simulation.

As a future work, we will perform simulations to validate our policy of sending base multiple descriptions only on disjoint paths and justify our ON-OFF model with peer tune outs. Also, we will compare the performance of our system to layered streaming solutions.

BIBLIOGRAPHY

[1] B.-F. Hung; C.-L. Huang, Content-based FGS coding mode determination for video streaming over wireless networks, *IEEE Jour. on Selected Areas in Comm.*, vol. 21, no. 10 , pp. 1595-1603, Dec. 2003.

[2] Secker, A. and Taubman, D "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Transactions on Image Processing,* volume 12 (number 12), pages 1530-1542.

[3] B. Girod, M. Kalman, Y.J. Liang, and R. Zhang, "Advances in channel-adaptive video streaming," *Wireless Communications and Mobile Computing*, vol. 2, no 6, pp. 573–84, Sept. 2002.

[4] Y.C. Su, C.S. Yang, and C.W. Lee, " Optimal FEC Assignment for Scalable Video Transmission over Burst Error Channel with Loss Rate Feedback," *Signal Processing: Image Communication*, vol. 18, pp. 537-547, 2003

[5] R. Zhang, S. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, June 2000.

[6] A. Reibman, "Optimizing multiple description video coders in a packet loss environment," *Packet Video Workshop*, April 2002.

[7] MPEG documents, "Registered Responses to the Call for Proposals on Scalable Video Coding," *ISO/IEC JTC1/SC29/WG11 MPEG04/M10569.*

[8] T. Wiegand, G. Sullivan and A. Luthra, "Draft ITU-T Recommendation and Final DraftInternational Standard of Joint Video Specification (*ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)*," May 27, 2003.

[9] J. Ohm, "Three dimensional subband coding with motion compensation," *IEEE Trans. Image Processing*, vol. 3, pp. 559-571, September 1994.

[10] S. Choi and J.W. Woods, "Motion Compensated 3-D Subband Coding Of Video*," IEEE Trans. Image Processing*, vol. 8, pp. 155-167, February 1999.

[11] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. on Circuits and Systems for Video Technology*, Oct 2004

[12] B. Pesquet-Popescu, and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," *ICASSP*, vol. 3, pp 1793–1796, Salt Lake City, 2001.

[13] A. Secker and D. Taubman, "Highly scalable video compression using a lifting-based 3D wavelet transform with deformable mesh motion compensation," *IEEE International Conference on Image Processing*, Vol. 3, Rochester, NY, 2002, pp. 749–752.

[14] L. Luo, J. Li, S. Li,Z. Zhuang, and Y. Zhang., "Motion compensated lifting wavelet and its application in video coding," in Proceedings of the IEEE International Conference on Multimedia and Expo 2001, Tokyo, Japan, August 2001, pp. 481-484.

[15] J.Xu, R.Xiong, B.Feng, G. Sullivan, M.-C. Lee, F. Wu, S. Li, "3D Sub-band Video Coding using Barbell lifting," ISO/IEC JTC/WG11 M10569, S05

[16] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Trans. Image Proc.*, vol. 12, Dec. 2003.

[17] J. R. Ohm, M. van der Schaar, J. W. Woods, "Interframe wavelet coding—motion picture representation for universal scalability", Signal Processing: Image Communication, Special Issue on Subband/Wavelet Interframe Video Coding, Vol:19/7, Aug 2004

[18] E. Akyol, A.M. Tekalp, M.R. Civanlar, "Motion Compansated Temporal Filtering Within the H.264/AVC Standard" , *IEEE International Conference on Image Processing 2004*

[19] _____, Optimum Scaling Operator Selection in Scalable Video Coding", *Picture Coding Symposium*, December 2004, San Francisco

[20] _____ ,"Content-Adaptive Scalability Type Selection For Bitrate Adaptation of Embedded Video", , submitted to *IEEE Trans. on Circuits Systems and Video Technology*, Jan 2005

[21] _____, "Scalable Multiple Description Video Coding with Flexible Number of Descriptions", *to be presented at IEEE International Conference on Image Processing 2005*

[22] _____,"Optimal Bit Allocation in Scalable Multiple Description Coding for Packet Loss Resilience", *to be presented at EUSIPCO 2005*

[23] M.Flierl and B.Girod, "Investigation of motion compensated lifted wavelet transforms" in *Proceedings of the IEEE International Conference on Image Processing*, 2,pp1029-1032 (Thessaloniki, Greece), Oct.2001

[24] L. Luo, F. Wu, S. Li, Z. Xiong, and Z. Zhuang, "Advanced motion threading for 3D wavelet video coding", *Signal Processing: Image Communication*, Special Issue on Subband/Wavelet Interframe Video Coding, Vol:19/7, Aug 2004.

[25] J. Xu, Z. Xiong, S. Li, Y.-Q. Zhang, Three-dimensional embedded subband coding with optimal truncation (3D ESCOT), *Appl. and Comput. Harmonic Analysis*, 10 (2001) 290–315.

[26] R. Kumar Rajendran, M. van der Schaar, S. F. Chang, "FGS+: Optimizing the Joint Spatio Temporal Video Quality in MPEG-4 Fine Grained Scalable Coding," *International Symposium on Crcuits and Systems (ISCAS*), Phoenix, Arizona, May 2002.

[27] Y. Wang, T.-T. Ng, M. van der Schaar, S.-F. Chang, Predicting optimal operation of MC-3DSBC multi-dimensional scalable video coding using subjective quality measurement. *SPIE Video Comm. and Image Processing (VCIP),* San Jose, CA, Jan. 2004.

[28] C. Kuhmünch, G. Kühne, C. Schremmer and T. Haenselmann, "A video-scaling algorithm based on human perception for spatio-temporal stimuli," *Proceedings of SPIE, Multimedia Computing and Networking (MMCN*), SPIE Press, Jan. 2001.

[29] B.-F. Hung; C.-L. Huang, Content-based FGS coding mode determination for video streaming over wireless networks, *IEEE Jour. on Selected Areas in Comm.*, vol. 21, no. 10 , pp. 1595-1603, Dec. 2003.

[30] S. Wolf and M. H. Pinson, "Spatial-Temporal Distortion Metrics for In-Service Quality Monitoring of Any Digital Video System." *Proc. of SPIE Int. Symp. on Voice, Video, and Data Communications*, Boston, MA, Sept. 11-22, 1999.

[31] E. C. Reed and J. S. Lim, "Optimal multidimensional bit-rate control for video communications," *IEEE Trans. Image Processing*, vol. 11, no. 8, pp. 873-885, Aug. 2002.

[32] A. Vetro, Y. Wang, H, Sun, "Rate-distortion optimized video coding considering frameskip," *Proc. IEEE Int. Conf. on Image Proc*., Vol. 3, pp. 534-537, Oct. 2001.

[33] Y. Wang, J.-G. Kim, and S.-F. Chang, Content-based utility function prediction for real-time MPEG-4 transcoding, *ICIP 2003,* September 14-17, 2003, Barcelona, Spain.

[34] P. Yin, A. Vetro, M Xia and B. Liu, "Rate Distortion Models for Video Transcoding", *IS&T/SPIE's Symposium on Electronic Imaging*, Santa Clara, Jan. 2003

[35] M. Masry, S.S. Hemami, "Models for the perceived quality of low bit rate video," *IEEE International Conference on Image Processing*, Rochester, NY, 2002.

[36] B. Girod, "What's wrong with mean-squared error," A.B.Watson (Ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, MA, 1993, pp. 207–220.

[37] S. Winkler: Vision Models and Quality Metrics for Image Processing Applications. *Ph.D. Thesis* #2313, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2000.

[38] S. Winkler, C. J. B. Lambrecht, M. Kunt: "Vision and Video: Models and Applications." in C.-J.-B Lambrecht (ed.), *Vision Models and Applications to Image and Video Processing,* chap. 10, Kluwer Academic Publishers, 2001.

[39] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi: "Perceptual blur and ringing metrics: Application to JPEG2000" in *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163-172, February 2004.

[40] A. A. Webster, C. T. Jones, M.H. Pinson, S.D.Voran and S. Wolf, " An objective video quality assessment system based on human perception," *Human Vision, Visual Processing, and Digital Display IV*, SPIE Proceedings, 1913, 15-26 (1993).

[41] L. Shapiro and G. Stockman, *Computer Vision*, Prentice-Hall, Upper Saddle River, N.J., 2000

[42] K.T. Tan, M. Ghanbari*, "*A multi-metric objective picture-quality measurement model for MPEG video," *Circuits and Systems for Video Technology, IEEE Transactions* on , Volume: 10 Issue: 7 Oct. 2000 Pages:1208 – 1213

[43] S. Winkler, R. Campos, "Video quality evaluation for Internet streaming applications," in Proc. SPIE, vol. 5007, pages 104–115, Santa Clara, CA, Jan. 21–24, 2003.

[44] T. Vlachos, "Detection of blocking artifacts in compressed video," Electronics Letters, 36(13):1106–1108, 2000.

[45] Z. Yu; H.R. Wu; S. Winkler.; T. Chen; "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," *Proceedings of the IEEE* , Volume: 90, Issue:1 , Jan. 2002 Pages:154 – 169

[46] F. Pan, X. Lin, S. Rahadja, W. Lin, E. Ong, S. Yao, Z. Lu and X. Yang, "A locally adaptive algorithm for measuring blocking artifacts in images and videos," *Signal Processing: ImageCommunication, vol. 19*, no. 6, pp. 499-506, July 2004.

[47] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and video clues," *IEEE Signal Processing Magazine*, vol. 17, pp. 12-36, Nov. 2000.

[48] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. on Image Processing*, vol. 12, no. 7, pp. 796-807, June 2003.

[49] S. F. Chang, P. Bocheck, "Principles and Applications of Content-Aware Video Communication," *IEEE International Symposium on Circuits and Systems (ISCAS-2000)*, Geneva, Switzerland, May 2000.

[50] S. F. Chang, D. Zhong, and R. Kumar, "Real-Time Content-Based Adaptive Streaming of Sports Video," *IEEE Workshop on Content-Based Access to Video/Image* Library, Hawaii, Dec. 2001.

[51] T. Ozcelebi, A. M. Tekalp, and R.Civanlar, "A multi-objective optimization framework for minimum delay content adaptive video streaming," *ICIP 2004, Singapore.*

[52] T. Frajka and K. Zeger, "Downsampling dependent upsampling of images," *Signal Processing: Image Communication (19)*, No. 3, March 2004, pp. 257-265.

[53] A. M. Tekalp, 'Digital Video Processing', Prentice Hall, 1995

[54] A. P. Hekstra, J.G Beerends, D.Ledermann, F.E. Caluwe, S. Kohler, R.H Koenen, S. Rihs, M Ehrsam and D. Schlauss, "PVQM: A perceptual video quality measure", *Signal Proces.: Image Communication (17)* No. 10, Nov. 2002, pp. 781-798.

[55] Test videos, available at http://home.ku.edu.tr/~eakyol/research/opt_path.htm

[56] Methodology for the Subjective Assessment of the Quality of Television Pictures, Recommendation ITU-R BT.500-10, ITU Telecom. Standardization Sector of ITU, August 2000

[57] J. Devore, "*Probability and Statistics for Engineering and the Sciences*," Duxbury Press, December 1999

[58] VQM software, available at http://www.its.bldrdoc.gov/n3/video/vqmsoftware.htm

[59] S.R Gulliver and G.Ghinea, "Changing frame rate, changing satisfaction?", *Proceedings of IEEE International Conference on Multimedia and Expo* (ICME 2004), June 2004

[60] Y. Wang, A. Reibman, and S. Lin, "Multiple Description Coding for Video Communications," *Proceedings of the IEEE*, Jan. 2005.

[61] Y. Wang and S. Lin, "Error resilient video coding using multiple description motion compensation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 438-52, June 2002

[62] D. Comás, R. Singh, A. Ortega, F. Marqués. "Unbalanced multiple description video coding based on a rate-distortion optimization". *EURASIP Journal on Applied Signal Processing*, 2003(1):81-90, January 2003

[63] M. Pereira, M. Antonini, and M. Barlaud "Multiple Description Image and Video Coding for Wireless Channels" *EURASIP Signal Processing: Image Communication Special Issue on Recent Advances in Wireless Video*, 18(10):925-945, November 2003

[64] M. Schaar and D. S. Turaga, "Multiple Description Scalable Coding Using Wavelet-Based Motion Compensated Temporal Filtering," in *IEEE International Conference on Image Processing*, vol. 3, Barcelona, Spain, September 2003

[65] Rohit Puri, Kang-Won Lee, Kannan Ramchandran and Vaduvur Bharghavan. "Forward Error Correction (FEC) Codes Based Multiple Description Coding for Internet Video Streaming and Multicast" *Signal Processing: Image Communication*, Vol.~16, No.~8, pp~745-762, May 2001

[66] J. Kim, R. M. Mersereau, and Y. Altunbasak, "Distributed video streaming using multiple description coding and unequal error protection," *IEEE Transactions on Image Processing*, July 2005

[67] J. Apostolopoulos, T. Wong, W. Tan, and S. Wee. "On Multiple Description Streaming with Content Delivery Networks" in *Proc. of IEEE INFOCOM*, June 2002.

[68] R. Rejaie, A. Ortega "PALS:Peer-to-Peer Adaptive Layered Streaming", in *Proc. of NOSSDAV'03* Monterey, CA, June 2003.

[69] X. Xu, Y. Wang, S.S. Panwar, and K.W. Ross, "A peer-to peer video-on-demand system using multiple description coding and server diversity", in *IEEE International Conference on Image Processing* 2004.

[70] Y. Shen, Z. Liu, S. P. Panwar, K. W. Ross and Y. Wang," Streaming Layered Encoded Video using Peers", *IEEE International Conference on Multimedia and Expo, 2005*

[71] V. N. Padmanabhan, H. J. Wang, and P. A. Chou, "Resilient Peer-to-Peer Streaming," *IEEE International Conference on Network Protocols*, Atlanta, GA, November 2003.

[72] R. Sood, M. van der Schaar, "Optimal Upload Policies for P2P networks in the presence of network imposed constraints", *IEEE International Conference on Acoustics Speech and Signal Processing 2005* .

[73] S.Sen, J. Wang, "Analyzing Peer-to-Peer Traffic Across Large Networks" *IEEE/ACM Transactions on Networking Volume* 12, Issue 2 ,April 2004

[74] T. Nguyen and A. Zakhor, "Multiple Sender Distributed Video Streaming" in *IEEE Transactions on Multimedia*, Vol. 6, No. 2, April 2004

[75]] D. S. Taubman and M. W. Marcellin, "JPEG 2000: Image Compression Fundamentals, Standards, and Practice" Kluwer International Series in Engineering and Computer Science,2002

[76] RFC 1889 - RTP: A Transport Protocol for Real-Time Applications

[77] O. Tickoo, S. Kalyanaraman, J. Woods, "Efficient path aggregation and error control for video streaming", *IEEE International Conference on Image Processing 2004*

[78] Y. Liang, J. G. Apostolopoulos, and B. Girod, "Analysis of Packet Loss for Compressed Video: Does Burst Loss Matter?" *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-03,* Hongkong, China, April 2003.

[79] S. Floyd, M. Handley, J. Padhye, and J. Widmer, “Equation-based congestion control for unicast applications,” presented at the *ACM SIGCOMM*, Aug. 2000.

[80] The Network Simulator NS-2 http://www.isi.edu/nsnam/ns

VITAE

Emrah Akyol graduated from Ankara Science High School as valedictorian in 1999 and obtained BS degree from Electrical Engineering Department of Bilkent University in 2003 with high honors. His research interests include image-video compression and multimedia communication. He has authored 4 international, 4 national conference publications, one submitted and one to be submitted journal papers. He is a reviewer of Signal Processing: Image Communication Journal since 2004. He is a student member of IEEE and IEEE Signal Processing Society He is going to start pursuing a Ph.D. degree on video communications in Fall 2005 at UCL A.