UNIVERSITY OF CALIFORNIA, SAN DIEGO

Wyner-Ziv Video Coding: Adaptive Rate Control, Key Frame Encoding and Correlation Noise Classification

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Ghazaleh R. Esmaili

Committee in charge:

Professor Pamela Cosman, Chair
Professor William S. Hodgkiss
Professor Laurence B. Milstein
Professor Truong Nguyen
Professor Steven Swanson

2011

The dissertation of Ghazaleh R. Esmaili is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2011

DEDICATION

To my mother, and the memory of my father
my husband, Ali, and my daughter Bahar.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Figures vii

List of Tables ix

Acknowledgements x

Vita and Publications xii

Abstract of the Dissertation xiii

Chapter 1 Introduction 1
1.1 Introduction 1
1.1.1 Slepian-Wolf Theorem 2
1.1.2 Wyner-Ziv Theorem 3
1.1.3 Low Complexity Video Encoding 4
1.1.4 Thesis Motivation and Organization 6

Chapter 2 Transform Domain Wyner-Ziv Coding 8
2.1 Introduction 8
2.2 Encoder 8
2.2.1 Quantization 9
2.2.2 Slepian-Wolf Codec 10
2.3 Decoder 10

Chapter 3 Correlation Noise 12
3.1 Introduction 12
3.2 Correlation Noise Estimation 13
3.3 Correlation Noise Classification Based on Matching Success 14
3.4 Simulation Results 17
3.5 Conclusion 20

Chapter 4 Key Frame Encoding and Hierarchical Coding 32
4.1 Introduction 32
4.2 Key Frame Encoding Based on Frequency Band Classification 33

4.2.1 Intra Coding 34
4.2.2 Coding Mode Selection and Side Information Refinement 34
4.3 Hierarchical Coding 37
4.3.1 Key frames 39
4.3.2 WZ-4 frames 39
4.3.3 WZ-2 frames 39
4.4 Simulation Results 40
4.5 Conclusion 43

Chapter 5 Adaptive Rate Control 57
5.1 Introduction 57
5.2 Dependence between Key and WZ frame quality 58
5.2.1 High motion activity 63
5.2.2 Low motion activity 63
5.3 Rate control 64
5.3.1 Motion activity classification 65
5.3.2 The relation between rate and Qstep size for key frames 66
5.3.3 Choosing Qstep size for key frames 67
5.3.4 Choosing Qstep size for WZ frames 68
5.4 Rate control simulation results 68
5.5 Subjective quality 69
5.5.1 Experimental design 73
5.5.2 Subjective Quality results 73
5.6 Conclusion 75

Chapter 6 Conclusions 76

Bibliography 79

LIST OF FIGURES

Figure 1.1: Video surveillance system 2
Figure 1.2: Block diagram of lossless distributed video coding with separate encoding and joint decoding 3
Figure 1.3: Admissible rate region for lossless distributed source coding of two sources 4
Figure 1.4: Block diagram of lossy encoding of X when Y is available at both encoder and decoder 4
Figure 1.5: Block diagram of lossy encoding of X when Y is available only at the decoder 4

Figure 2.1: Transform domain Wyner-Ziv video codec 9

Figure 3.1: Approximated Laplacian distribution for frequency band (1,2): (a) without classification; (b) for class No.1 18
Figure 3.2: PSNR vs. rate for different coding methods for *Claire* @ 30fps 19
Figure 3.3: PSNR vs. rate for different coding methods for *Mother-daughter* @ 30fps 20
Figure 3.4: PSNR vs. rate for different coding methods for *Carphone* @ 30fps 21
Figure 3.5: PSNR vs. rate for different coding methods for *Foreman* @ 30fps 22
Figure 3.6: PSNR vs. rate for different coding methods for *Claire* @ 15fps 23
Figure 3.7: PSNR vs. rate for different coding methods for *Mother-daughter* @ 15fps 24
Figure 3.8: PSNR vs. rate for different coding methods for *Carphone* @ 15fps 25
Figure 3.9: PSNR vs. rate for different coding methods for *Foreman* @ 15fps 26
Figure 3.10: PSNR vs. rate for different coding methods for *Soccer* @ 15fps 27
Figure 3.11: PSNR vs. rate for different coding methods for *Claire* @ 10fps 28
Figure 3.12: PSNR vs. rate for different coding methods for *Mother-daughter* @ 10fps 29
Figure 3.13: PSNR vs. rate for different coding methods for *Carphone* @ 10fps 30
Figure 3.14: PSNR vs. rate for different coding methods for *Foreman* @ 10fps 31

Figure 4.1: Proposed video codec with frequency band coding mode selection for key frames 35
Figure 4.2: PSNR of key frames vs. rate for different numbers of frequency bands considered as low bands 37
Figure 4.3: Proposed hierarchical coding 38
Figure 4.4: PSNR vs. rate for different coding methods for *Claire* @ 30fps 44

Figure 4.5: PSNR vs. rate for different coding methods for *Mother-daughter* @ 30fps 45
Figure 4.6: PSNR vs. rate for different coding methods for *Carphone* @ 30fps 46
Figure 4.7: PSNR vs. rate for different coding methods for *Foreman* @ 30fps 47
Figure 4.8: PSNR vs. rate for different coding methods for *Claire* @ 15fps 48
Figure 4.9: PSNR vs. rate for different coding methods for *Mother-daughter* @ 15fps 49
Figure 4.10: PSNR vs. rate for different coding methods for *Carphone* @ 15fps 50
Figure 4.11: PSNR vs. rate for different coding methods for *Foreman* @ 15fps 51
Figure 4.12: PSNR vs. rate for different coding methods for *Soccer* @ 15fps 52
Figure 4.13: PSNR vs. rate for different coding methods for *Claire* @ 10fps 53
Figure 4.14: PSNR vs. rate for different coding methods for *Mother-daughter* @ 10fps 54
Figure 4.15: PSNR vs. rate for different coding methods for *Carphone* @ 10fps 55
Figure 4.16: PSNR vs. rate for different coding methods for *Foreman* @ 10fps 56

Figure 5.1: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Claire* 59
Figure 5.2: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Mother-daughter* 60
Figure 5.3: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Foreman* 61
Figure 5.4: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Soccer* 62
Figure 5.5: Polynomial fit to data points of high and low motion convex hull 64
Figure 5.6: Rate-distortion performance of the WZ video codec applying our proposed rate control algorithm and the Discover method . 70
Figure 5.7: The achieved bit rate at each GOP for 15 fps sequences 71

LIST OF TABLES

Table 3.1: Lookup table of α parameters for 16 DCT bands of different classes 16

Table 4.1: Average fraction of time key frame high bands are Wyner-Ziv coded. 42

Table 5.1: Minimum, Maximum and Average of D_{min} over all frames of four sequences 65

Table 5.2: *Average* PSNR and Rate for key and WZ frames at different RD points of convex hull 72

Table 5.3: Five selected rates of different sequences for subjective test . . . 73

Table 5.4: Average score of subjective test over all participants for different sequences 74

Table 5.5: *Average* PSNR for five different rates for key and WZ frames . . 74

ACKNOWLEDGEMENTS

There are many people who have contributed one way or the other to make this dissertation possible. I can not possibly acknowledge all of them in a few lines.

First I offer my sincerest gratitude to my advisor, Prof. Pamela Cosman, who has supported me throughout my thesis with her vast knowledge, academic experience and constructive criticism, whilst allowing me the room to work in my own way. She was always available for discussions whenever I had questions or issues and that was very valuable and important to me.

I would like to thank my committee members, Prof. William S. Hodgkiss, Prof. Laurence B. Milstein, Prof. Truong Nguyen and Prof. Steven Swanson, for their useful comments and advice. I am very grateful to the late Prof. Jack Wolf. My dissertation is based on his theorem and I had a great joy of learning coding theory with his unforgettable teaching.

As I look further back in time, I realize that much of what I achieved would not be possible without the impact of several teachers during my studies to whom I am forever indebted: Mrs. Farhangmehr, Mrs. Mandana Mozafarinejad, Mr. Mehdizadeh, Mr. Hedayati, Dr. Ahmad Feyz, Dr. Ahmad Reza Sharafat and Dr. Mohammad Gharavi-Alkhansari. I would also like to thank Dr. Navid Lashgarian for his great help and advice to find my path to pursue my dreams.

I am especially grateful to my dear friend Mona Mahmoudi for all the emotional support, sisterhood and caring she has provided me. During the past few years I was blessed with many other great friends. In particular, I am thankful to Shadi Sagheb, Koohyar Minoo, Zeinab and Hossein Tagahvi, Hamidreza Chitsaz, Solmaz Kheiravar, Shahram Mahdavi, Sara Motamedi, Ehsan Saberian, Reza Ghanimati, Maryam Mohsennzadeh, Parham Minoo, Simin Kazemi, Mohammad Reza Gharavi, Elham Parsayan, Faryar Farokhi, Hamed, Alireza, and Maryam Masnadi, Omid Momeni, Maryam Sharif, Shahin Mehdizad and Niloofar Raisian.

I would also like to thank Ms. Zahra Tavakolian not only for being a great friend to us but also for being a great teacher and care giver with endless love to my daughter. If it was not for her, I could have never been fully concentrated on my studies without being worried about my daughter.

My deepest gratitude goes to my family. I am indebted to my mother and my father for their unconditional love, guidance, and consistent support. I can not find a suitable word to express my feeling and appreciation to my mother. I could never be where I am today if it had not been for her sacrifice, efforts and prayers. I am very thankful to my brother, Pedram, and my sisters, Mona and Saba for their love, friendship and support. I would also like to thank my in-laws Hossein, Morteza, Zohreh and Shahram for their love and kindness. And a special thanks to my mother and father in-law for their love and raising such a kind and responsible son.

Most importantly, I would like to express my sincere appreciation to my husband Ali Afsahi for his patience, understanding, encouragement and love. I could never gone this far without his great support.

The last but dearest is my daughter Bahar whom I would like to thank for her patience and filling my life with lots of hope, joy and happiness.

Chapter 3 and 4 are in part a reprint of the material in the paper: G. Esmaili and P. Cosman, "Wyner-Ziv Video Coding with Classified Correlation Noise Estimation and Key Frame Coding Mode Selection", *IEEE Transactions on Image Processing*, 2011.

Chapter 5 is in part a reprint of the material in the paper: G. Esmaili and P. Cosman, "Adaptive Rate Control for Wyner-Ziv Video Coding: Performance and Subjective Quality", *submitted to IEEE Transactions on Image Processing*, 2011.

VITA AND PUBLICATIONS

1997	B. S. in Electrical Engineering (Control), University of Tehran, Tehran, Iran.
2002	M. Sc. in Electrical Engineering (communications), Tarbiat Modares University, Tehran, Iran.
2011	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California, San Diego, United states.

G. Esmaili and P. Cosman, "Adaptive Rate Control for Wyner-Ziv Video Coding: Performance and Subjective Quality", *submitted to IEEE Transactions on Image Processing*, 2011.

G. Esmaili and P. Cosman, "Wyner-Ziv Video Coding with Classified Correlation Noise Estimation and Key Frame Coding Mode Selection", *IEEE Transactions on Image Processing*, 2011.

G. Esmaili and P. Cosman, "Frequency Band Coding Mode Selection for Key Frames of Wyner-Ziv Video Coding", *11th IEEE International Symposium on Multimedia (ISM)*, 2009.

G. Esmaili and P. Cosman, "Low Complexity Spatio-Temporal Key Frame Encoding for Wyner-Ziv Video Coding", *Data Compression Conference (DCC)*, 2009.

G. Esmaili and P. Cosman, "Correlation noise classification based on matching success for transform domain Wyner-Ziv Coding", *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

ABSTRACT OF THE DISSERTATION

Wyner-Ziv Video Coding: Adaptive Rate Control, Key Frame Encoding and Correlation Noise Classification

by

Ghazaleh R. Esmaili
Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)
University of California San Diego, 2011
Professor Pamela Cosman, Chair

Wyner Ziv (WZ) video coding is based on the results of the Slepian-Wolf and Wyner-Ziv theorems where the temporal correlation between neighboring frames is exploited at the decoder. This approach enables designing low-cost, low complexity encoders which are essential for some recent applications such as video surveillance and mobile camera phone. Despite recent advances in WZ video coding, the rate distortion performance is still far from that of predictive coding. One of the main factors affecting the WZ coding performance is the accuracy of modeling the dependency (correlation noise) between a frame to be coded and its estimation (side information) at the decoder. In existing transform domain Wyner-Ziv video coding methods, blocks within a frame are treated uniformly to estimate the correlation noise even though the success of generating side information is different for each block. This thesis proposes a method to estimate the correlation noise by differentiating blocks within a frame based on the accuracy of the side information. A second contribution of this dissertation is exploiting the temporal correlation between key frames which are usually simply intra coded. We propose a frequency band coding mode selection for key frames to exploit similarities between adjacent key frames at the decoder to improve the overall performance. Furthermore, the advantage of applying both schemes in a hierarchical order is investigated.

Rate control is an important issue in many video applications. In Wyner-

Ziv video coding architectures, the available bit budget for each GOP is shared between key frames and Wyner-Ziv frames. In this dissertation, we propose a model to express the relationship between the quantization step size of key and WZ frames based on their motion activity. Then we apply this model to propose an adaptive algorithm adjusting the quantization step size of key and WZ frames to achieve and maintain a target bit rate. The objective and subjective quality of the proposed method is evaluated.

Chapter 1

Introduction

1.1 Introduction

Raw or uncompressed video typically contains a large amount of data which requires expensive resources, such as magnetic storage or transmission bandwidth for videos. For example, an uncompressed HDTV signal needs about 1.3 Gb/s while a typical channel bandwidth might allow reliable transmission of only 20 Mb/s. Video compression has the goal of removing redundant data to reduce the number of bits used to represent the video without significantly affecting the viewer's perception of visual quality.

There are two types of redundancy in a video: spatial and temporal redundancy. Spatial redundancy is due to similarity between neighboring pixels, and temporal redundancy is due to correlation between adjacent frames. In standard video compression techniques, spatial redundancy is mainly exploited by applying a Discrete Cosine Transform (DCT), whereas temporal redundancy is exploited by motion estimation and compensation techniques. For motion estimation, a frame to be encoded is divided into small blocks. For each block, the best matching block is found in the reference frame by searching over all possible blocks within the search range. A motion vector represents the displacement between the current block and the matching block in the reference frame. So, only the the prediction error (residual between the current block and the best matching block in the reference frame) and motion vectors need to be sent. The process of finding the best

match for each block which can be done by full exhaustive search or some sub-optimal methods is very computationally intensive and makes the encoder usually 5 to 10 times more complex than the decoder [1]. This is a practical solution for downlink applications where a video is encoded once and decoded many times. But for some recent applications such as sensor networks, video surveillance, and mobile camera phones, many simple and low cost encoders are required while a high-complexity decoder can be used. For example as shown in Fig. 1.1 in each video surveillance system there are a number of encoders and just one decoder. So, the encoder which is implemented in a simple camera should be low cost and low complexity. Wyner-Ziv video coding which is founded on the Slepian-Wolf [2] and Wyner-Ziv [3] theorems is a promising solution for such applications. In this approach, the complexity is largely shifted from the encoder to the decoder by encoding individual frames independently (intraframe encoding) but decoding them conditionally (interframe decoding).

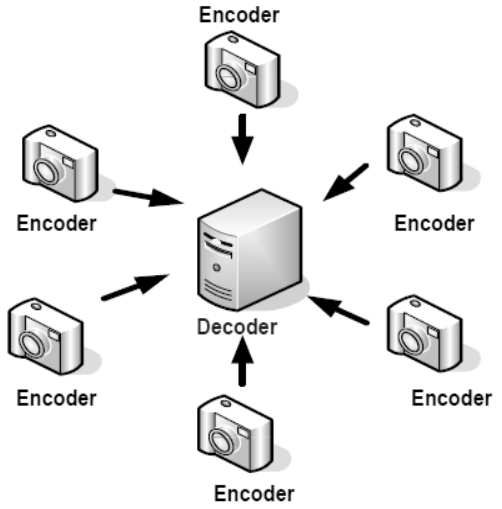


Figure 1.1: Video surveillance system

1.1.1 Slepian-Wolf Theorem

In Fig. 1.2, $\underline{X}$ and $\underline{Y}$ are statistically dependent where each of them is an independent identically distributed (i.i.d.) source. With separate encoders and decoders, the minimum achievable rates to encode $\underline{X}$ and $\underline{Y}$ losslessly are R_X and

R_Y, respectively. With separate encoders but a joint decoder, the Slepian-Wolf theorem [2] showed that R_X and R_Y should satisfy the following set of inequalities for a reconstruction of $\underline{X}$ and $\underline{Y}$ with an arbitrarily small error probability. H denotes the entropy.

$$R_X \geq H(X|Y), \quad R_Y \geq H(Y|X), \quad R_X + R_Y \geq H(X,Y) \tag{1.1}$$

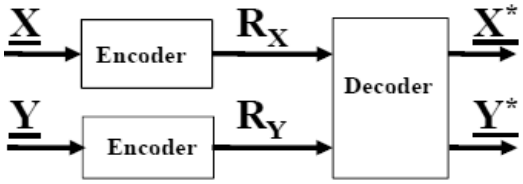


Figure 1.2: Block diagram of lossless distributed video coding with separate encoding and joint decoding

The set of points that satisfies these inequality is called the admissible rate region. Fig. 1.3 shows the admissible rate region for the pairs of rates, (R_X, R_Y).

Corners A and B are referred to as noiseless source coding with side information at the decoder only. In general, the correlated information that is available at the decoder is called side information. For example, for point A, $\underline{Y}$ which is correlated with $\underline{X}$ is the side information at the decoder. The Slepian-Wolf theorem showed that the minimum rate to encode $\underline{X}$ losslessly when $\underline{Y}$ is available only at the decoder ($R_Y = H(Y)$) is the same as the one when $\underline{Y}$ is available at both encoder and decoder, and is equal to $H(X|Y)$.

1.1.2 Wyner-Ziv Theorem

The Wyner-Ziv theorem [3] is the lossy version of the Slepian-Wolf theorem. As shown in Fig. 1.4 and Fig. 1.5, the distortion D is allowed. R_1 is the minimum rate to encode $\underline{X}$ to achieve the distortion D when $\underline{Y}$ is available at both encoder and decoder. R_2 is the minimum rate to encode $\underline{X}$ to achieve the distortion D when $\underline{Y}$ is available only at the decoder. The Wyner-Ziv theorem proved that $R_2 \geq R_1$ and the equality is possible when the two sources are jointly Gaussian and the distortion function is a squared error distortion function.

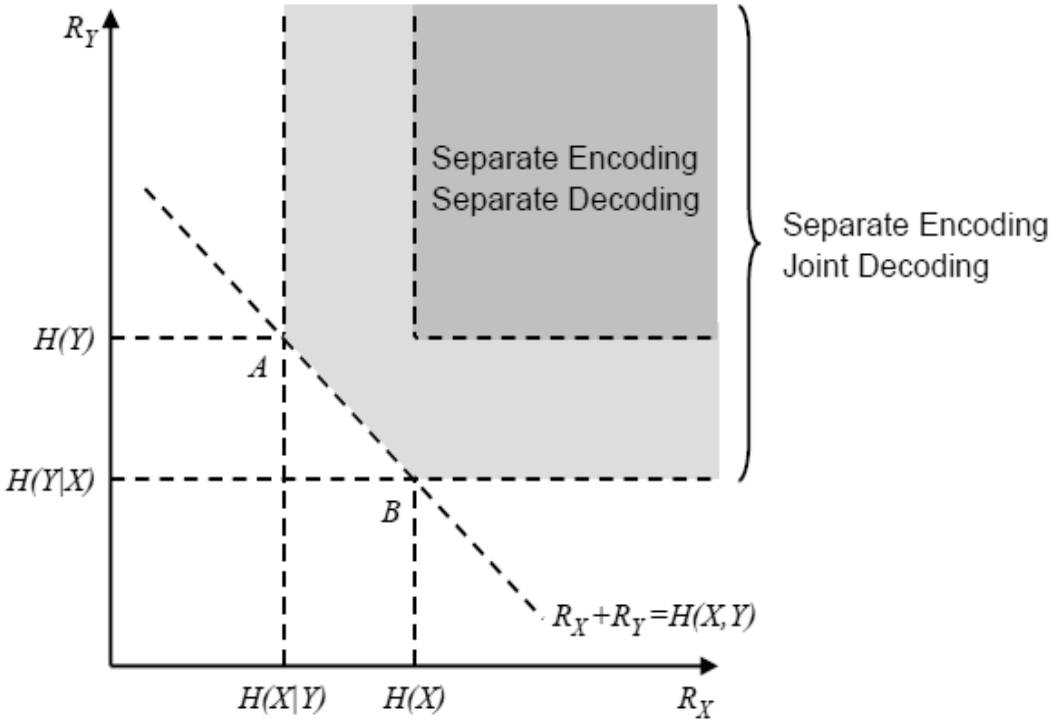


Figure 1.3: Admissible rate region for lossless distributed source coding of two sources

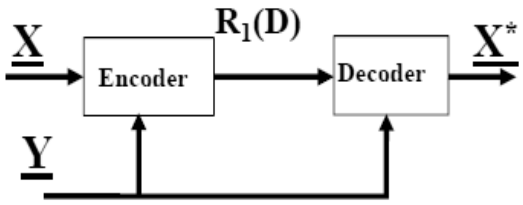


Figure 1.4: Block diagram of lossy encoding of X when Y is available at both encoder and decoder

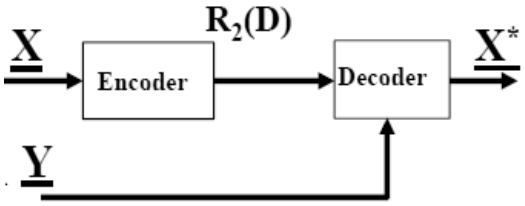


Figure 1.5: Block diagram of lossy encoding of X when Y is available only at the decoder

1.1.3 Low Complexity Video Encoding

Based on the Slepian-Wolf and Wyner-Ziv theorems, separate encoding of $\underline{X}$ and $\underline{Y}$ but joint decoding can be as efficient as jointly encoding and jointly decoding

them. In video we can consider, for example, the even frames of a sequence as $\underline{X}$ and the odd frames as $\underline{Y}$. The challenge is how to design a video coding system which encodes $\underline{X}$ and $\underline{Y}$ independently but decodes them jointly with a performance closer to conventional predictive video coding than low complexity intra coding. Two research groups introduced initial implementation of Wyner-Ziv (WZ) video coding. Puri and Ramchandran [4] proposed a syndrome-based video coding scheme which deployed block-level coding primitives, and no feedback was required. The complexity can be flexibly distributed between encoder and decoder. Motion estimation can be performed at the encoder or the decoder. It was upgraded to a more practical solution in [5]. Aaron et al. [1] proposed a feedback-required frame-based scheme using a Slepian-Wolf codec. In this approach, frames are grouped into two different classes: Wyner-Ziv (WZ) and key frames. In a GOP size of 2, key frames occur every other frame, and Wyner-Ziv frames are the frames in between. Key frames are intra coded with a conventional video intra coder and used to generate side information which is an estimation of the WZ frame to be encoded. A channel coding technique (turbo codes or LDPC) is used to encode WZ frames and correct the estimation errors in the side information. In this system, providing sufficient rate for successful decoding of WZ frames is usually achieved by using a feedback channel. They proposed their algorithm in both the pixel and transform domains [6] which became the basis for considerable further research. Improvement was achieved in [7] by sending hash codewords of the current frame to aid the decoder in accurately estimating the motion. In [8], high-frequency coefficients, which are encoded using run-length and Huffman coding, help in the decoding of the low-frequency coefficients which are encoded using Wyner-Ziv coding. In [9], Wyner-Ziv residual coding was proposed; it encodes the residual of a frame with respect to a known reference at the encoder. In [10], Brites et al. outperformed [6] by adjusting the quantization step size and applying an advanced frame interpolation for side information generation. Later, in [11] and [12], enhanced frame interpolation techniques were proposed to achieve better performance. In [13], Weerakkody et al. proposed a spatial-temporal refinement algorithm to improve the initial side information obtained by motion extrapolation.

In [14], a new side information refinement technique for transform domain Wyner-Ziv coding was proposed. In [15], Zhang et al. proposed a transform domain classification method to differentiate low motion blocks from high motion blocks to exploit additional video statistics. In [16], an iterative algorithm considered which blocks should use intra coding and which Wyner-Ziv coding. In this method, the complexity of the encoder is increased (compared to simply intra coding) by adding a buffer, generating some form of side information at the encoder, processing the iterative algorithm, and compressing mode information bits.

1.1.4 Thesis Motivation and Organization

The main goal of this thesis is to enhance the Transform Domain Wyner-ziv video coding (TDWZ) in different ways. In Chapter 2, we review the TDWZ coding adopted in this dissertation and explain its components in detail. In distributed video coding which is a special case of distributed source coding, the correlation between source and side information at the decoder should be modeled. This modeling together with applying channel coding techniques enables the decoder to exploit dependency between source and side information. Modeling the correlation between source and side information in the context of WZ video coding is called correlation noise estimation, since side information is considered to be a noisy version of the source at the decoder. In Chapter 3, after reviewing existing correlation noise estimation techniques, we describe our method to improve this modeling.

As mentioned, in WZ video coding, key frames are usually intra encoded and decoded so the inter-frame correlation between them is not exploited. In Chapter 4, we extend the idea of Wyner-Ziv coding to key frames as well to exploit their temporal correlation and improve the rate-distortion performance. In addition, we propose a hierarchical coding applying both of the proposed methods of correlation noise estimation and key frame coding.

In most existing WZ video codecs, without considering any bit rate constraint, the quantization parameters (QPs) of key frames and WZ frames are selected offline by exhaustive search to provide maximum coding efficiency with

similar quality for key and WZ frames. The offline exhaustive search approach is not viable for online applications. Also, in a video sequence, there are usually different scenes with different content and motion characteristics which are treated uniformly by this method. In Chapter 5, we first propose a method to efficiently distribute the bit budget between key and WZ frames by modeling the relationship between the quantization step size of a WZ frame and its neighboring key frames. Then, we apply this model to propose an adaptive algorithm to meet and maintain a target bit rate by dynamically adjusting the quantization parameters of key and WZ frames based on the residual energy between the WZ frame and the estimation of the side information at the encoder. The subjective quality of our proposed method is also evaluated and compared with one of the most referenced methods in the literature. Chapter 6 is reserved for the conclusion.

Chapter 2

Transform Domain Wyner-Ziv Coding

2.1 Introduction

In this chapter, we explain different components of the transform domain Wyner-Ziv (TDWZ) video codec architecture proposed in [6] which is adopted in this dissertation. As depicted in Fig. 2.1, in TDWZ, key frames are encoded and decoded by a conventional intraframe codec (H.264 in this work). The frames between them (Wyner-Ziv frames) are also encoded independently of any other frame, but their decoding makes use of other frames.

2.2 Encoder

At the encoder, a $(M \times M)$ blockwise discrete cosine transform (DCT) (8×8 or 4×4) is applied on Wyner-Ziv frames. If there are N blocks in the image, X_k (for $k = 1$ to M^2) is a vector of length N obtained by grouping together the k^{th} DCT coefficient from each block.

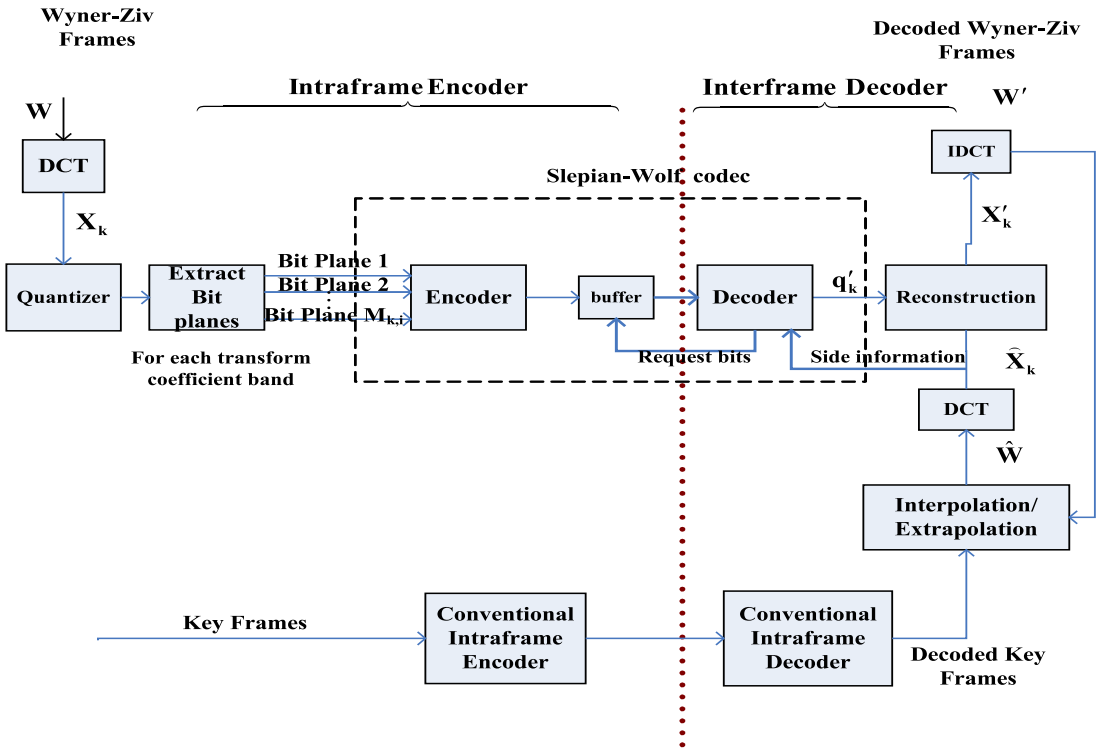


Figure 2.1: Transform domain Wyner-Ziv video codec

2.2.1 Quantization

A uniform quantizer is applied on each frequency band. Since human eyes are less sensitive to high frequency brightness variation, the amount of information in high frequency components can be reduced. A quantization matrix (QM) is used to provide finer quantization to more perceivable frequency components. We use QS_{WZ} to denote the quantization step size of a WZ frame, and it is used to quantize the DCT coefficients of WZ frames as follows:

$$Q(a_{i,j}) = round\,(\frac{a_{i,j}}{QS_{WZ} \times c_{i,j}}) \tag{2.1}$$

where $Q(a_{i,j})$ and $a_{i,j}$ are the quantized and unquantized coefficients at position (i, j), respectively. $c_{i,j}$ is the element of the QM at position (i, j).

The coefficients of X_k are quantized to form a vector of quantized symbols, q_k. That is, q_k is the vector of quantization step indices for the elements of X_k. After representing the quantized values in binary form, bit plane vectors $M_{k,i}$ $(i = 1$

to I_k) are extracted, where I_k is the maximum number of bit planes for frequency band k. The maximum number of bit planes for frequency band k is calculated by:

$$I_k = \begin{cases} \lfloor log_2 |v_k|_{max} + 1 \rfloor & \text{if } k = 1 \\ \lfloor log_2 |v_k|_{max} + 1 \rfloor + 1 & \text{otherwise} \end{cases} \tag{2.2}$$

where $|v_k|_{max}$ is the highest absolute value within frequency band k. The encoder lets the decoder know the maximum number of bit planes for each frequency band within a frame.

2.2.2 Slepian-Wolf Codec

Each bit-plane vector then enters the Slepian-Wolf encoder. In this dissertation, we have adopted the low-density parity-check accumulate (LDPCA) code from [17] which is a rate adaptive code. This code is used in conjunction with a feedback channel from the decoder. The LDPCA encoder consists of an LDPC syndrome-former following by an accumulator. The accumulated syndrome bits are stored in the buffer and sent in chunks (upon decoder request through the feedback channel) until a desired bit error rate is met. We assume ideal error detection.

2.3 Decoder

At the decoder, $\hat{W}$ is the estimate of W (Wyner-Ziv frame) which is generated by applying extrapolation or interpolation techniques on decoded key frames. For a GOP of size 2, a motion compensation frame interpolation (MCFI) technique as explained in [18] is applied on the previous and next key frames to estimate the Wyner-Ziv frame in between.

A blockwise $M \times M$ DCT is applied on $\hat{W}$ to produce $\hat{X}$. $\hat{X}_k$, the side information corresponding to X_k, is generated by grouping the transform coefficients of $\hat{X}$. When all the bit-planes are decoded, the bits are regrouped to form a vector

of reconstructed quantized symbols, $\acute{q}_k$. At the end, the reconstructed coefficient band $X\acute{}_k$ is calculated as $E(X_k|\acute{q}_k, \hat{X}_k)$.

Chapter 3

Correlation Noise

3.1 Introduction

The Wyner-Ziv decoder needs some model for the statistical dependency between the source and the side information to make use of the side information. The statistical model is necessary for the conditional probability calculations in the Slepian-Wolf (Turbo or LDPC) decoder as well as for the conditional expectation in the reconstruction block. Accurate modeling of correlation has a strong impact on performance by exploiting the statistics between source and side information [19]. The dependency between source and side information is modeled by $Y = X + Z$ where Y denotes the side information and X denotes the source. Z is called the correlation noise. In [20], the correlation noise was modeled by Laplacian, Guassian, two sided Gamma, and Generalized Guassian distributions and the relationship between the compression ratio and sensitivity of the estimated channel model parameter was investigated. In most approaches, the perobability density function (pdf) of Z is approximated by a Laplacian distribution and its corresponding parameters are estimated by plotting the residual histogram of several sequences. In these methods, the estimated Laplacian parameter is the same for all blocks within a frame, even though the accuracy of the side information varies based on the MCFI success. In [21], a method was proposed to estimate the pixel domain correlation noise by online adjustment of the Laplacian parameter for each block. In [22] and [23], some methods at frame, block, and pixel lev-

els were suggested for online parameter estimation of pixel and transform domain Wyner-Ziv coding. Their proposed method for transform domain correlation noise estimation was improved by Huang et al. in [24] by utilizing cross-band correlation. In this chapter, we propose a simple and effective method to estimate the correlation noise based on MCFI success at the decoder. We define a practical criterion to evaluate the block matching success at the decoder. Also, we make use of block matching information in an efficient way to elevate the accuracy of correlation noise estimation.

This chapter is organized as follows. In Section 3.2, we explain the most common approach in the literature to estimate the correlation noise. In Section 3.3, we propose a simple and effective method to differentiate blocks within a frame to estimate the correlation noise based on MCFI success at the decoder. Simulation results are presented in Section 3.4. Section 3.5 concludes this chapter.

3.2 Correlation Noise Estimation

In most existing methods, the Slepian-Wolf decoder and reconstruction block assume a Laplacian distribution to model the statistical dependency between X_k and $\hat{X}_k$. Although more accurate models such as a generalized Gaussian can be applied, the Laplacian is selected for good balancing of accuracy and complexity. The distribution of d can be approximated as

$$f(d) = \frac{\alpha}{2} e^{-\alpha|d|} \tag{3.1}$$

where d denotes the difference between corresponding elements of X_k and $\hat{X}_k$. In existing approaches [1] to [10], a different α parameter is assigned for each frequency band. These α parameters are estimated by plotting the residual histogram of several sequences using motion compensated frame interpolation (MCFI) for the side information. For example, for frequency band k, the differences between corresponding elements in X_k and $\hat{X}_k$ of several sequences are grouped to form a set F_k. The α parameter is calculated by $\frac{\sqrt{2}}{\sigma_k}$ where σ_k is the square root of the

variance of the F_k values.

3.3 Correlation Noise Classification Based on Matching Success

More accurate estimation of the dependency between source and side information means that fewer accumulated syndrome bits need to be sent, resulting in improved rate-distortion performance. Traditional estimation of Laplacian distribution parameters treats all frames and blocks within a frame uniformly, even though the quality of the side information varies spatially and temporally. General MCFI methods are based on the assumption that the motion is translational and linear over time among temporally adjacent frames. This assumption often holds for relatively small motion, but tends to give a poor estimation for high motion regions. The general approach to estimate a given block B in the interpolated frame F_t is to find the motion vector of the co-located block in F_{t+1} with reference to frame F_{t-1} where t, $t-1$ and $t+1$ are time indexes. In [25], the motion vectors obtained by block matching in the previous step are refined by a bidirectional motion estimation technique. A spatial smoothing algorithm is then used to improve the accuracy of the motion field. If $v = (v_x, v_y)$ is the final motion vector, where v_x and v_y are the x and y components of v, then the interpolated block is obtained by averaging the pixels in F_{t-1} and F_{t+1} pointed to by $\frac{v}{2}$ and $-\frac{v}{2}$. These blocks of pixels in F_{t-1} and F_{t+1} are called forward and backward interpolations, and the interpolated block is calculated as their average:

$$F_t(x,y) = \frac{F_{t-1}(x+\frac{v_x}{2}, y+\frac{v_y}{2}) + F_{t+1}(x-\frac{v_x}{2}, y-\frac{v_y}{2})}{2} \qquad (x,y)\epsilon B \qquad (3.2)$$

The residual energy between the forward and backward interpolations is:

$$E = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} [F_{t-1}(x+\frac{v_x}{2}, y+\frac{v_y}{2}) - F_{t+1}(x-\frac{v_x}{2}, y-\frac{v_y}{2})]^2 \qquad (3.3)$$

where M and N represent the block size (in our case $M = N = 4$). In [23], the residual between forward and backward interpolations was applied to estimate the correlation noise. $R(x, y)$ is the residual frame and is calculated as

$$R(x,y) = \frac{F_{t-1}(x + \frac{v_x}{2}, y + \frac{v_y}{2}) - F_{t+1}(x - \frac{v_x}{2}, y - \frac{v_y}{2})}{2} \tag{3.4}$$

$T(u, v) = DCT(R(x, y))$ is defined. The α parameter for frequency band b and frame s is $\alpha_{b,s} = \frac{\sqrt{2}}{\sigma_{b,s}}$ where $\sigma_{b,s}$ is the square root of the variance of the elements of $|T|$. At the coefficient level, to have more accurate correlation noise estimation, each coefficient of frame $|T|$ was classified into inlier or outlier classes. As explained in [23], inlier coefficient values are close to the corresponding DCT band average value $\widehat{\mu}_b^2$. Outlier coefficients are those whose value is far from $\widehat{\mu}_b^2$. The α parameter for inlier coefficients was taken to be $\sigma_{b,s}$ which was the frame level α parameter. The α parameter for outlier coefficients was taken to be $\sqrt{\frac{2}{[D_b(u,v)]^2}}$ where

$$D_b(u,v) = |T|_b(u,v) - \widehat{\mu}_b \tag{3.5}$$

With this approach for blocks/regions where the residual error is high, $[D_b(u, v)]^2$ is used instead of $\sigma_{b,s}$ to give less confidence to areas where MCFI is less successful. But for well interpolated blocks/regions, coefficient level estimation is not better than frame level estimation. In our method, every block within a frame is classified in order to estimate the correlation noise. By a training stage and offline classification, we are able to estimate the dependency between source and side information based on the residual energy of a given block. We give different levels of confidence to different blocks based on how well interpolated they are.

In our method, we divide our sample of data into several classes of residual energy. The residual energy between forward and backward interpolation of every block within a frame for all Wyner-Ziv frames of several sequences is calculated to form a set R. We classify elements of this set into m different classes using $m - 1$

thresholds T_i where $i \in \{1, ..., m-1\}$. Class i is chosen when $T_i < r < T_{i+1}$ where $r \in R$. To help ensure statistically reliable classification, the threshold values are set such that classes have roughly the same numbers of elements. All coefficients corresponding to frequency band j of all blocks labeled with class i are grouped together to form a set $v_{i,j}$. The α parameter of the set $v_{i,j}$ is calculated by $\frac{\sqrt{2}}{\sigma_{i,j}}$ where $\sigma_{i,j}$ is the square root of the variance of the $v_{i,j}$ elements.

Based on the above procedure, there are m different classes of correlation estimation for each frequency band. We have therefore an m by 16 (since a 4×4 DCT is applied) lookup table of α parameters at the decoder. The component i, j of this table represents the α parameter of frequency band j of class i where $i \in \{1, ..., m\}$ and $j \in \{1, ..., 16\}$. Development of this table is done offline.

During decoding, for a given block of the Wyner-Ziv frame, the decoder evaluates the matching success of MCFI by calculating the residual energy between forward and backward interpolation and chooses one of the defined m classes by comparing to the threshold values. Once the block class is determined, the α

Table 3.1: Lookup table of α parameters for 16 DCT bands of different classes

class	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$
1	1.1	1.3	1.4	1.7	1.3	1.9	2.0	2.2	1.7	2.2	2.3	2.6	2.1	2.6	2.7	3.1
2	1.0	1.1	1.1	1.2	1.1	1.4	1.4	1.6	1.2	1.6	1.7	2.0	1.5	1.9	2.1	2.4
3	0.8	1.0	1.0	1.0	0.9	1.1	1.1	1.2	1.0	1.1	1.3	1.5	1.2	1.5	1.6	1.9
4	0.6	0.9	0.9	0.9	0.7	0.9	0.9	1.0	0.7	0.9	1.0	1.2	0.9	1.1	1.2	1.5
5	0.4	0.6	0.6	0.6	0.4	0.6	0.6	0.7	0.5	0.6	0.6	0.8	0.6	0.7	0.8	1.0
6	0.2	0.4	0.4	0.5	0.3	0.4	0.5	0.5	0.4	0.5	0.5	0.6	0.4	0.5	0.6	0.7
7	0.2	0.3	0.3	0.4	0.2	0.3	0.4	0.4	0.3	0.3	0.4	0.5	0.4	0.4	0.4	0.5
8	0.1	0.2	0.2	0.3	0.1	0.2	0.2	0.3	0.2	0.2	0.2	0.3	0.2	0.3	0.3	0.3
Unique	0.2	0.3	0.4	0.4	0.3	0.4	0.4	0.5	0.3	0.4	0.4	0.5	0.4	0.5	0.5	0.6

parameter of each frequency band is found through the lookup table. In our simulation, the number of classes is set to 8 since in that case, as discussed below, we have enough elements in each class to have a reliable distribution model. Threshold values are calculated offline for each quantization parameter, separately. Table 3.1 shows the computed lookup table for quantization parameter equal to 0.4. Each row represents the α parameter of different DCT bands of a given class. The last

row represents the calculated α parameter of different DCT bands based on the existing method where there is no classification. As we can see, going from class 1 down to class 8 in each column, the α parameter of each DCT band is a monotonically decreasing function of residual energy satisfying our expectation. Also, the α parameter of each class is an increasing function of frequency in each direction meaning that the α parameters of $f_{i,j}, f_{i,j+1}, ..., f_{i,j+3}$ and $f_{i,j}, f_{i+1,j}, ..., f_{i+3,j}$ are monotonically increasing. This suggests we have sufficient data within each class, since the α parameters follow the same trends as they do when there is no classification.

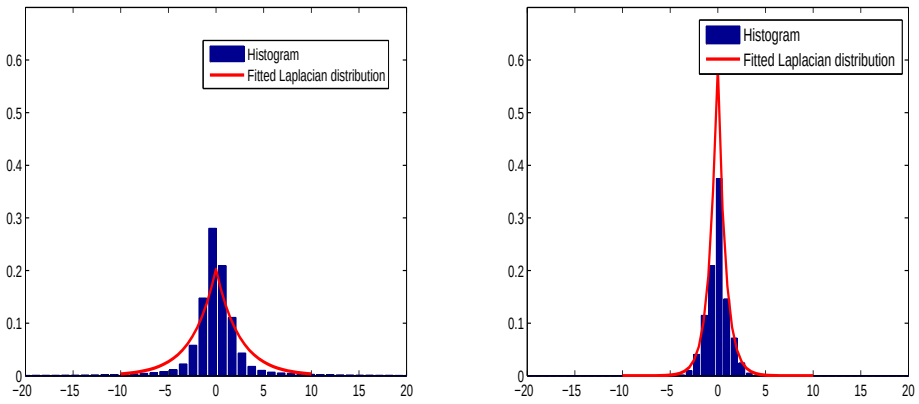
As shown in Table 3.1, the α parameters of the last row (corresponding to no classification) lie between class 6 and class 7. So for high motion sequences with most blocks classified to class 6 or higher, we expect less improvement than for low motion sequences with most blocks classified to class 5 or lower.

Fig. 3.1 (a) and (b) show the distribution of frequency band $f(1,2)$ corresponding to the traditional method (no classification) and corresponding to class 1, respectively. As we can see, the width of the approximated Laplacian distribution for frequency band $f(1,2)$ of class 1 is smaller than the width of the distribution for the traditional method, meaning that the prediction will be more accurate on average when using the classification.

3.4 Simulation Results

Fig. 3.2-3.5, Fig. 3.6-3.9 and Fig. 3.11-3.14 show the rate-distortion performance for the test sequences *Claire*, *Mother-daughter*, *Foreman* and *Carphone* QCIF (176×144) sequences at 30 frame per second (fps), 15fps and 10fps, respectively. Fig. 3.10 shows the rate-distortion performance for the *Soccer* QCIF sequence at 15 fps.

In all offline processes such as setting threshold values and correlation noise classification lookup tables, training video sequences are different from test video sequences. Our training sequences are *Container*, *Salesman*, *Coastguard* and *Akiyo*. In our simulation, a 4×4 blockwise discrete cosine transform (DCT) is



(a) (b)

Figure 3.1: Approximated Laplacian distribution for frequency band (1,2): (a) without classification; (b) for class No.1

applied on Wyner-Ziv frames. $QP\epsilon\{0.4, 0.85, 1.5, 2, 3, 3.5, 4\}$. The quantization matrix applied in our simulation is the initializing quantization matrix borrowed from H.264 JM 9.6, as follows:

$$C = \begin{bmatrix} 6 & 12 & 19 & 26 \\ 12 & 19 & 26 & 31 \\ 19 & 26 & 31 & 35 \\ 26 & 31 & 35 & 39 \end{bmatrix}$$

Threshold values $[T_1, T_2, ..., T_7] = [200, 290, 740, 860, 1200, 1500, 1800]$ corresponding to quantization parameters $QP = [0.4, 0.85, 1.5, 2, 3, 3.5, 4]$, were chosen as they work well for the training sequences with different characteristics.

"*Conventional*" is based on the method in [10] but we modified the algorithm in two ways. First, the assumption of availability of original key frames at the decoder is removed since it is not valid from a practical point of view. Second, the quantization part is replaced with the quantization procedure explained above. Our quantization method is applied for all proposed methods. We use the same quantization method for all the approaches in order to highlight the performance improvement due to correlation noise classification and key frame encoding. When Wyner-Ziv coding equipped with correlation noise classification is

applied for Wyner-Ziv frames of the conventional method, the method is called "$Conventional - WZ+$".

Simulation results show that applying the correlation noise classification proposed in Section 3.4 results in up to 2 dB improvement over "Conventional" and 1 dB improvement over the best proposed method (coefficient level) in [23] (*Claire* 10 fps at 240 Kbps).

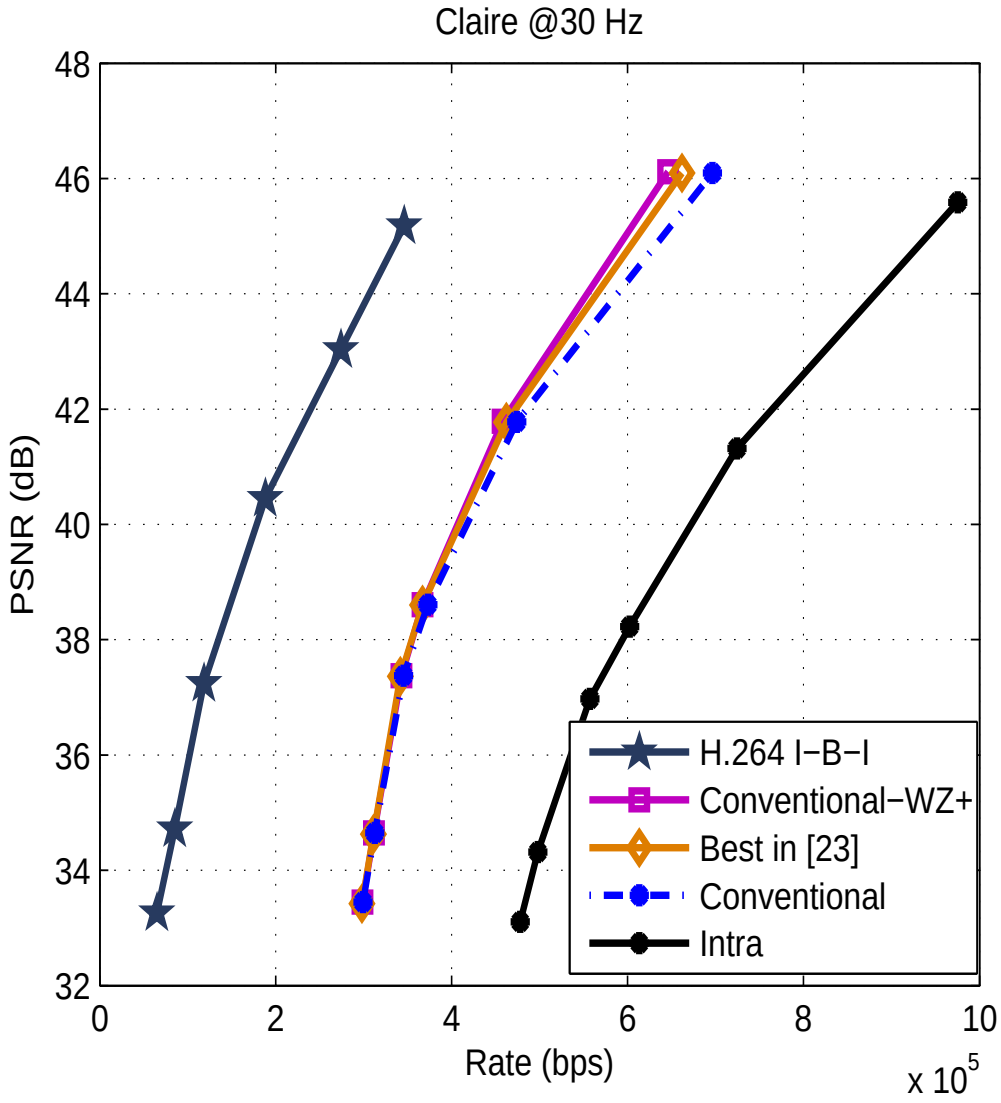


Figure 3.2: PSNR vs. rate for different coding methods for *Claire* @ 30fps

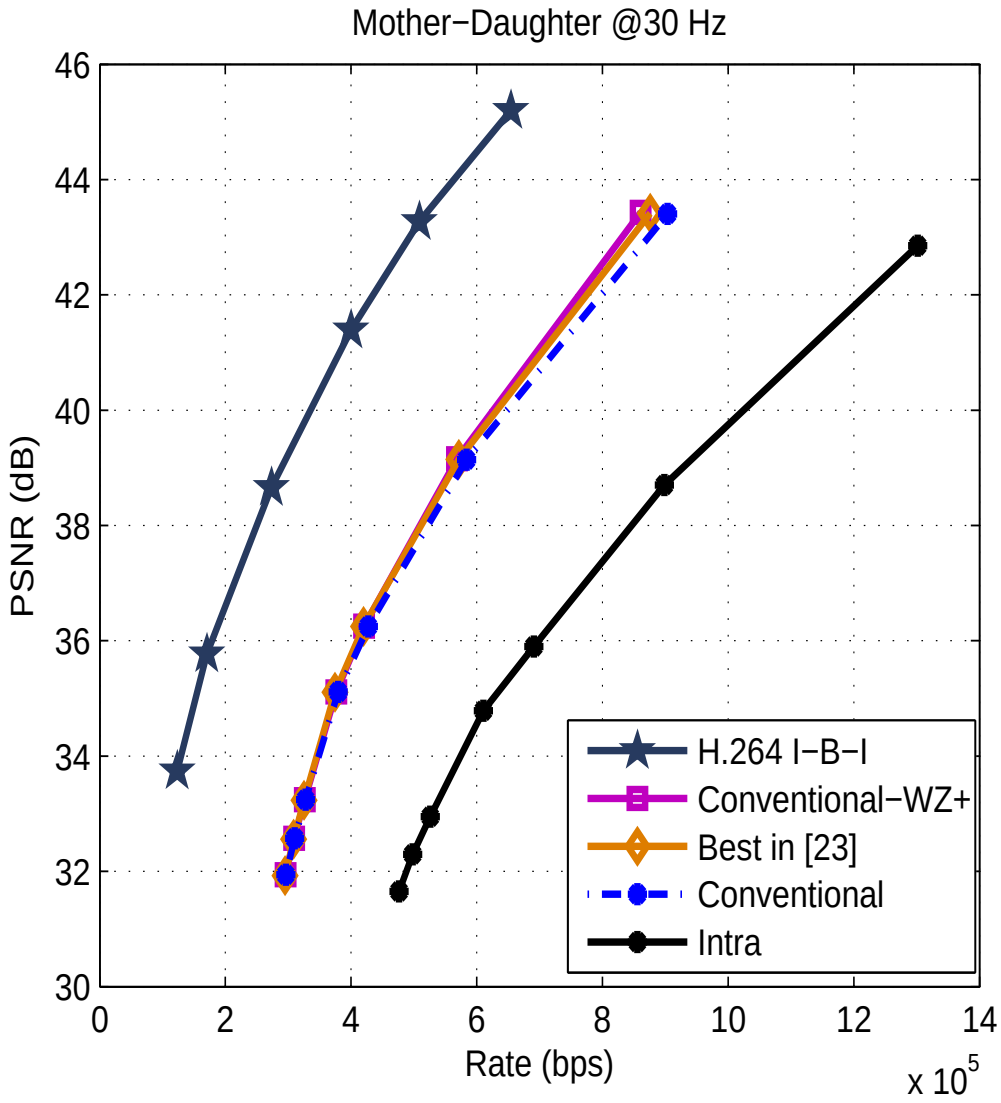


Figure 3.3: PSNR vs. rate for different coding methods for *Mother-daughter* @ 30fps

3.5 Conclusion

We proposed a new method of correlation noise estimation for TDWZ based on block matching classification at the decoder. We were able to exploit additional statistical dependency between source and side information by using the residual energy between forward and backward interpolation as the matching criterion. The proposed approach does not increase the complexity at the encoder. Simulation

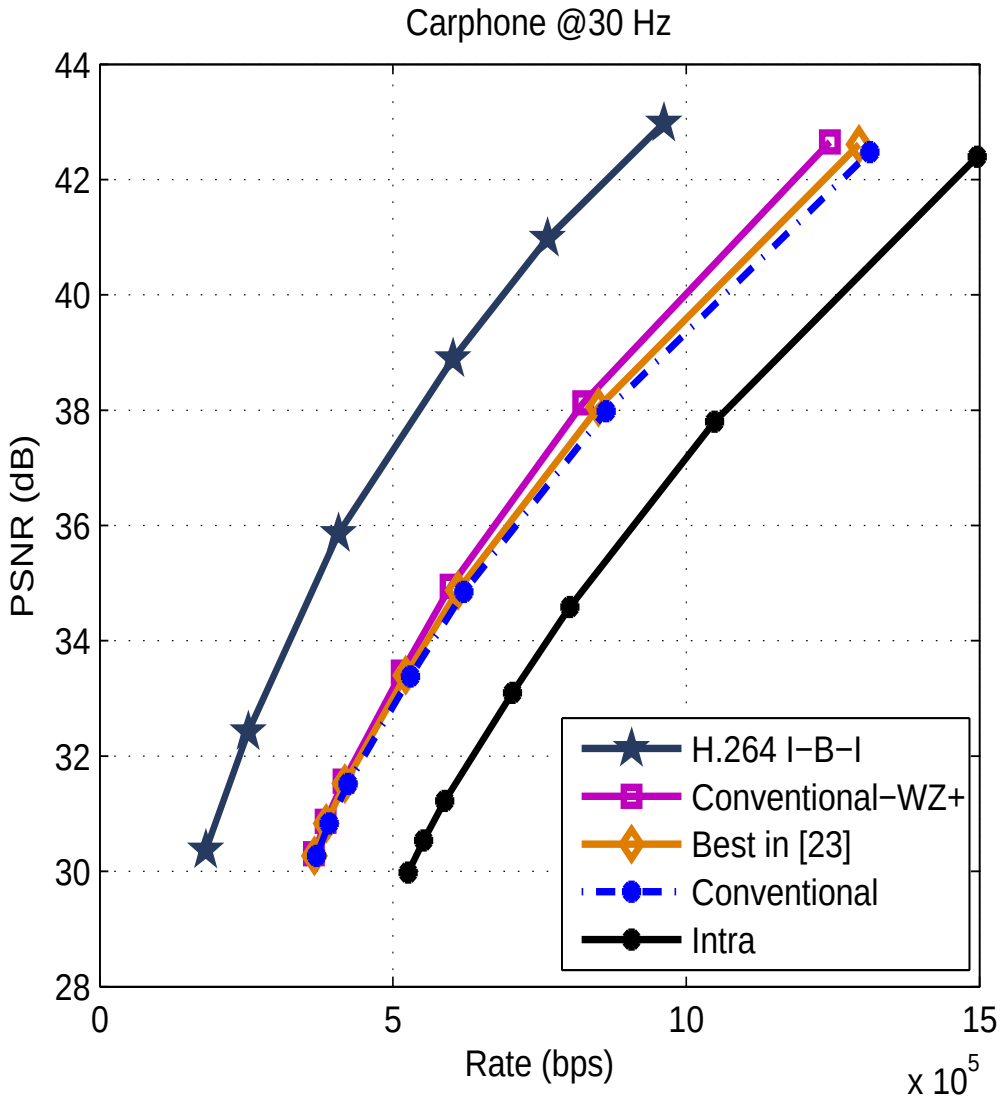


Figure 3.4: PSNR vs. rate for different coding methods for *Carphone* @ 30fps

results show up to 2 dB improvement over "Conventional" and 1 dB improvement over the best proposed method (coefficient level) in [23] (*Claire* 10 fps at 240 Kbps).

This chapter is in part a reprint of the material in the paper: G. Esmaili and P. Cosman, "Wyner-Ziv Video Coding with Classified Correlation Noise Estimation and Key Frame Coding Mode Selection", *IEEE Transactions on Image Processing*, 2011.

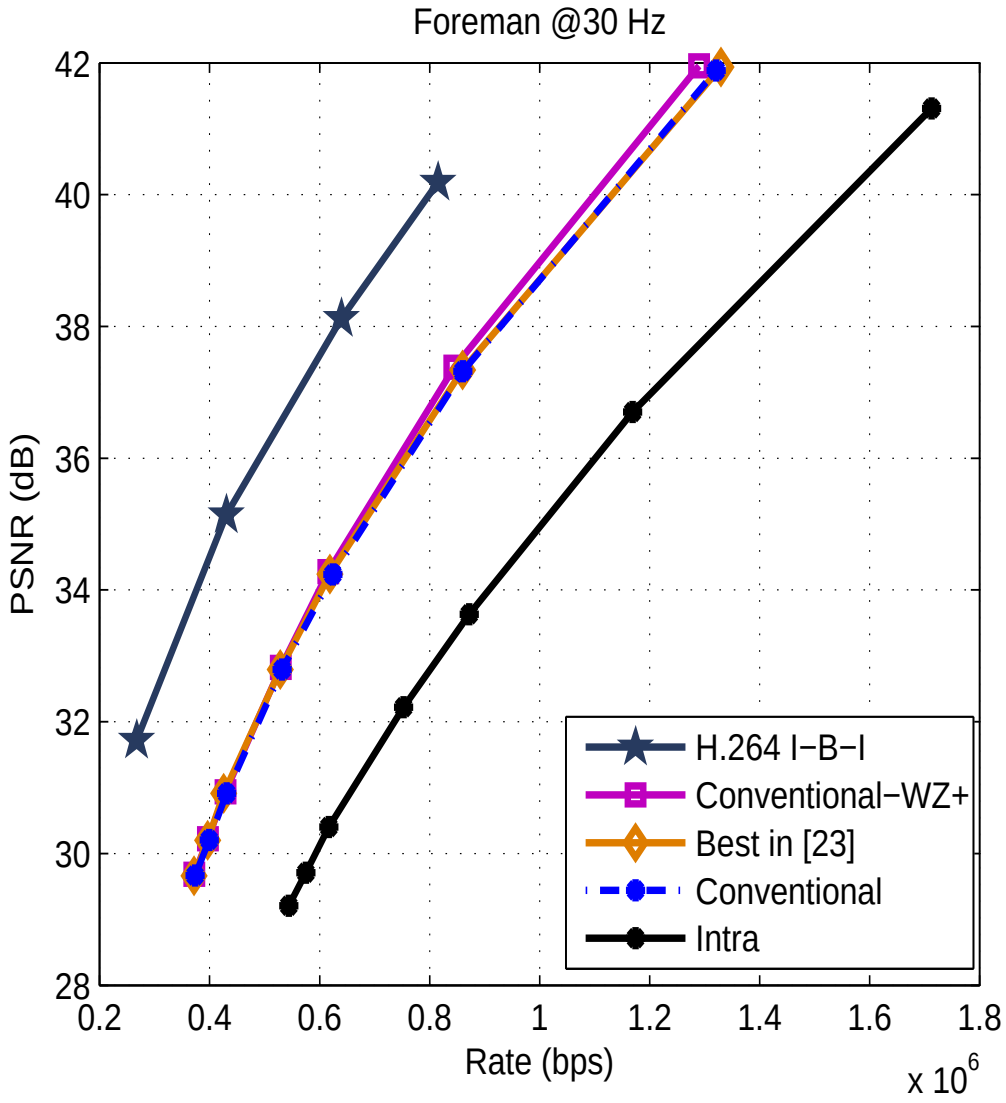


Figure 3.5: PSNR vs. rate for different coding methods for *Foreman* @ 30fps

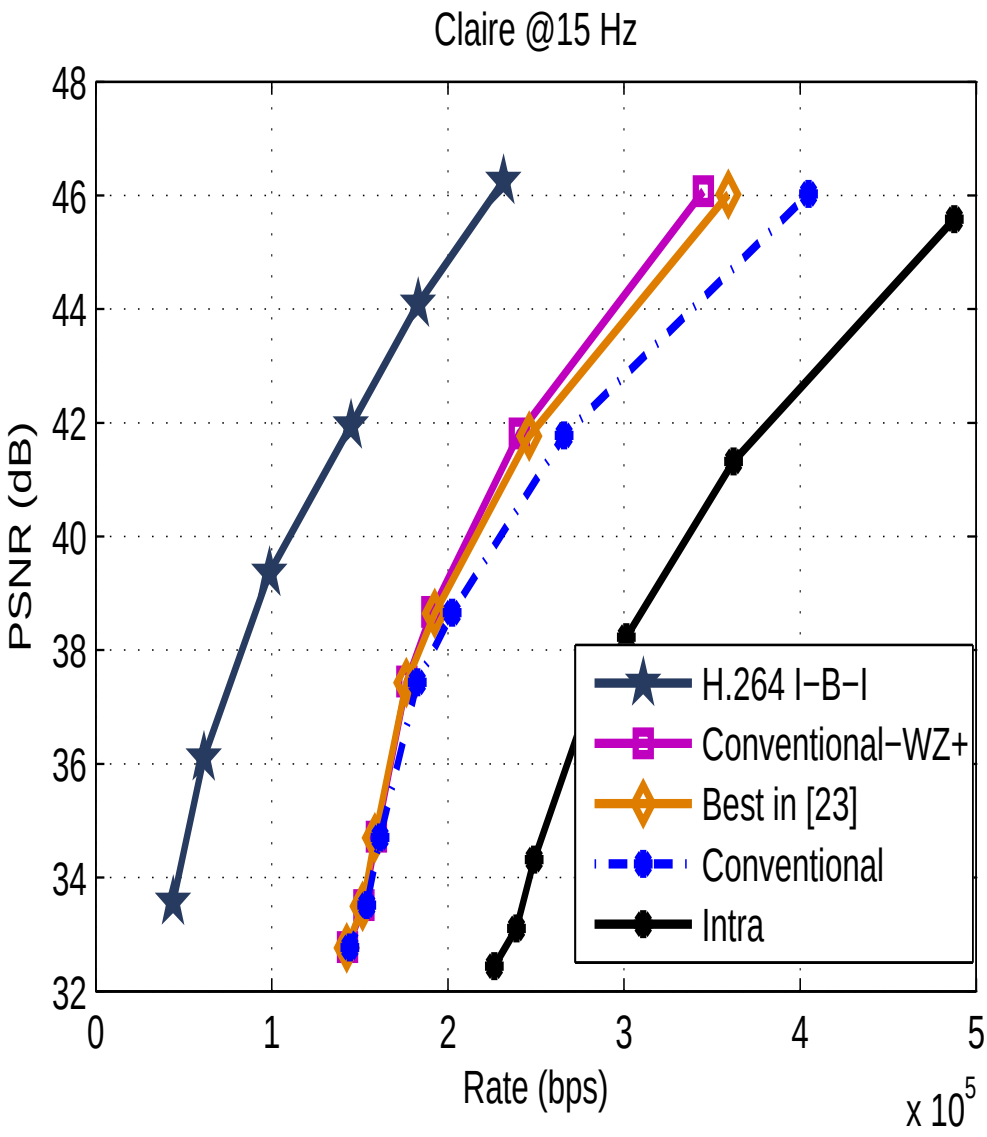


Figure 3.6: PSNR vs. rate for different coding methods for *Claire* @ 15fps

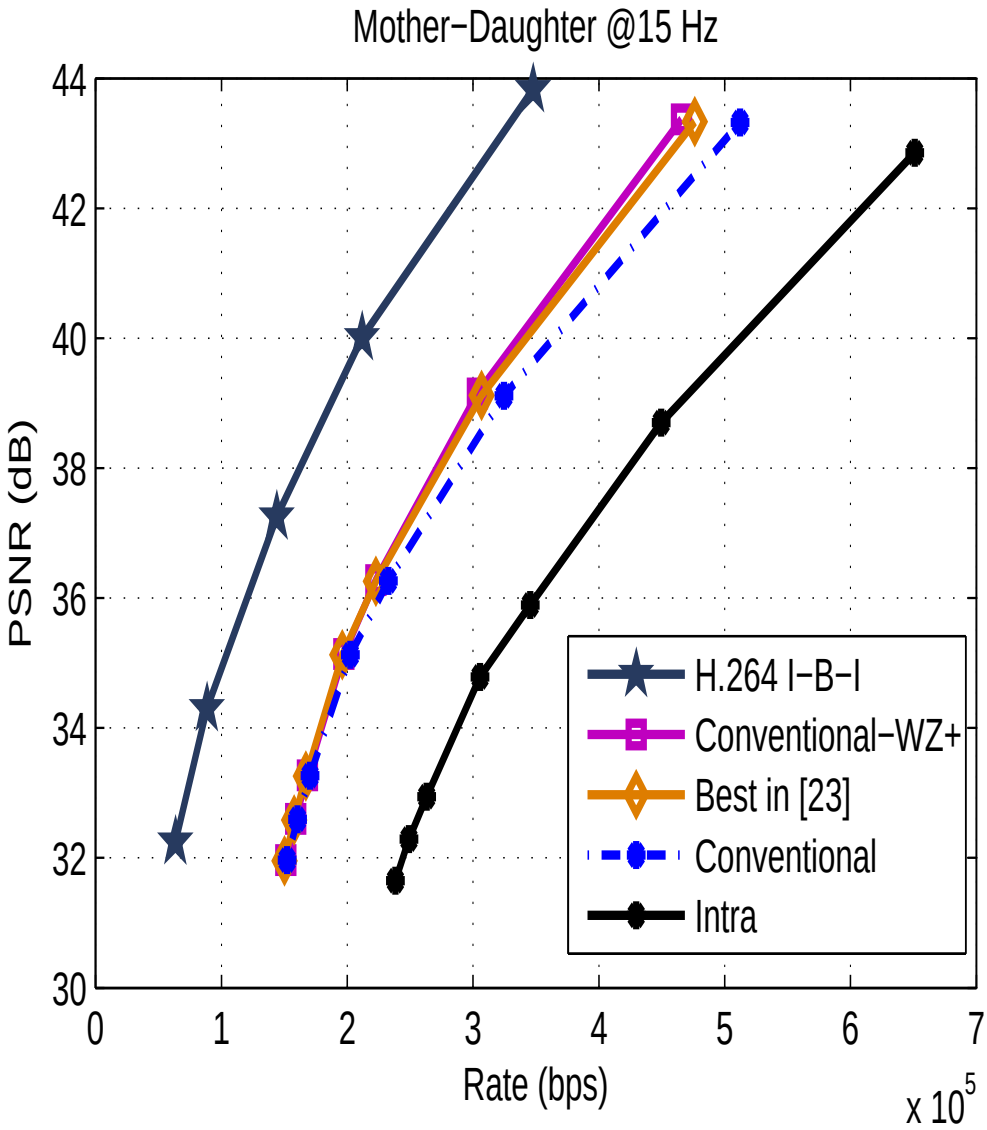


Figure 3.7: PSNR vs. rate for different coding methods for *Mother-daughter* @ 15fps

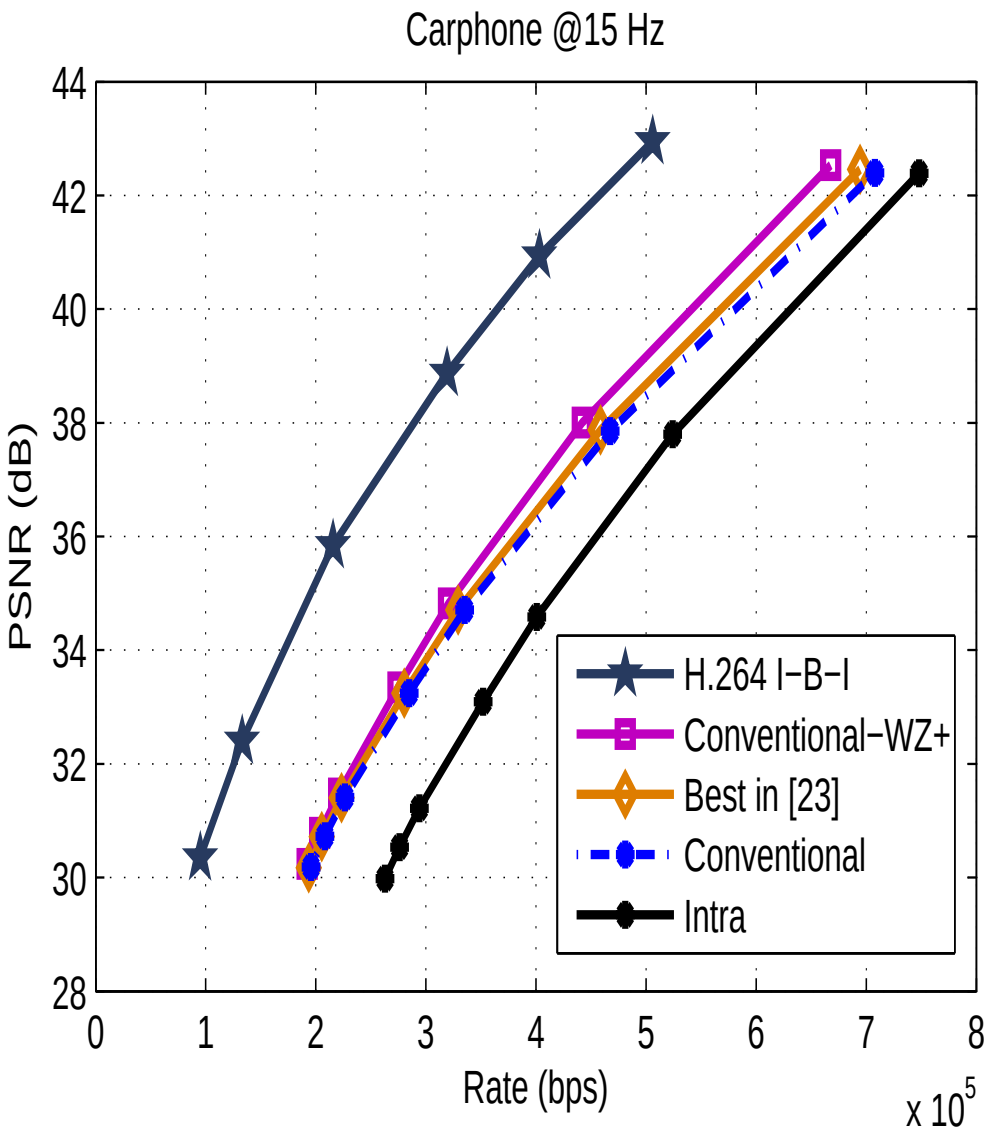


Figure 3.8: PSNR vs. rate for different coding methods for *Carphone* @ 15fps

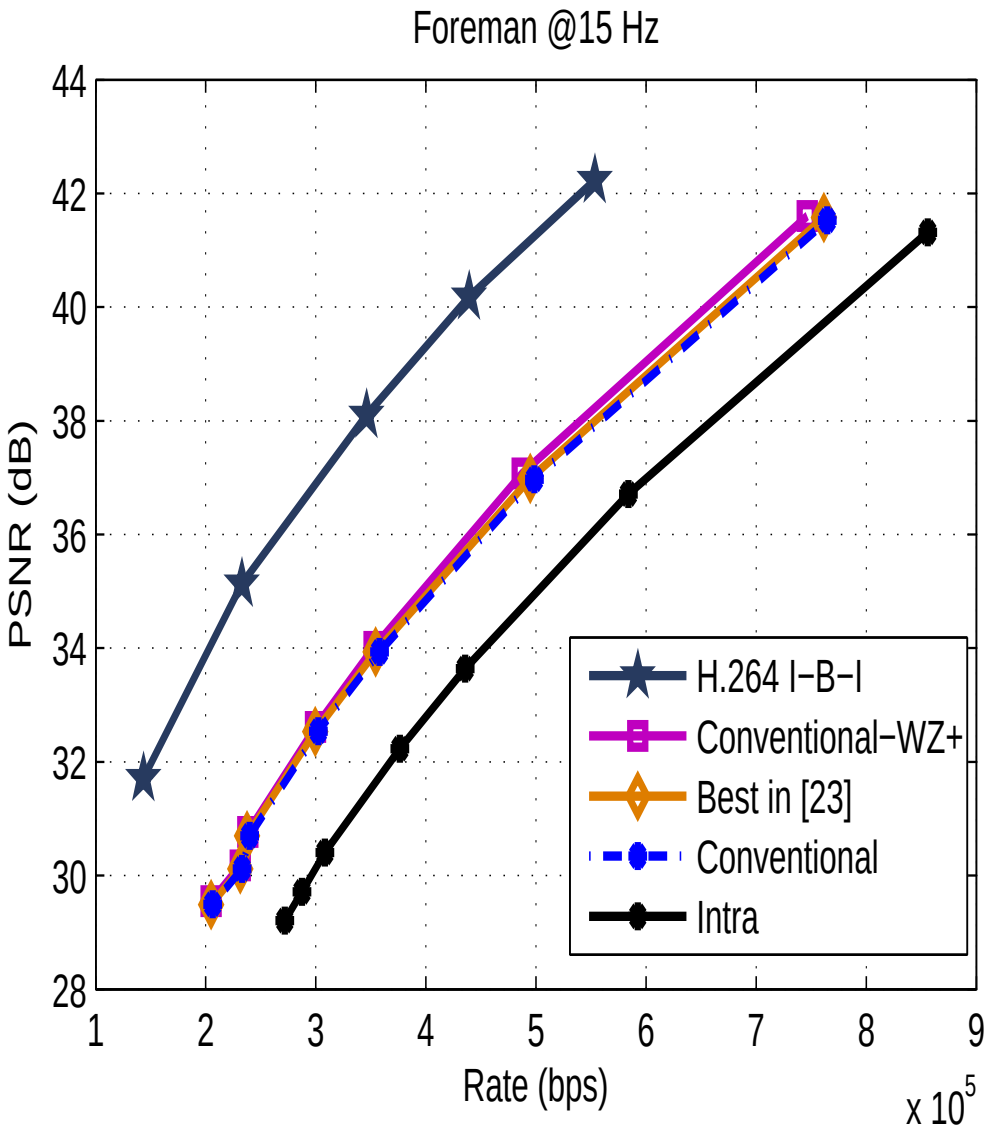


Figure 3.9: PSNR vs. rate for different coding methods for *Foreman* @ 15fps

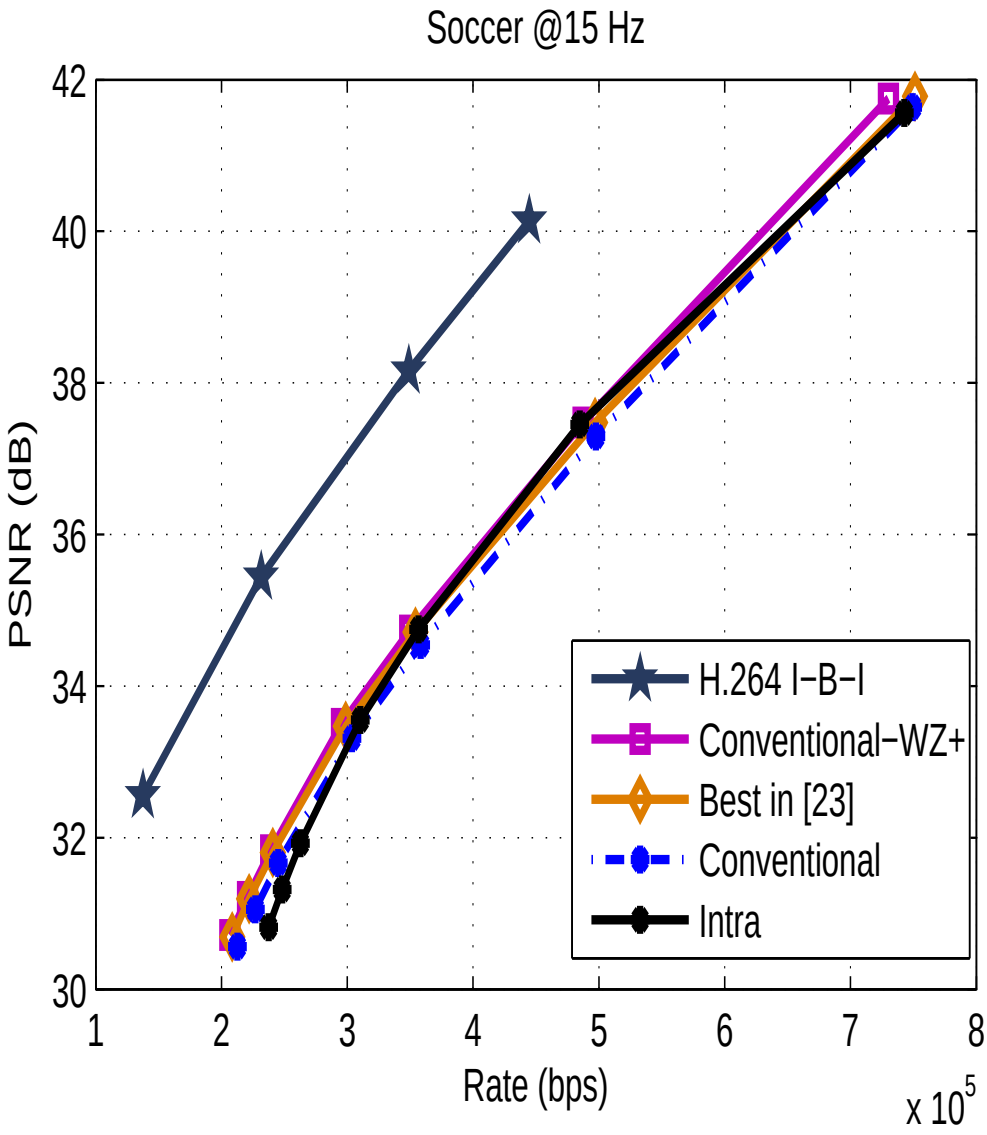


Figure 3.10: PSNR vs. rate for different coding methods for *Soccer* @ 15fps

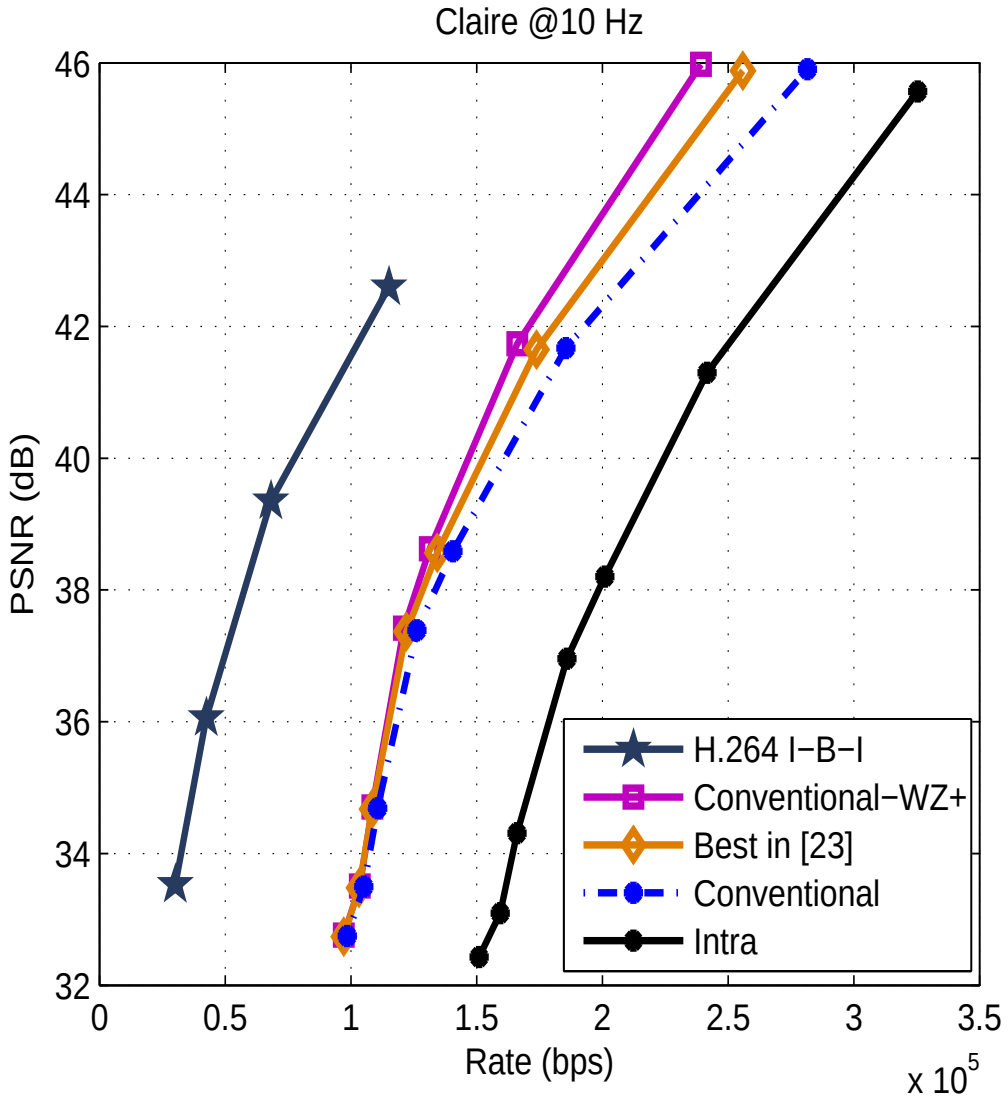


Figure 3.11: PSNR vs. rate for different coding methods for *Claire* @ 10fps

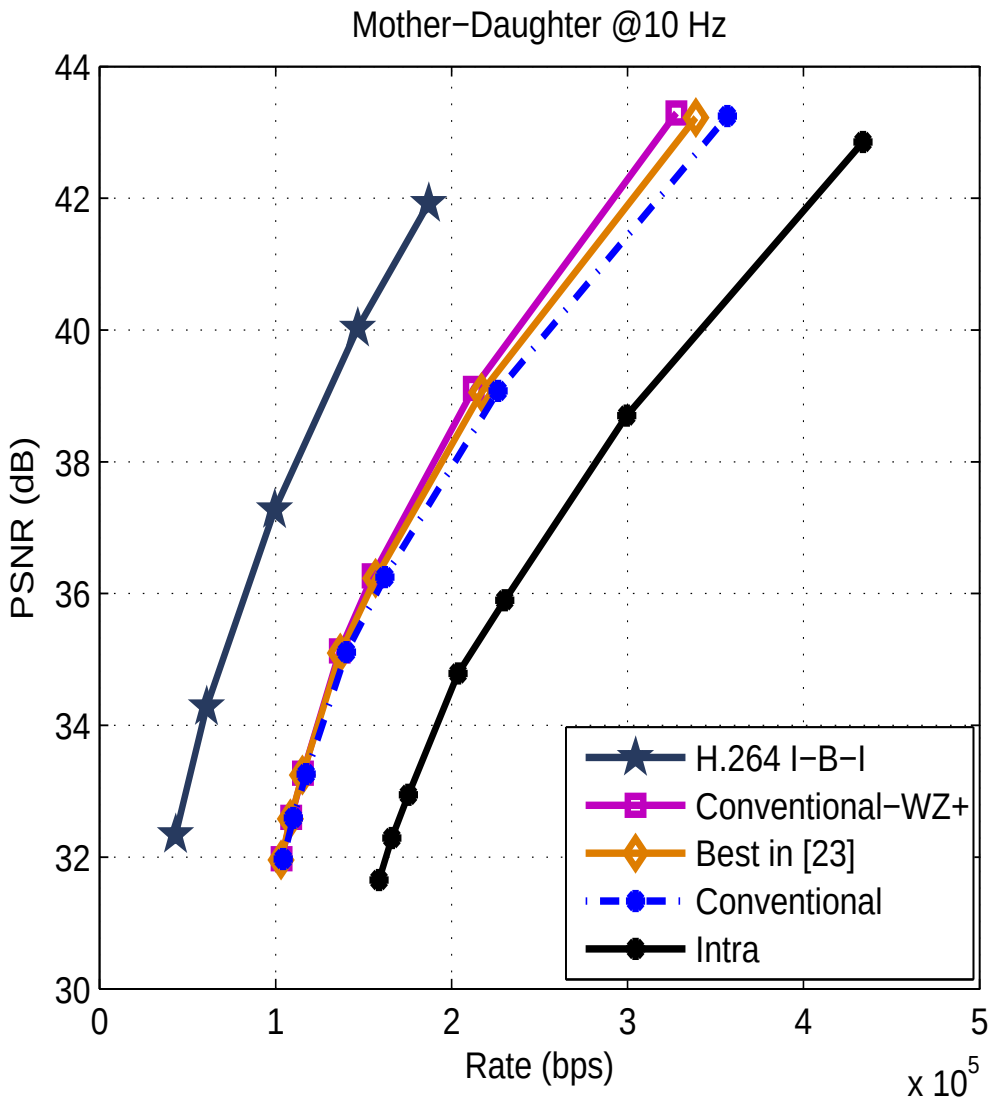


Figure 3.12: PSNR vs. rate for different coding methods for *Mother-daughter* @ 10fps

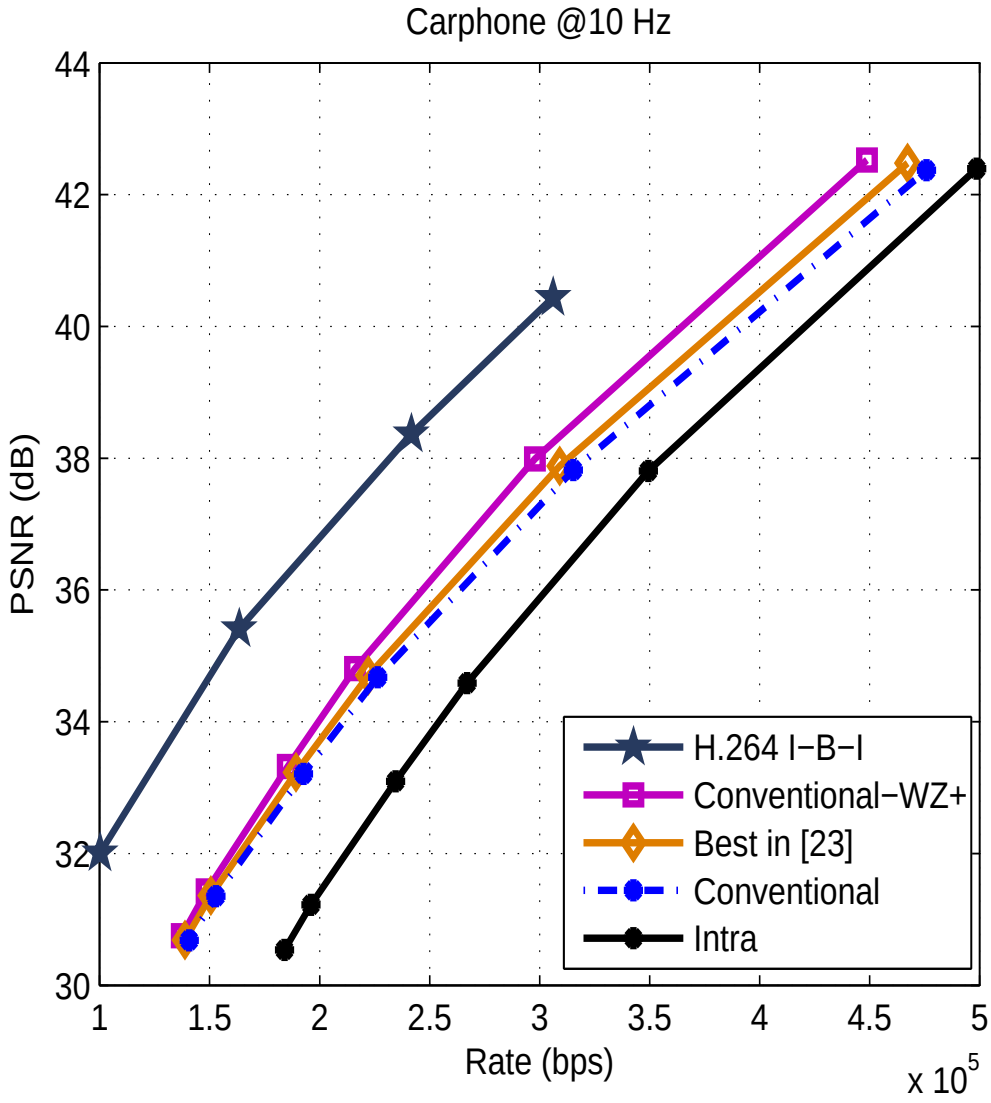


Figure 3.13: PSNR vs. rate for different coding methods for *Carphone* @ 10fps

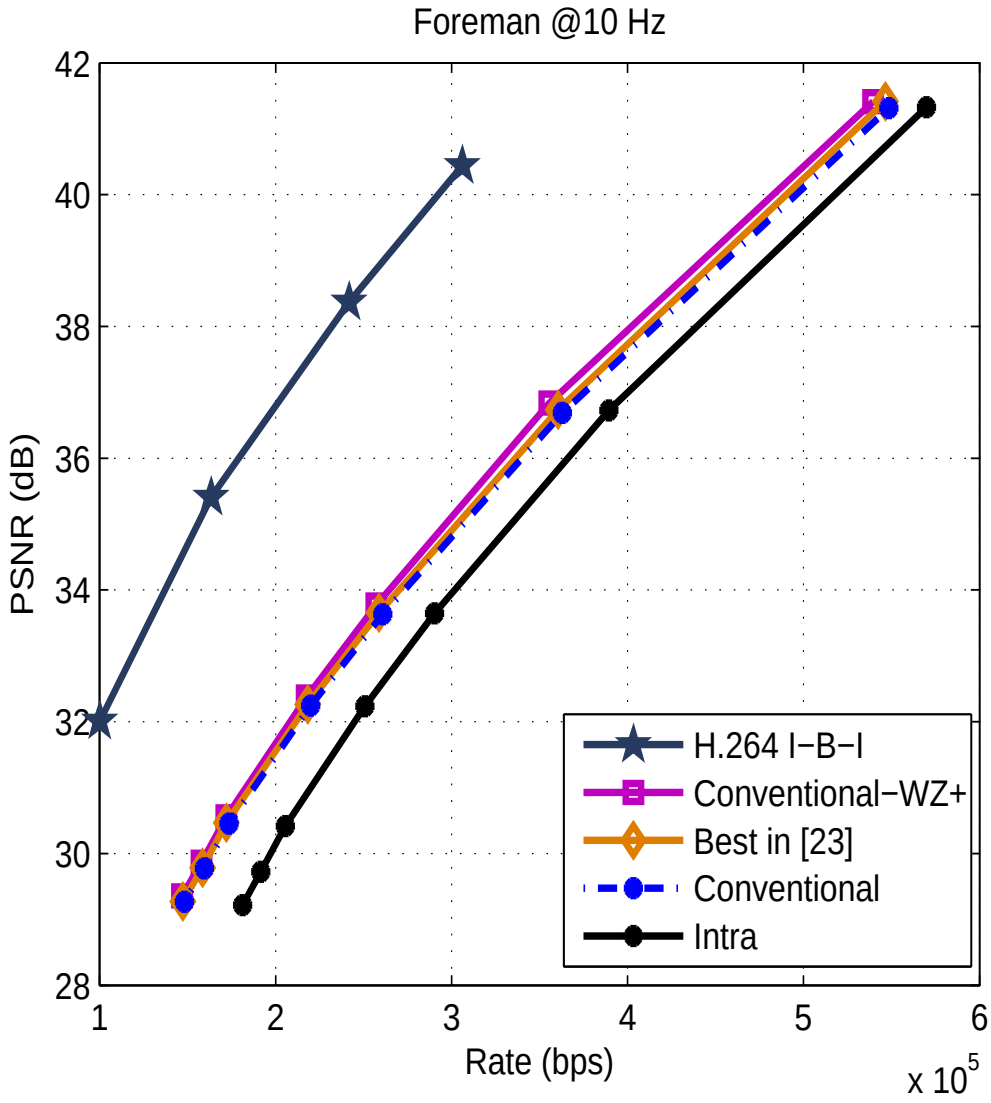


Figure 3.14: PSNR vs. rate for different coding methods for *Foreman* @ 10fps

Chapter 4

Key Frame Encoding and Hierarchical Coding

4.1 Introduction

As mentioned, key frames are usually intra encoded and decoded so the inter-frame correlation between them is not exploited. Extending Wyner-Ziv coding to key frames as well can help to exploit the temporal correlation and improve the rate-distortion performance. In [26], Wyner-Ziv coding was applied for key frames and the previously decoded key frame was considered as the pixel domain side information for the next key frame to be decoded. Their results showed improvement for two low motion sequences. However, as shown in [27] and [28], directly applying Wyner-Ziv coding on key frames can degrade the overall performance since Wyner-Ziv coding is capable of outperforming intra coding only when the side information is accurate enough. Using the previously decoded key frame as the side information for the next key frame to be encoded is usually not accurate enough, especially for high motion sequences.

In this chapter, the Wyner-Ziv coding method is extended to key frames by applying a coding mode selection technique which tries to select the proper coding method (Intra or Wyner-Ziv) based on the correlation characteristics of the low and high frequency bands of each frame to the past. In this method, the

decoder decides the coding mode and no complexity is added to the conventional Wyner-Ziv encoder. After decoding low bands, a new method is used to refine the side information corresponding to the remaining frequency bands. This chapter is organized as follows. In Section 4.2, our proposed key frame encoding is described in detail. In Section 4.3, a hierarchical coding applying both of the proposed methods of noise classification and key frame coding is explained. In Section 4.4, the performance of the proposed methods is evaluated. Section 4.5 concludes the Chapter.

4.2 Key Frame Encoding Based on Frequency Band Classification

In conventional transform domain Wyner-Ziv coding, key frames are encoded and decoded by a conventional intraframe coder. So, the spatial correlation within a block is exploited by applying a DCT, but the temporal correlation between adjacent key frames is not exploited [27]. To extend the Wyner-Ziv coding idea to key frames to exploit similarities between them, previously decoded key frames can be used as the side information. If the side information is not a sufficiently accurate estimate of the source, Wyner-Ziv coding can do worse than intra coding. So, we need tools to evaluate the quality of the side information to select the proper coding method. Wyner-Ziv coding and Intra coding blocks are already part of existing Wyner-Ziv codecs, therefore applying a method switching between Wyner-Ziv and intra coding to exploit inter-frame correlation between consecutive key frames does not add complexity to the encoder as long as the decision step is done at the decoder. Since the temporal correlation of low frequency bands is usually high, Wyner-Ziv coding can often outperform Intra coding. For high frequency bands, measuring the distortion between source and side information of the low frequency bands at the decoder can help to estimate the accuracy of the side information for high frequency bands [28]. Side information which is simply a previous decoded key frame can be refined to a more accurate one for high frequency bands by using decoded low frequency bands.

4.2.1 Intra Coding

For the Intra mode, the quantized DCT coefficients are arranged in a zigzag order to maximize the length of zero runs. The codeword represents the run-length of zeros before a non-zero coefficient and the size of that coefficient. A Huffman code for the pair (Run, Size) is used because there is a strong correlation between the size of a coefficient and the expected run of zeros which precedes it. In our simulation, Huffman and run length coding tables are borrowed from the JPEG standard.

4.2.2 Coding Mode Selection and Side Information Refinement

Fig. 4.1 shows our proposed codec applying coding mode selection for key frames. To separate different frequency bands of the key frame to be encoded, first a DCT is applied. For frequency band k, the k^{th} DCT coefficients from all blocks are grouped to form vector X_k. Low frequency bands are encoded and decoded by Wyner-Ziv coding. The previously decoded key frame is used to generate the side information for low frequency bands. To provide the corresponding side information for each frequency band, a DCT is applied on the previously reconstructed key frame and the k^{th} DCT coefficients from all blocks are grouped to form vector $\hat{X_k}$. Once the decoder receives and decodes all low bands, a block matching algorithm is used for motion estimation of each block with reference to the previously decoded key frame. In block matching algorithms, each macroblock in the new frame is compared with shifted regions of the same size from the previous frame, and the shift which results in the minimum error is selected as the best motion vector for that macroblock. Since here only reconstructed low bands of the new key frame are available at the decoder, the best match is found using the mean squared error (MSE) of low frequency components. The MSE of low bands of two blocks A and B with $n \times n$ pixels is calculated as

$$MSE = \frac{1}{K_{low}} \sum_{(i,j)\epsilon Low freq.} (U(i,j) - V(i,j))^2 \tag{4.1}$$

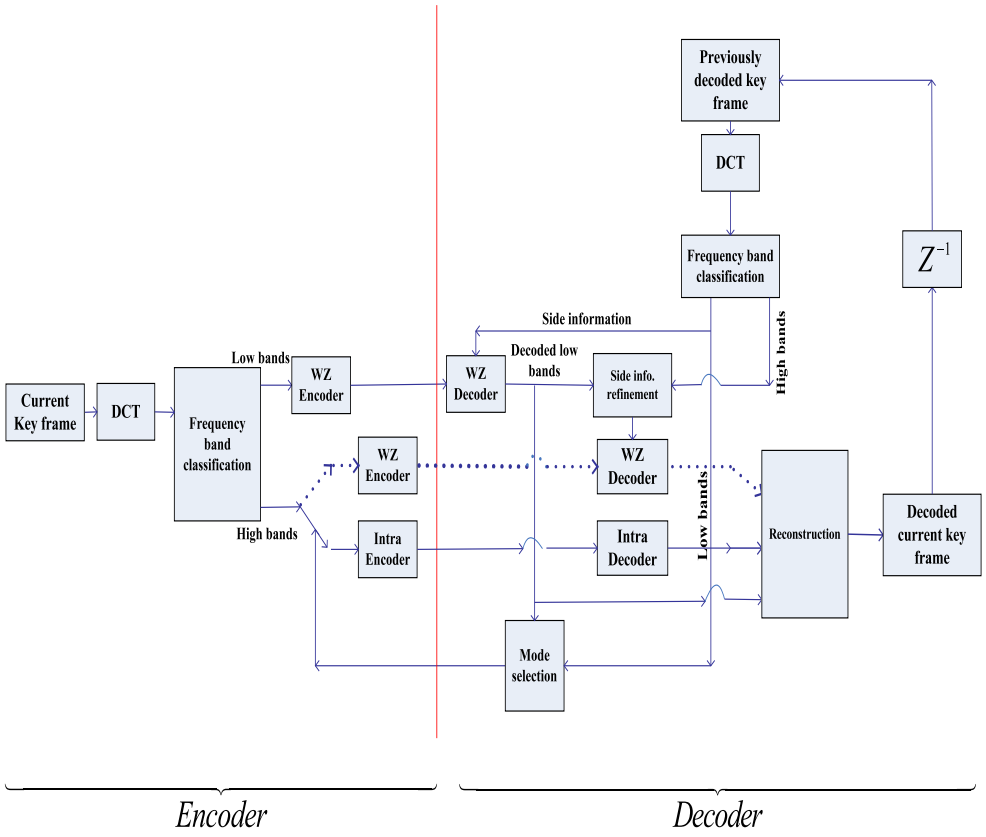


Figure 4.1: Proposed video codec with frequency band coding mode selection for key frames

where K_{low} is the total number of low bands and U and V are the DCT transform of A and B, respectively. The motion compensated frame is the new side information for the remaining frequency bands. In our simulation, motion estimation for the refinement step is a full search in a ± 2 pixel search area. To select the proper coding method for high frequency bands we need to estimate the accuracy of the side information. At this point, decoded low bands constitute the only available information of the frame to be encoded. Since the side information is a noisy version of the source, measuring the distortion between decoded low bands of the current key frame and those of the motion compensated one at the decoder can help to give an estimation of the distortion for high bands. This distortion is calculated as

$$D = \frac{1}{K_{low} \times L} \sum_{k \epsilon Low freq.} \sum_{l=1}^{L} (\acute{X}_k(l) - \tilde{X}_k(l))^2 \tag{4.2}$$

where $\acute{X}_k$ denotes the reconstructed X_k at the decoder and $\tilde{X}_k$ denotes a vector formed by grouping the k^{th} DCT coefficient from all blocks of the motion compensated frame at the decoder. L is the number of elements in each frequency band which is the number of DCT blocks in a frame. If D is less than a threshold T_D, the side information is likely accurate enough that Wyner-Ziv coding can outperform Intra coding for high frequency bands. Otherwise, Intra coding is applied for them.

The decoder sends a single bit per frame through the feedback channel to indicate the selection. The added effect of sending a single bit per frame through the feedback channel on the latency of the system is negligible, since in traditional Wyner-Ziv coding, feedback bits might be sent for each bit plane to request more accumulated syndrome bits to meet the desired bit error rate. The conventional Distributed Video Coding (DVC) decoder allows for all the bands to be decoded in parallel whereas the proposed scheme essentially cuts in half the amount of parallelization that could be done. So instead of having a time S in which to decode (in parallel) all the bands, the decoder would have to decode the low bands in $\frac{S}{2}$ and then the high bands in $\frac{S}{2}$. To allow random access and limit error propagation, we can switch off our proposed key frame encoding once in a while to use intra coding instead, as is done in conventional IPPP or IBBP type coders. The whole process of Wyner-Ziv coding of low bands, side information refinement, and finding the proper coding method for high bands is called adaptive coding for the rest of this chapter. As more bands are considered to be low, more accuracy is expected for the side refinement step in this method, although there would be some exceptions based on video content. But if we increase the number of low bands, fewer bands would be left to take advantage of the improved side information. As depicted in Fig.4.2, the performance is improved when $f(1,1)$, $f(1,2)$ and $f(2,1)$ are considered as low bands compared with the case that only $f(1,1)$ is considered. However the

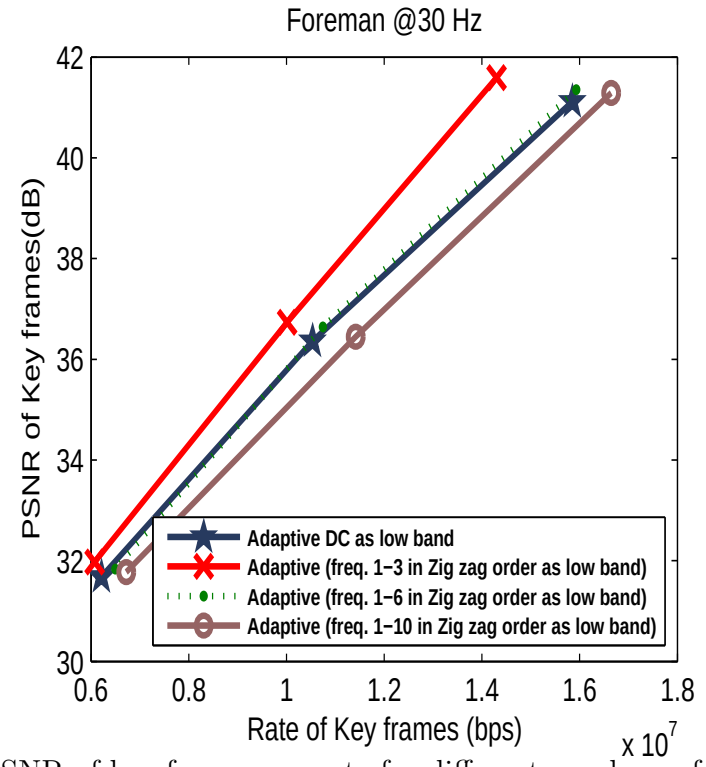


Figure 4.2: PSNR of key frames vs. rate for different numbers of frequency bands considered as low bands

performance is degraded by considering the 6 lowest frequencies of the 4×4 DCT in zig zag order as low bands. Therefore, in our simulation $f(1,1)$, $f(1,2)$ and $f(2,1)$ are considered low frequency bands, and the rest are considered high frequency bands.

4.3 Hierarchical Coding

In traditional Wyner-Ziv coding, key frames occur every other frame and are intra coded to provide high quality side information for the Wyner-Ziv frames in between. Many key frames encoded as intra leads to increasing rate, and to overall rate distortion degradation. MCFI methods tend to be less successful when the distance between frames gets higher, so less frequent key frames results in less accurate side information for the corresponding Wyner-Ziv frame. Less accurate side information means more accumulated syndrome bits need to be sent to satisfy the bit error expectation. In the previous section, we proposed a method to exploit

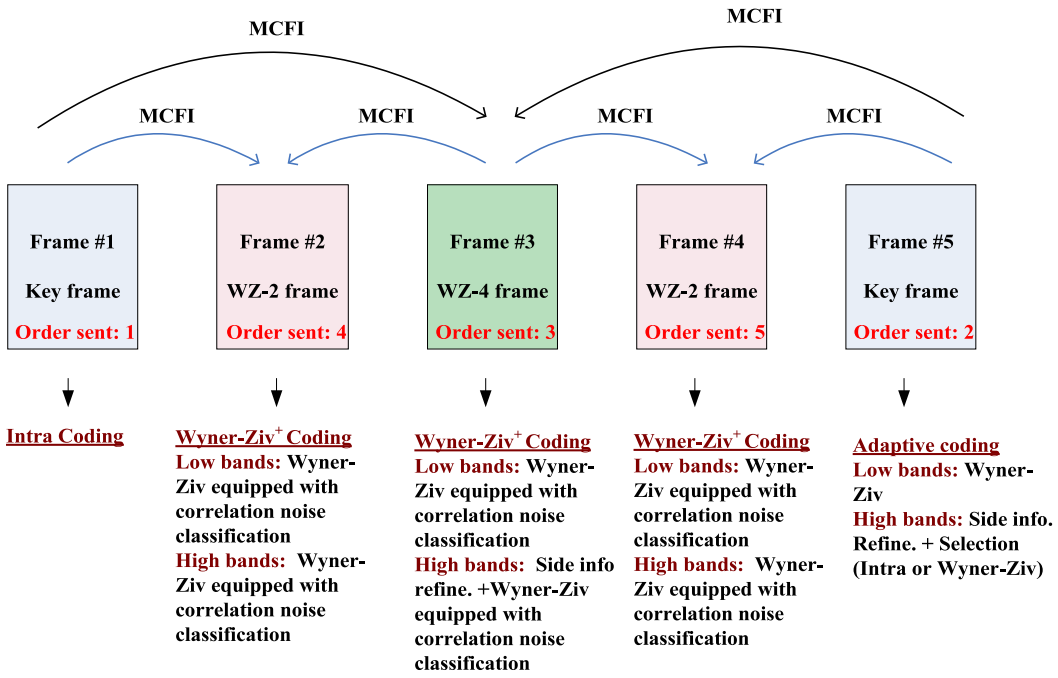


Figure 4.3: Proposed hierarchical coding

similarities between key frames. Here, we propose a more practical structure taking advantage of both adaptive coding and correlation noise classification techniques. Most WZ coders consider key frames every 2 frames. We started with this spacing and saw what improvement could be obtained by key frame prediction. The next step beyond this is key frame spacing of 4. As shown in Fig. 4.3, in this hierarchical arrangement, key frames occur every 4 frames and there are two types of Wyner-Ziv frames: Wyner-Ziv frames with 4 frame distance, WZ-4, and Wyner-Ziv frames with 2 frame distance, WZ-2, which will be explained in detail. Lookup tables of correlation noise classification for the two types are different, and are obtained offline by using several sequences as training data.

Compared to the traditional structure with one frame delay, latency in this structure is increased to a delay of 3 frames. In traditional Wyner-Ziv video coding where key frames occur every other frame, decoding of a Wyner-Ziv frame can not be started unless the previous and next key frames were decoded.

4.3.1 Key frames

As depicted in Fig. 4.3, key frames occur every 4 frames and they are used to generate side information corresponding to WZ-4 frames which will be explained later. The first key frame is intra coded since no other information is available. Applying the proposed adaptive coding method in Section 4.2 will be very helpful to exploit temporal correlation of key frames in high frame rate videos or low motion sequences. Otherwise simply applying intra coding would be a better choice. In Fig. 4.4- 4.16, both methods are applied for key frames, and results for different types of video content and frame rates are compared.

4.3.2 WZ-4 frames

As shown in Fig. 4.3, WZ-4 frames are at 2 frame distance from key frames and 4 frame distance from each other. The MCFI method proposed in [25] is applied on previous and next key frames to generate their corresponding side information. Since here the side information comes from both temporal directions and MCFI is applied, we can apply the proposed correlation noise classification method in Chapter 3. For a given block of a WZ-4 frame, the decoder evaluates the matching success of MCFI by calculating the residual energy between forward and backward interpolation and chooses one of the defined m classes by comparing to the threshold values. Once the block class is determined, the α parameter of each frequency band is found through the lookup table. Once low bands are reconstructed at the decoder, they are used to refine the side information, and the rest of the frequency bands are Wyner-Ziv encoded with the refined side information.

4.3.3 WZ-2 frames

As depicted in Fig. 4.3, these frames lie between key frames and WZ-4 frames. The MCFI method proposed in [25] is applied on their key frame and WZ-4 frame immediate neighbors which are at 1 frame distance from them. For this type of frame also, side information comes from both sides, so the correlation noise classification technique is applicable. Since here the frame distance is only

1 frame from each side, the obtained side information is more accurate than for WZ-4 frames. Empirically, for WZ-2 frames, having low bands is not very helpful to provide more accurate side information than the one attained by MCFI. So, the refinement step is not applied for them.

4.4 Simulation Results

Fig. 4.4-4.7, Fig. 4.8-4.11 and Fig. 4.13-4.16 show the rate-distortion performance for the test sequences *Claire*, *Mother-daughter*, *Foreman* and *Carphon* QCIF (176×144) sequences at 30 fps, 15fps and 10fps. Fig. 4.12 (e) shows the rate-distortion performance for the *Soccer* QCIF sequence at 15 fps. The same settings as in Section 3.4 are applied. For adaptive coding which is described in Section 4.2.2, for each one of these quantization parameters, a threshold value is set. We tried different values between 50 and 1800 with step sizes 20 to 100 for several video sequences at different quantization parameters. The value of the step size depends on the quantization parameter, with larger step sizes for larger quantization parameters. Threshold values $[T_1, T_2, ..., T_7] = [200, 290, 740, 860, 1200, 1500, 1800]$ corresponding to quantization parameters $QP = [0.4, 0.85, 1.5, 2, 3, 3.5, 4]$, were chosen as they work well for the training sequences with different characteristics. Threshold values are obtained for training sequences at 30 fps and used for test sequences at frame rates of 30, 15 and 10 fps. For correlation noise classification, for each type of Wyner-Ziv frame and each quantization step, a different lookup table is calculated.

Table 4.1 shows the average number of times that key frame high bands are Wyner-Ziv coded in the Adaptive coding method. In Fig. 4.4- 4.16, the results of applying different methods are compared. With "*Intra*", all frames are intra encoded and decoded by using the method explained in Section 4.2.1. The complexity of this method is as low as JPEG. In this work, whenever intra coding was needed, this method was used. "*Conventional*" is based on the method in [10] but we modified the algorithm in two ways. First, the assumption of availability of original key frames at the decoder is removed since it is not valid from a practical point of

view. Second, the quantization part is replaced with the quantization procedure explained in Section 3.4. Although not depicted in the figures, our simulation results show that this change in quantization method improves the performance of [10]. Our quantization method is applied for all proposed methods. We use the same quantization method for all the approaches in order to highlight the performance improvement due to correlation noise classification and key frame encoding. In the "*Conventional*" method, key frames (odd frames) are encoded and decoded as intra using the method explained in Section 4.2.1, and even frames are encoded as Wyner-Ziv frames. When Wyner-Ziv coding equipped with correlation noise classification is applied for Wyner-Ziv frames of the conventional method, the method is called "$Conventional - WZ+$". When the adaptive method is applied for key frames (odd frames), and the Wyner-Ziv method equipped with correlation noise classification is applied for even frames, the method is called "$AdaptiveWZ+$". "$Hierarchical - key - intra$" and "$Hierarchical - key - adaptive$" are the names of the methods explained in Section 4.3 where intra coding and adaptive coding are applied respectively for key frames. Results are also compared to "H.264 intra" and "H.264 I-B-I'. Since in this work all methods are using an intra method as low complexity as JPEG, to have a fair comparison, intra predictions and context adaptive binary arithmetic coding (CABAC) are turned off for I frames of "H.264 intra" and "H.264 I-B-I". Certainly, adding these features can improve the performance of all methods, at the cost of additional complexity.

The proposed adaptive method combined with correlation noise classification results in up to 5 dB improvement over "Conventional-WZ+" (*Claire* 30 fps at 400 Kbps). The gain is more for low motion and higher frame rate sequences where the inter-correlation is high. For high motion sequences at lower frame rate we do not expect improvement since the inter-correlation is very low. As shown in Fig. 4.11 and 4.16 for *Foreman*, as a moderately high motion sequence at 15 fps and 10 fps, the performance of "Adaptive-WZ+" is very close to that of "Conventional-WZ+" but with a slight degradation. For very high motion sequences such as *Soccer* at 15 fps where the MCFI method gives a poor side information, the whole idea of Wyner-Ziv coding fails, meaning that intra coding outperforms Wyner-Ziv

coding. For such cases, all of these methods for exploiting correlation between consecutive key frames are useless. "$Hierarchical-key-adaptive$" is capable of beating all methods for most cases and results in up to 1 dB additional improvement. The exceptions are *Foreman* at 15 fps, 10 fps and *Soccer* at 15 fps. For these high motion and low frame rate cases, since in the hierarchical structure, key frames are 4 frames apart, the temporal correlation between key frames is very low. So applying intra coding for key frames would be a better alternative. As shown in Fig. 4.8- 4.16, "$Hierarchical-key-intra$" can beat "$Hierarchical-key-adaptive$" for these cases. Although even "$Hierarchical-key-intra$" results in degradation for *Soccer* as the whole idea of Wyner-Ziv coding fails for this sequence.

Table 4.1: Average fraction of time key frame high bands are Wyner-Ziv coded.

(a)

	Mother-Daughter		
	@10Hz	@15Hz	@30Hz
QP=0.40	0.84	0.93	0.97
QP=0.85	0.90	0.93	0.97
QP=1.50	0.94	0.97	0.99
QP=2.00	0.98	0.99	0.99
QP=3.00	0.98	0.99	0.99
QP=3.50	0.98	0.99	0.99
QP=4.00	0.98	0.99	0.99

(b)

	Claire		
	@10Hz	@15Hz	@30Hz
QP=0.40	0.92	0.99	1.00
QP=0.85	0.99	0.99	1.00
QP=1.50	0.99	0.99	1.00
QP=2.00	0.99	0.99	1.00
QP=3.00	0.99	0.99	1.00
QP=3.50	0.99	0.99	1.00
QP=4.00	0.99	0.99	1.00

(c)

	Carphone		
	@10Hz	@15Hz	@30Hz
QP=0.40	0.31	0.42	0.64
QP=0.85	0.39	0.53	0.74
QP=1.50	0.55	0.70	0.86
QP=2.00	0.70	0.80	0.92
QP=3.00	0.77	0.90	0.97
QP=3.50	0.81	0.96	0.98
QP=4.00	0.89	0.98	0.98

(d)

	Foreman		
	@10Hz	@15Hz	@30Hz
QP=0.40	0.22	0.29	0.60
QP=0.85	0.22	0.39	0.64
QP=1.50	0.28	0.48	0.71
QP=2.00	0.40	0.57	0.75
QP=3.00	0.58	0.68	0.81
QP=3.50	0.62	0.71	0.83
QP=4.00	0.64	0.76	0.87

4.5 Conclusion

We proposed two new techniques to improve the overall rate distortion performance of Wyner-Ziv video coding: (1) an advanced mode selection scheme for frequency bands of key frames followed by side information refinement, (2) a hierarchical Wyner-Ziv coding approach including the other two schemes. Simulation results showed that with the possible cost of additional buffering at the encoder, the proposed key frame encoding with side information refinement combined with correlation noise classification results in up to 5 dB improvement over the Conventional method equipped with correlation noise classification. Experimental results showed that one can achieve up to 1 dB additional improvement by applying the hierarchical method at the cost of extra latency. All the proposed methods keep the encoder low complexity.

This chapter is in part a reprint of the material in the paper: G. Esmaili and P. Cosman, "Wyner-Ziv Video Coding with Classified Correlation Noise Estimation and Key Frame Coding Mode Selection", *IEEE Transactions on Image Processing*, 2011.

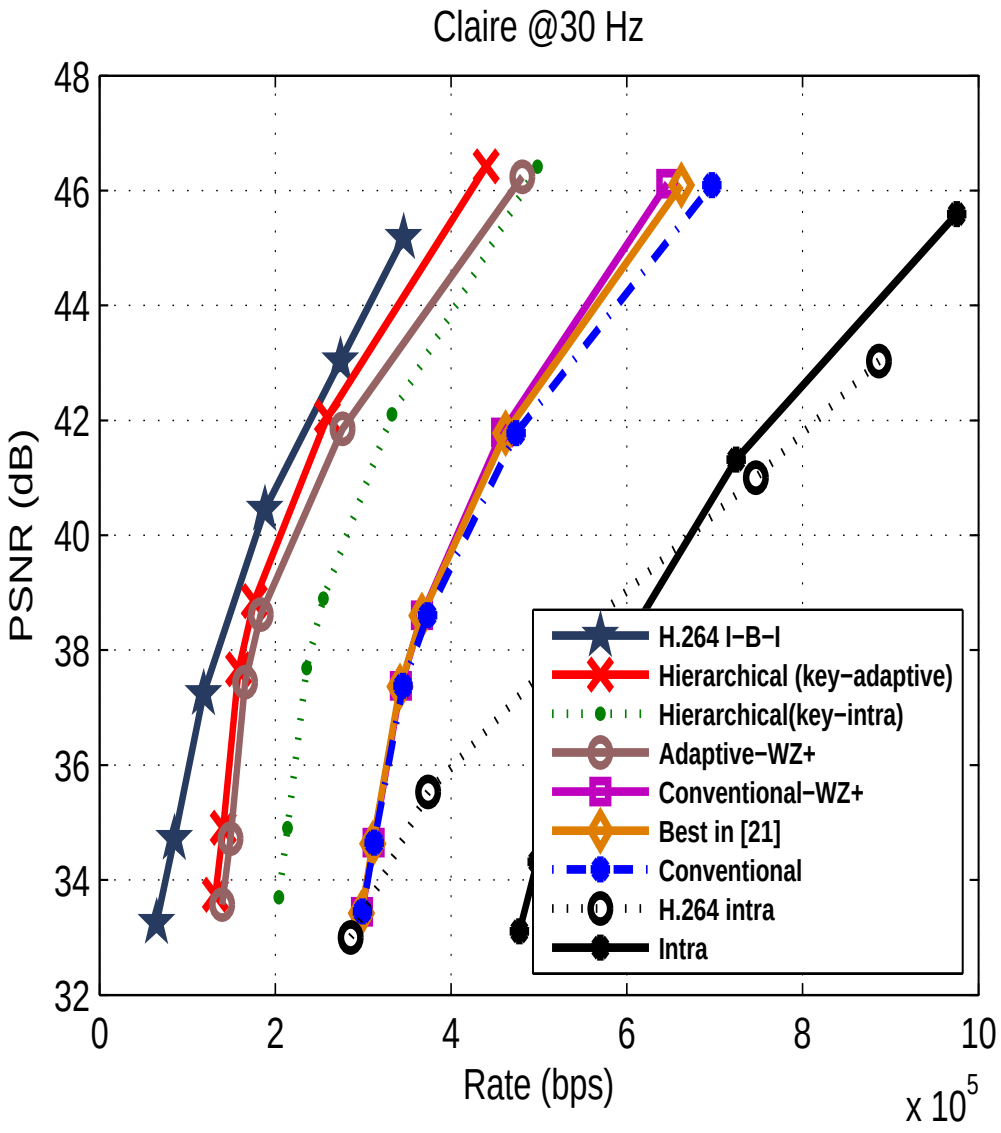


Figure 4.4: PSNR vs. rate for different coding methods for *Claire* @ 30fps

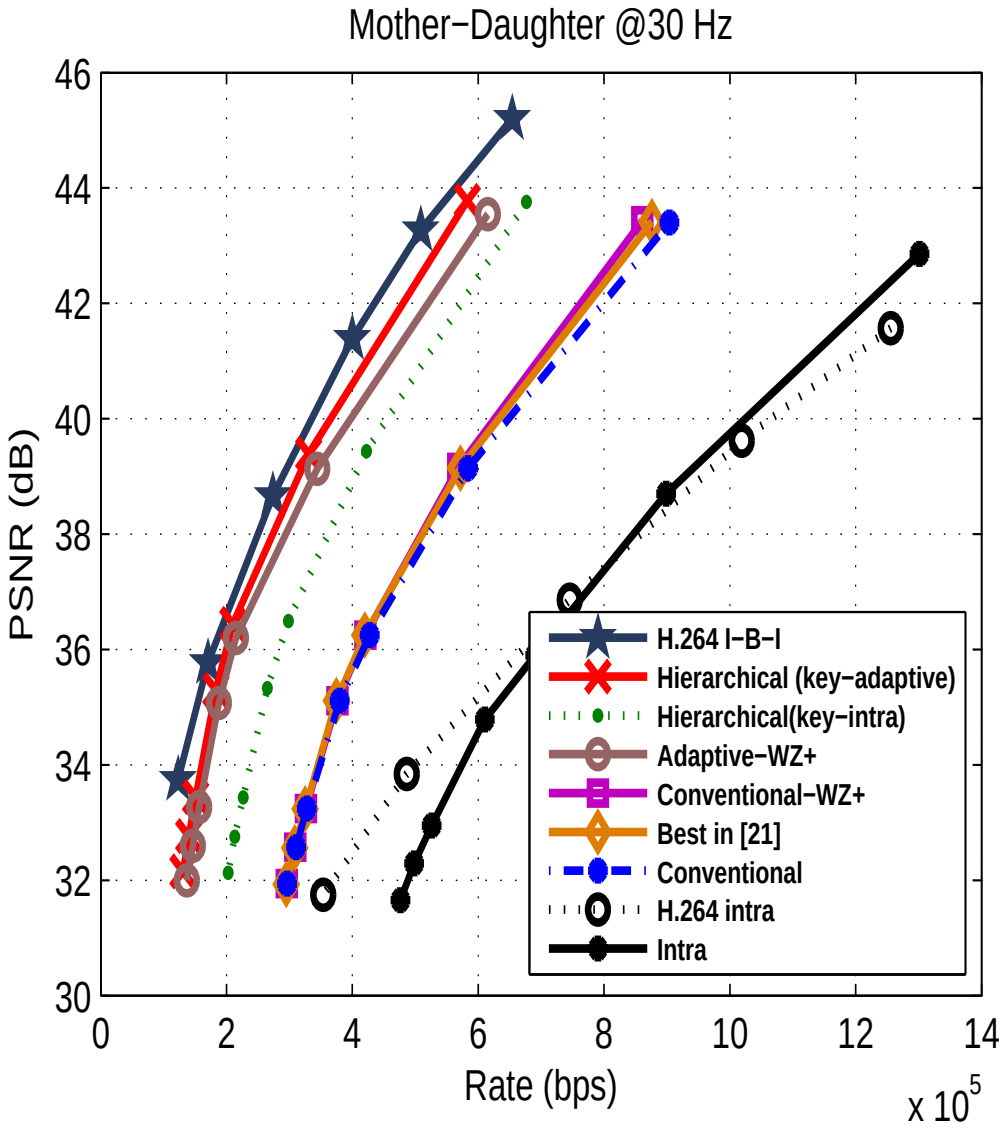


Figure 4.5: PSNR vs. rate for different coding methods for *Mother-daughter* @ 30fps

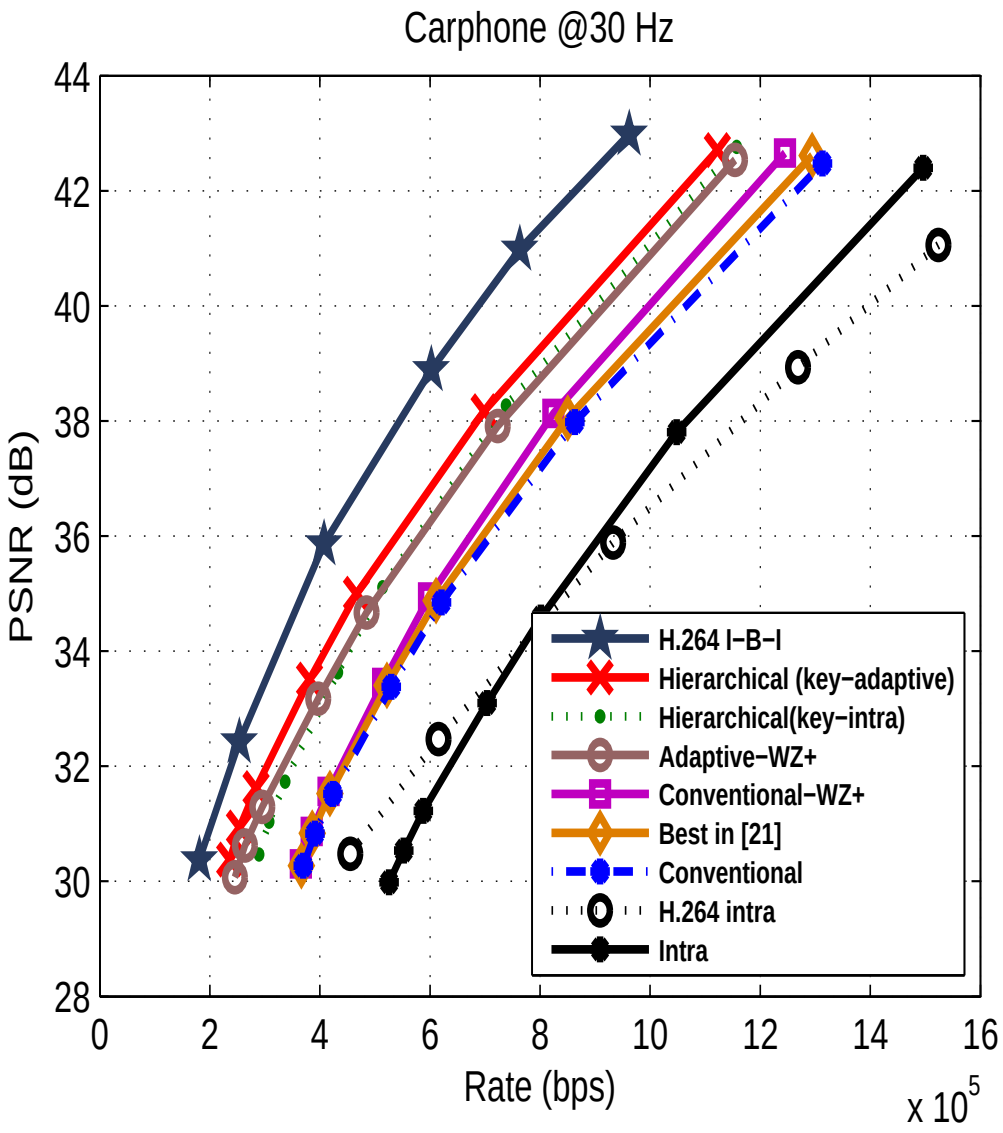


Figure 4.6: PSNR vs. rate for different coding methods for *Carphone* @ 30fps

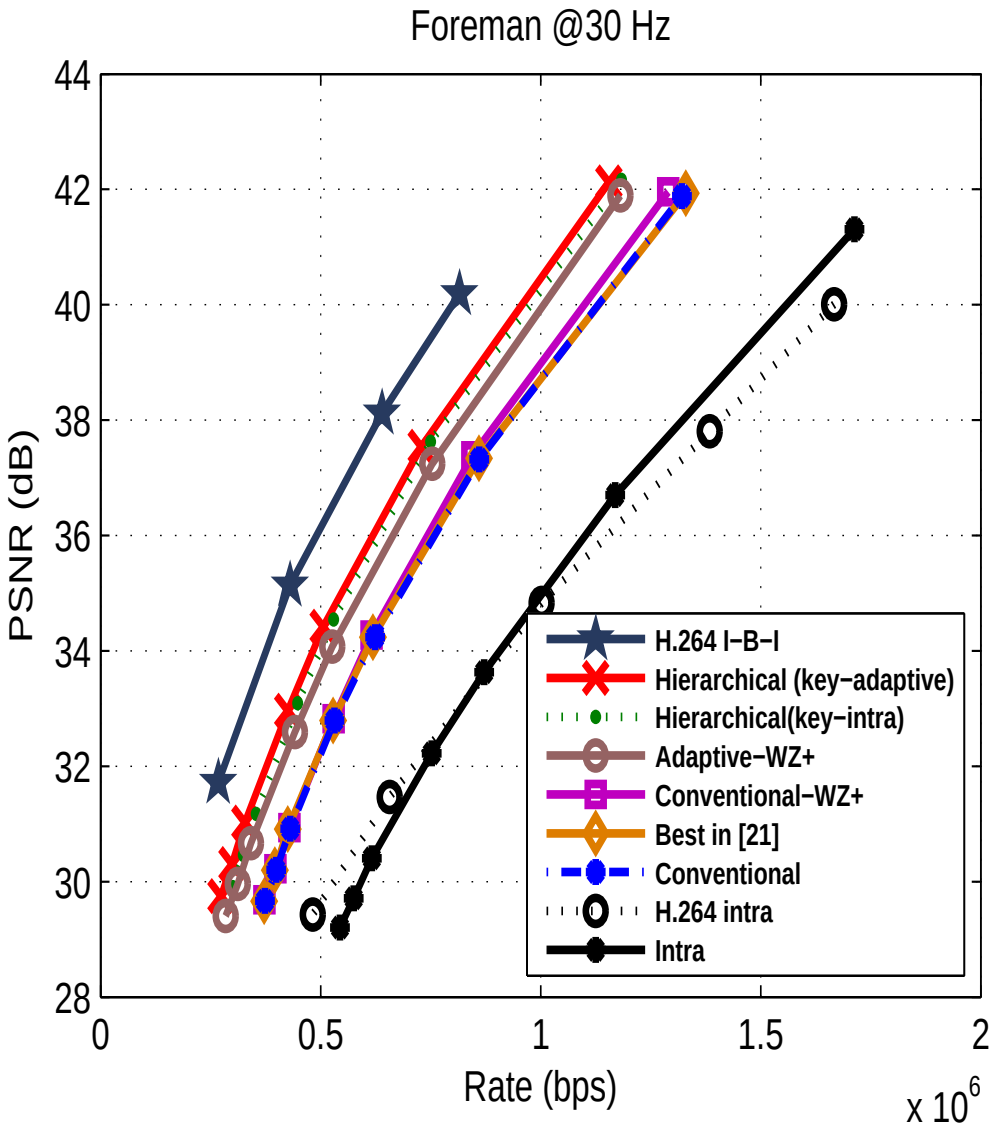


Figure 4.7: PSNR vs. rate for different coding methods for *Foreman* @ 30fps

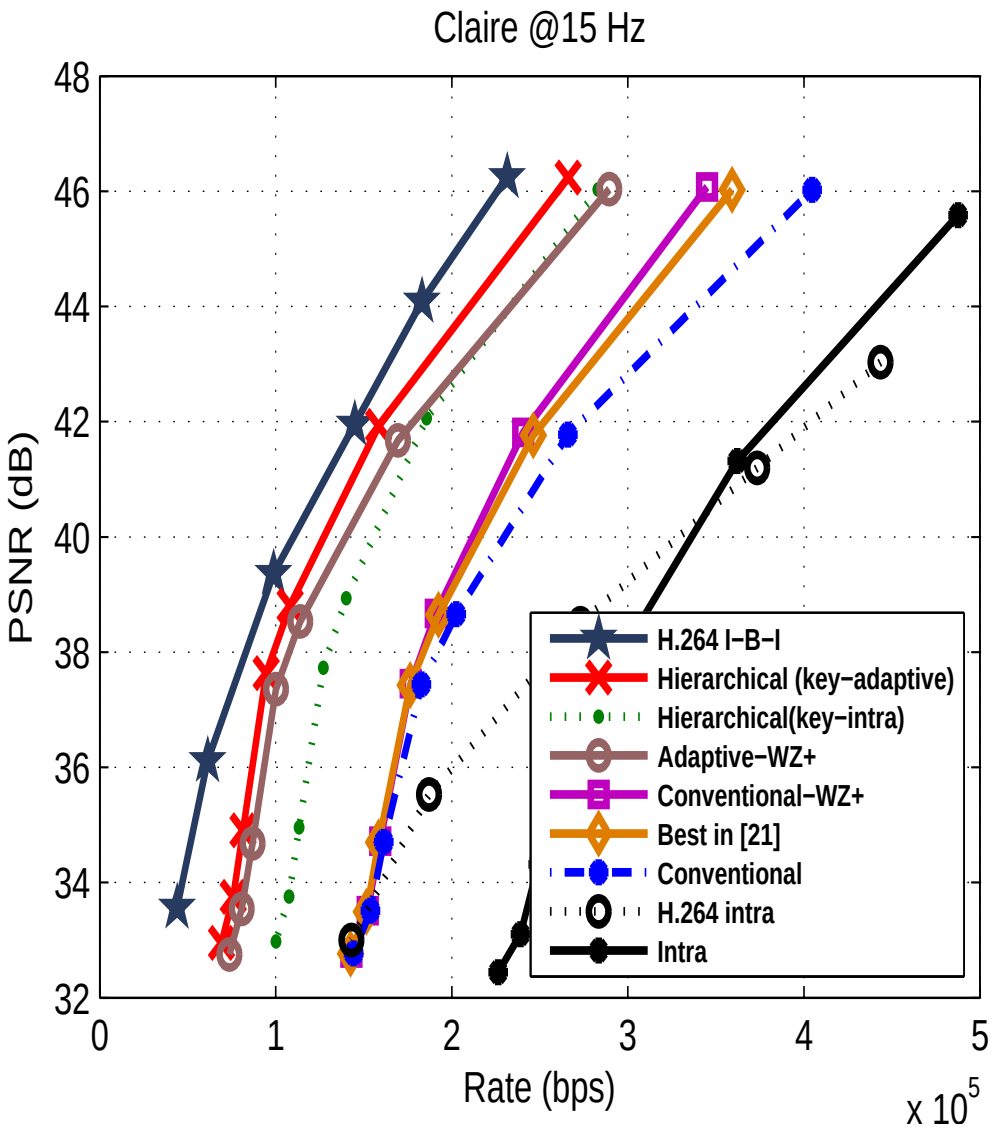


Figure 4.8: PSNR vs. rate for different coding methods for *Claire* @ 15fps

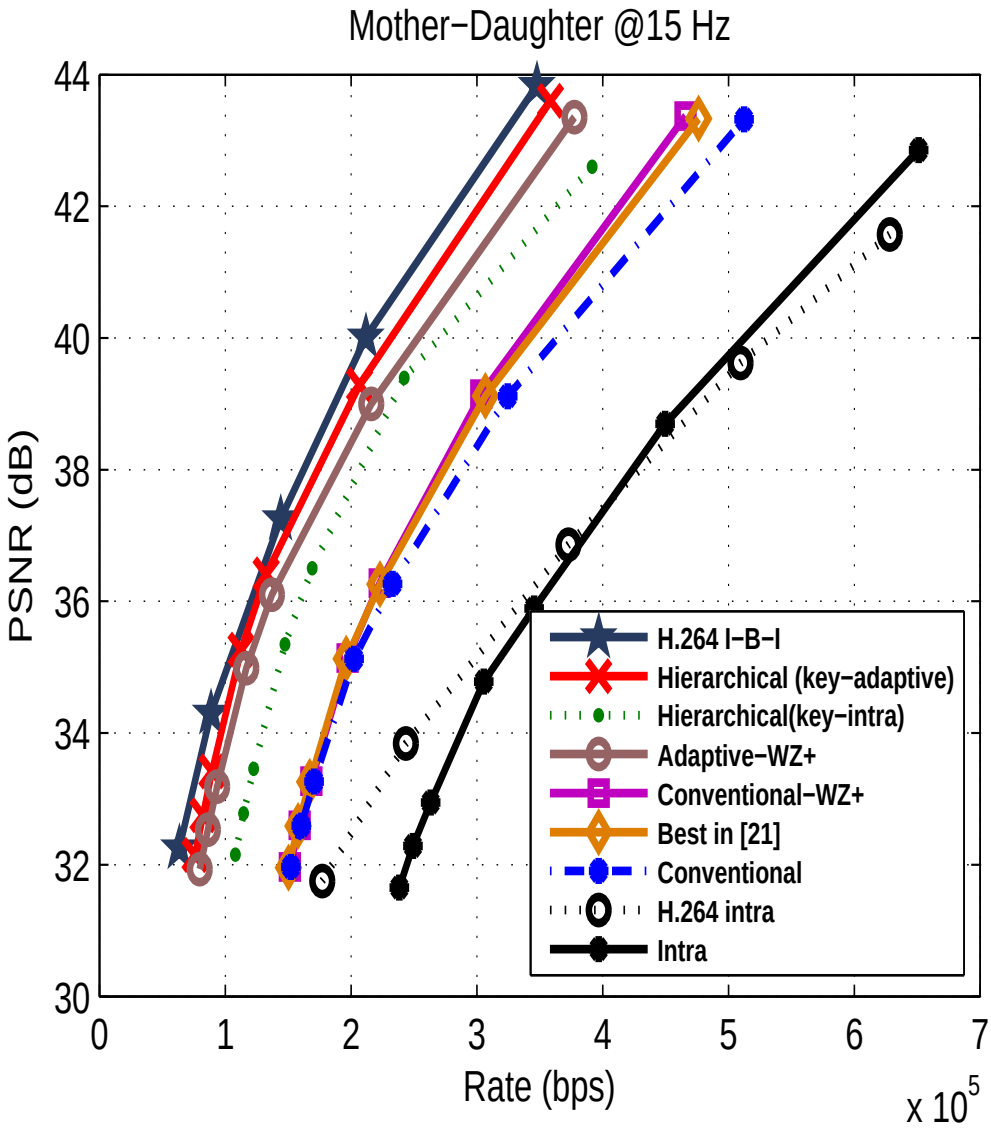


Figure 4.9: PSNR vs. rate for different coding methods for *Mother-daughter* @ 15fps

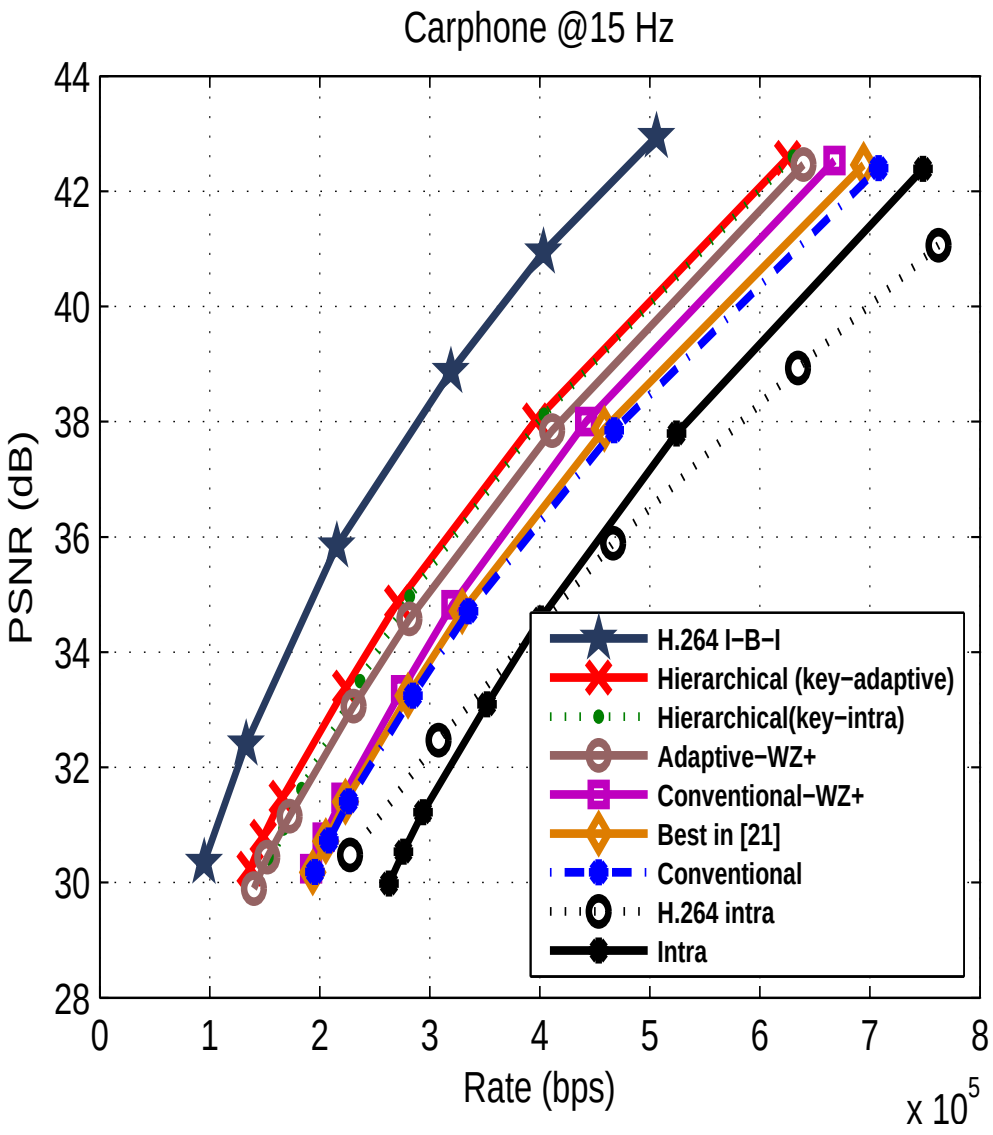


Figure 4.10: PSNR vs. rate for different coding methods for *Carphone* @ 15fps

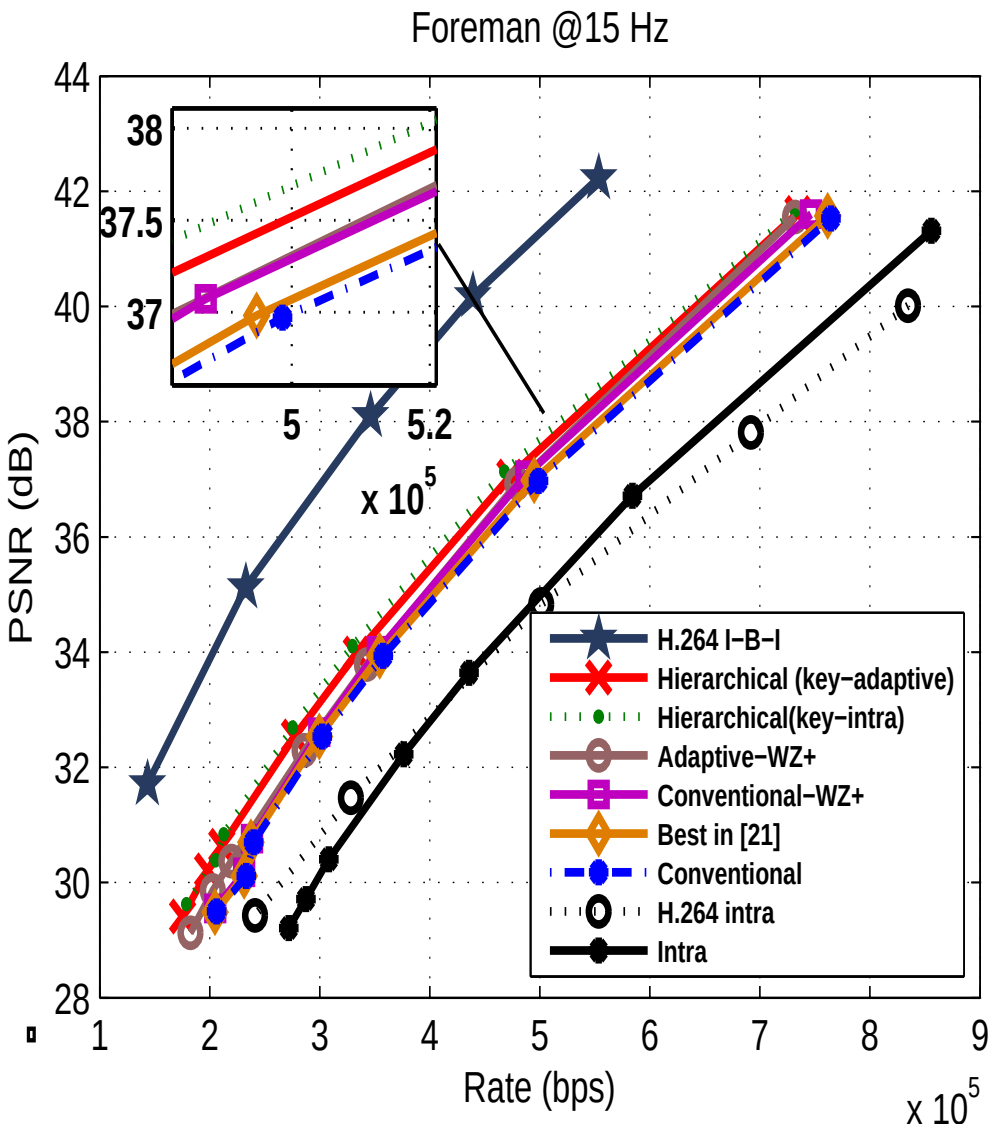


Figure 4.11: PSNR vs. rate for different coding methods for *Foreman* @ 15fps

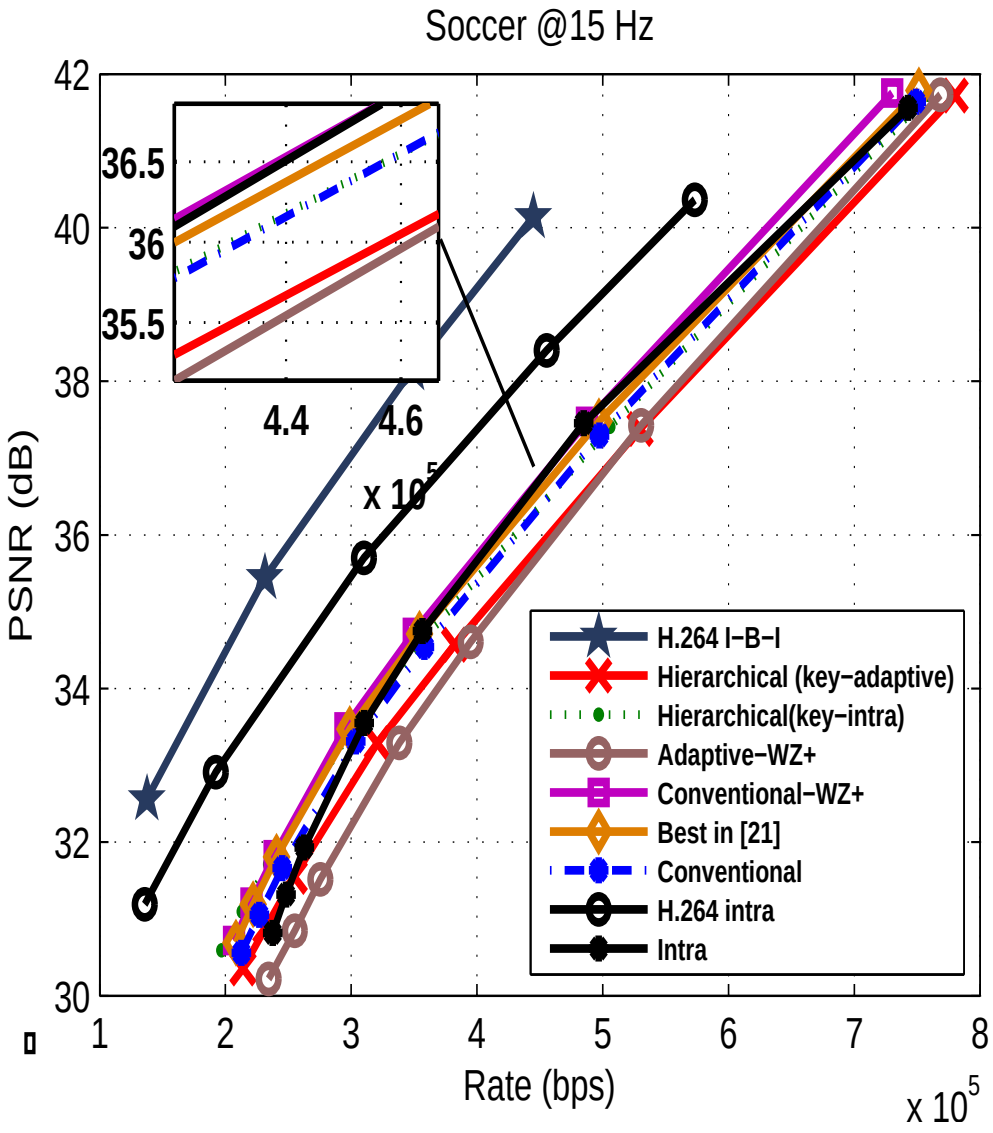


Figure 4.12: PSNR vs. rate for different coding methods for *Soccer* @ 15fps

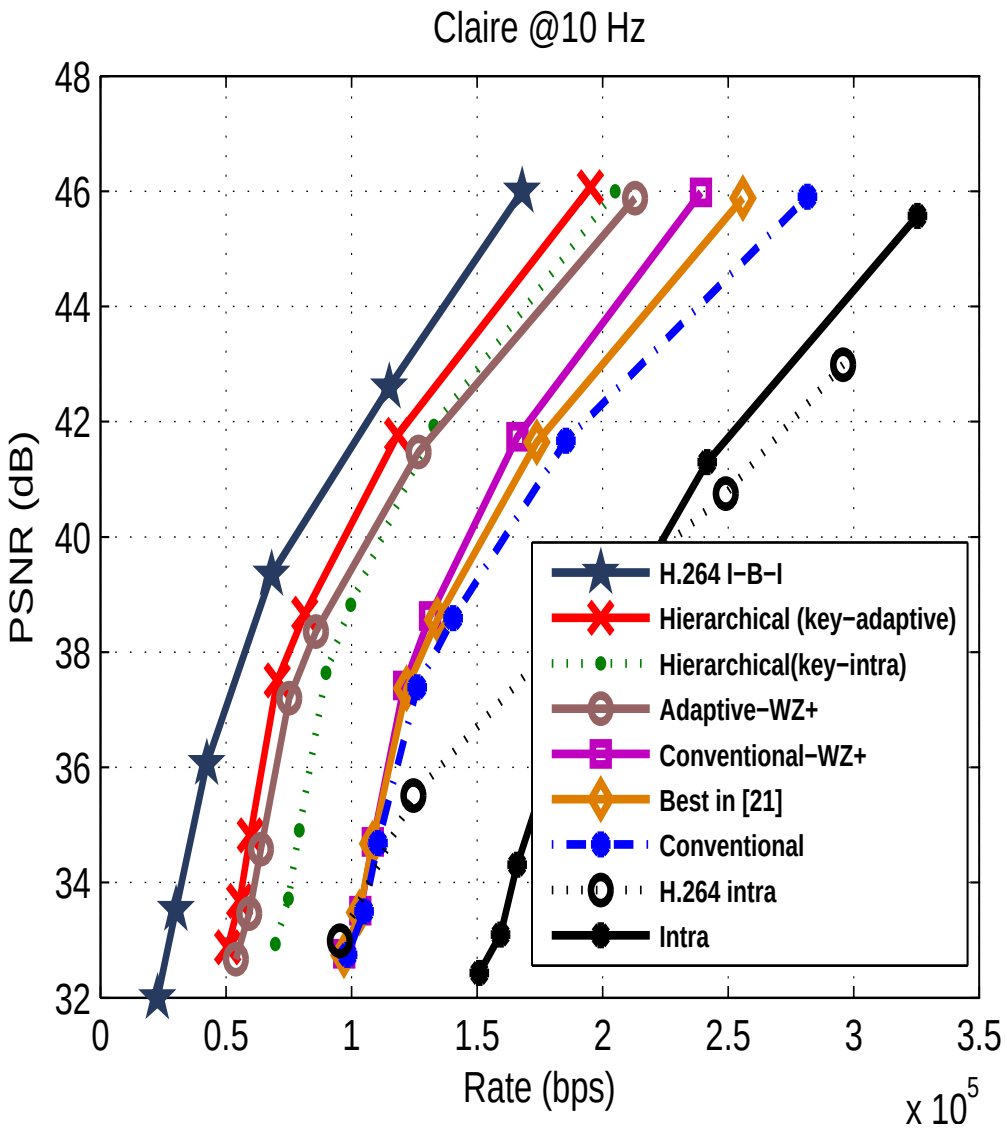


Figure 4.13: PSNR vs. rate for different coding methods for *Claire* @ 10fps

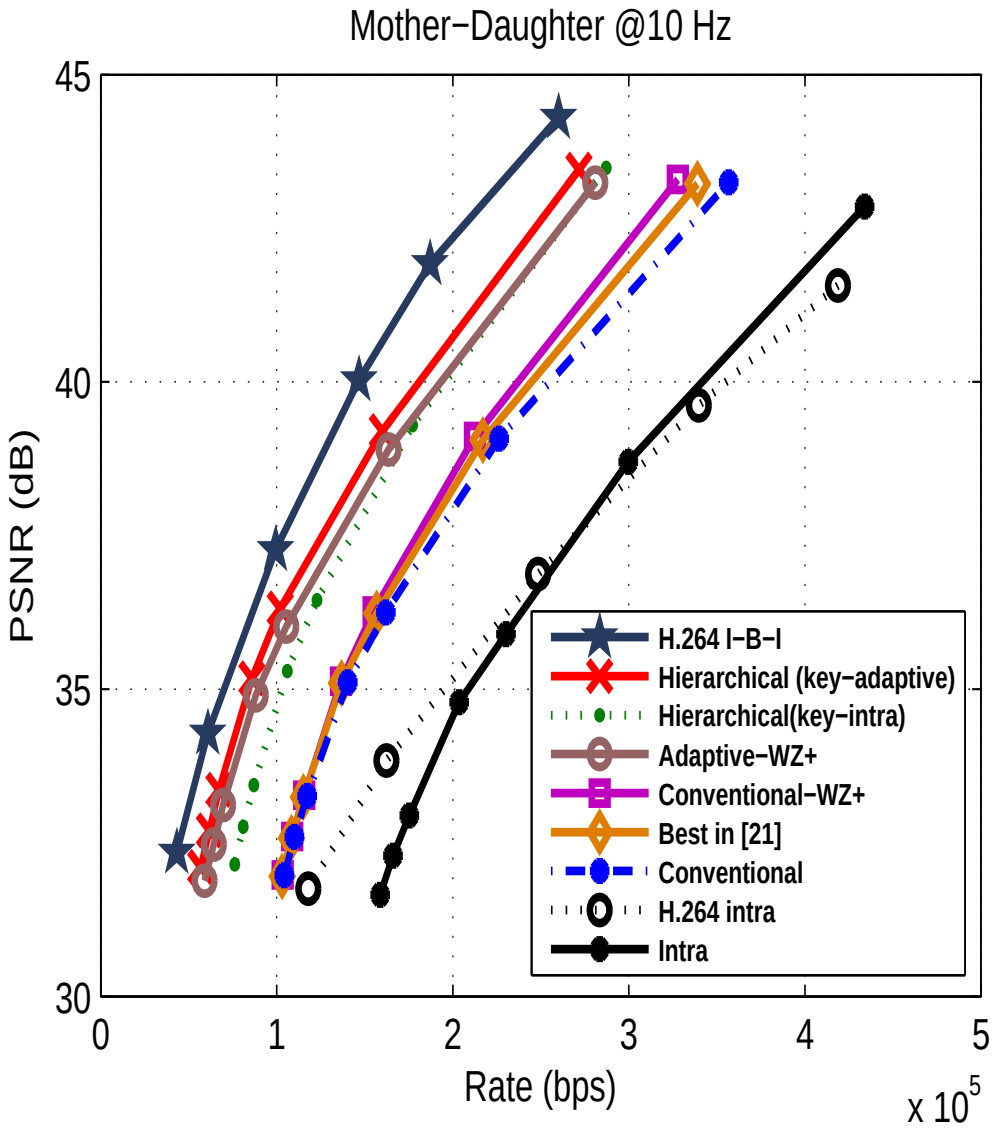


Figure 4.14: PSNR vs. rate for different coding methods for *Mother-daughter* @ 10fps

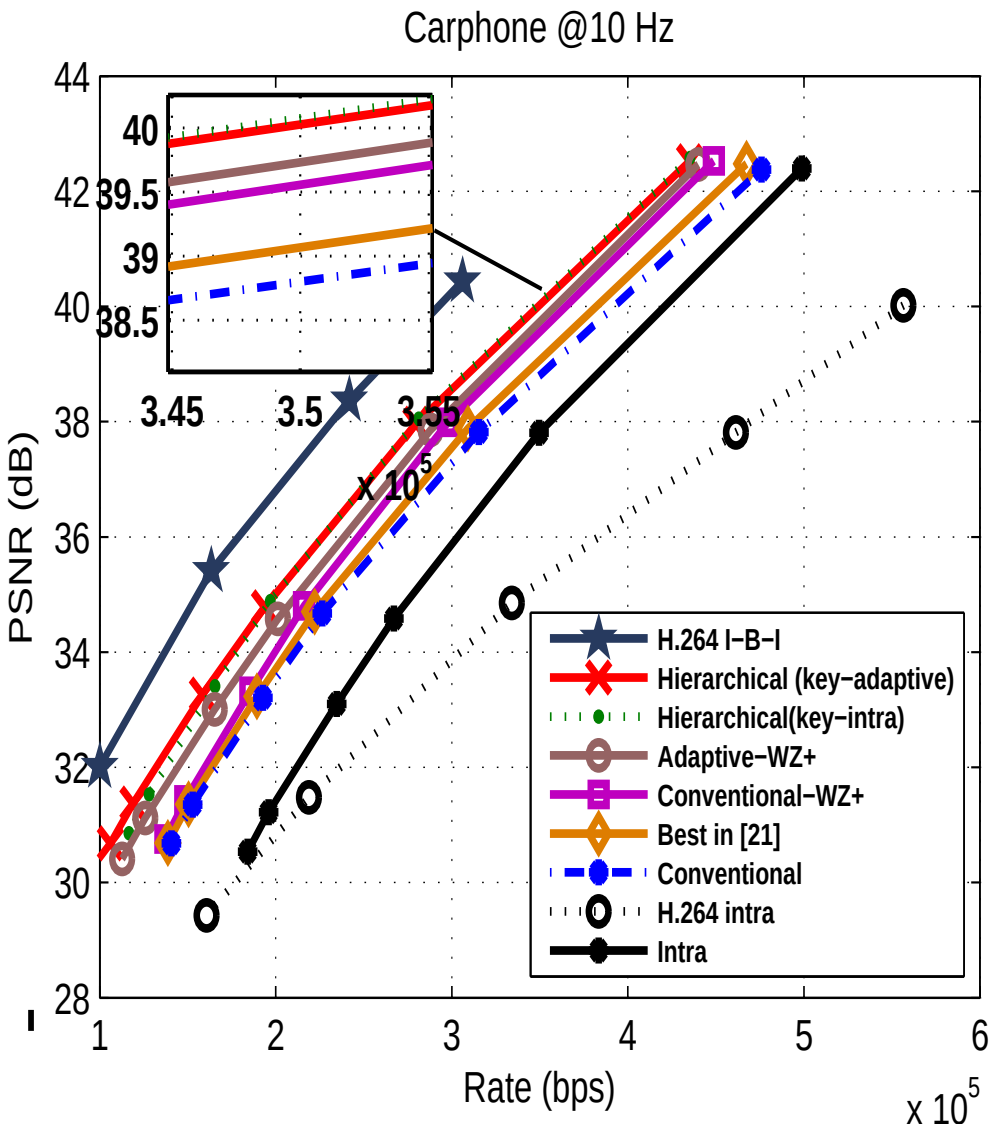


Figure 4.15: PSNR vs. rate for different coding methods for *Carphone* @ 10fps

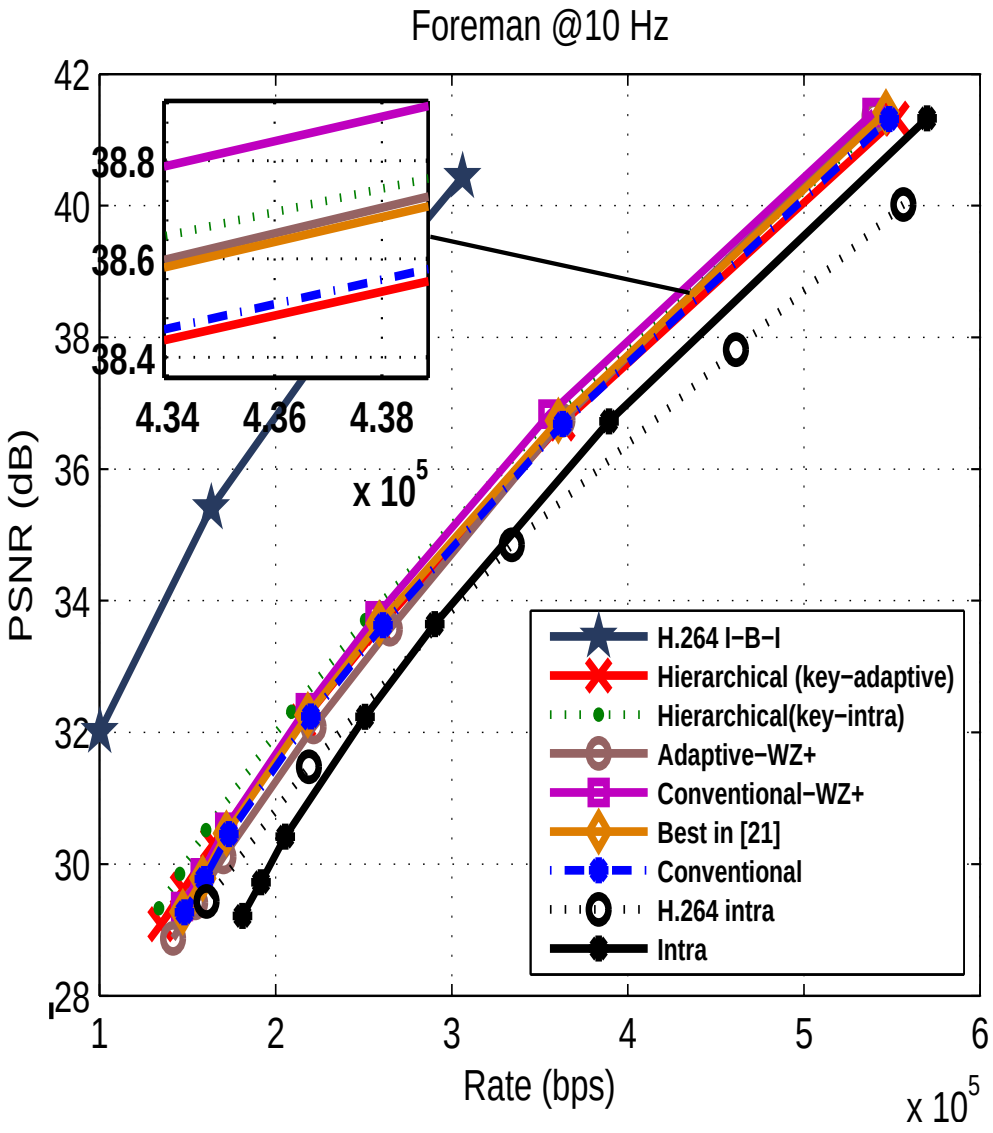


Figure 4.16: PSNR vs. rate for different coding methods for *Foreman* @ 10fps

Chapter 5

Adaptive Rate Control

5.1 Introduction

In most existing WZ video codecs, without considering any bit rate constraint, the quantization parameters (QPs) of key frames and WZ frames are selected offline by exhaustive search to provide maximum coding efficiency with similar quality for key and WZ frames. The offline exhaustive search approach is not viable for online applications. Also, in a video sequence, there are usually different scenes with different content and motion characteristics which are treated uniformly by this method. In [29], a quality control algorithm without any bit constraint was proposed in which the QP of key and WZ frames is dynamically adjusted to provide constant quality for both key and WZ frames. In this approach, the quantization level of each frequency band of the WZ frame needs to be obtained through an iterative loop which increases the complexity of the encoder. An RD performance loss of about 0.4 to 1.0 dB compared to the WZ coding solution without quality control was reported. In [30], a rate control algorithm for pixel-domain Wyner-Ziv video coding was proposed which estimated the rate and distortion of each video frame as a function of the coding mode and the QP. In [31], based on the motion activity between adjacent key frames, a table was suggested to select QPs of key and WZ frames for six different quality levels. No solution was suggested for an arbitrary target quality. In [32], to obtain similar quality for Intra and WZ frames, the relevant parameters are controlled jointly: QP for the

key frames and the quantization step size for the WZ frames.

In this chapter, we first propose a method to efficiently distribute the bit budget between key and WZ frames by modeling the relationship between quantization step size of a WZ frame and its neighboring key frames. We next apply this model to propose an adaptive algorithm to meet and maintain a target bit rate by dynamically adjusting the quantization parameters of key and WZ frames based on the residual energy between the WZ frame and the estimation of the side information at the encoder. We also evaluate the objective and subjective quality of our proposed method compared with the Discover method [33] where quantization parameters are predefined (offline exhaustive search).

This chapter is organized as follows: In Section 5.2, our method of finding the relationship between the quantization step size of key and WZ frames for efficient bit budget distribution is explained in detail. Our adaptive rate control algorithm is described in Section 5.3, and its objective performance is evaluated in Section 5.4. The subjective quality of our method is studied in Section 5.5. Section 5.6 concludes the chapter.

5.2 Dependence between Key and WZ frame quality

In Wyner-Ziv video coding, key frames are intra encoded and decoded by a conventional video codec. Therefore the rate distortion performance of key frames is independent of other frames. MCFI methods are applied on key frames to generate side information. Since side information is used in both the decoding and reconstruction blocks, the rate and distortion of WZ frames strongly depend on the quantization step size of the key frames, and on MCFI success. MCFI methods provide better estimation where the motion is smooth and translational. Therefore MCFI methods are usually more successful for low motion sequences than high motion ones. So, the compression efficiency of WZ frames is affected by their motion activity. In this section, we will investigate the relationship between the quantization step size of WZ frames and key frames in order to efficiently distribute

the bit budget between key and WZ frames. As a first step, the impact of the quality of key frames is studied for sequences with different motion characteristics. We used *Claire*, *Mother-daughter*, *Foreman*, and *Soccer* QCIF (176 × 144) sequences at 15fps which are different in content and motion characteristics.

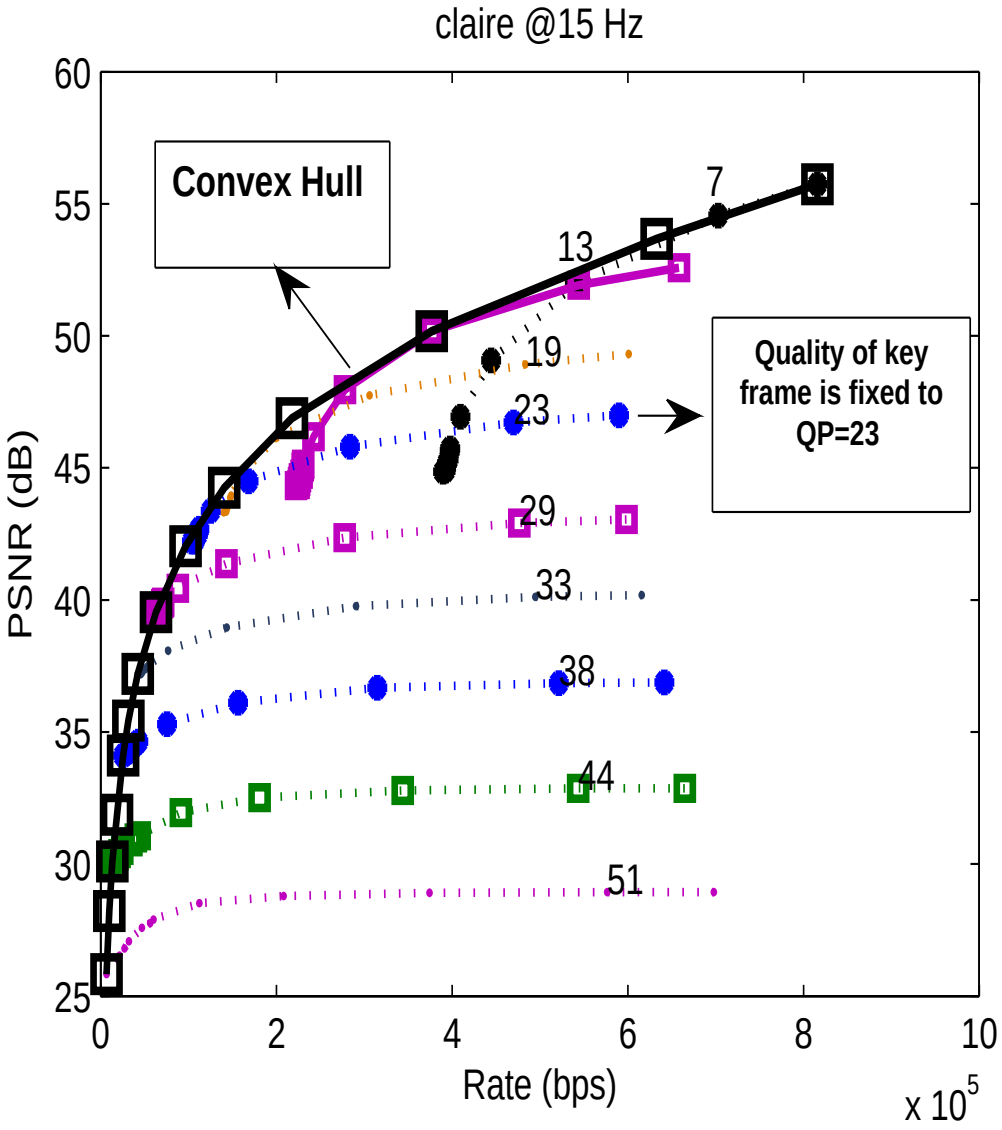


Figure 5.1: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Claire*

In Fig. 5.1- Fig. 5.4, the x axes show the average rate of all frames and the y axes show the average PSNR of all frames. Average PSNR was calculated

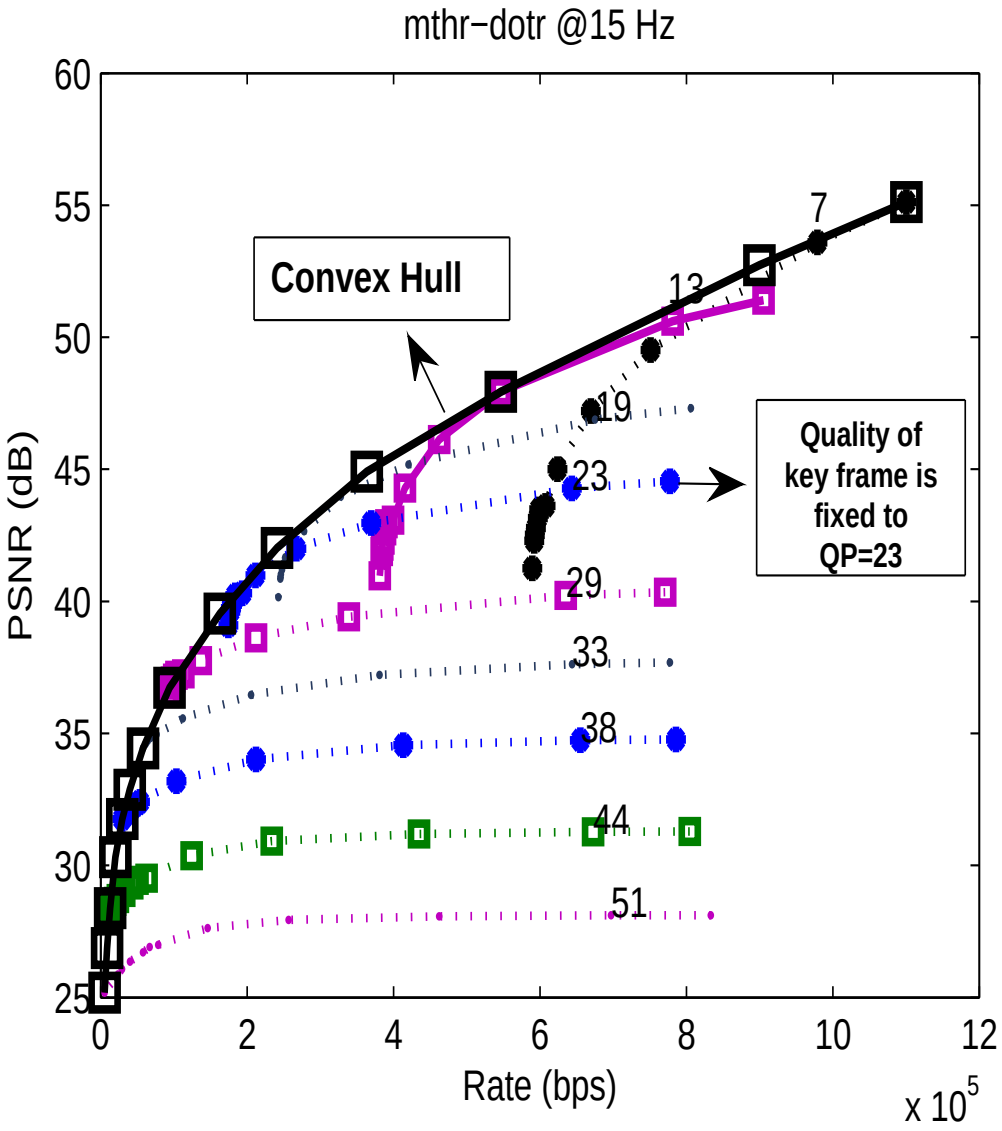


Figure 5.2: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Mother-daughter*

by first computing average MSE across all frames and then converting the final result to PSNR. The QP of key frames, QP_K, is fixed for each curve. There are several points on each curve representing different quantization step sizes of WZ frames. In H.264, QP values range from 0 to 51 and it is possible to calculate the equivalent quantization step size (Qstep) for each value of QP [34]. As QP increases, Qstep increases; in fact, Qstep doubles for every increase of 6 in QP [34].

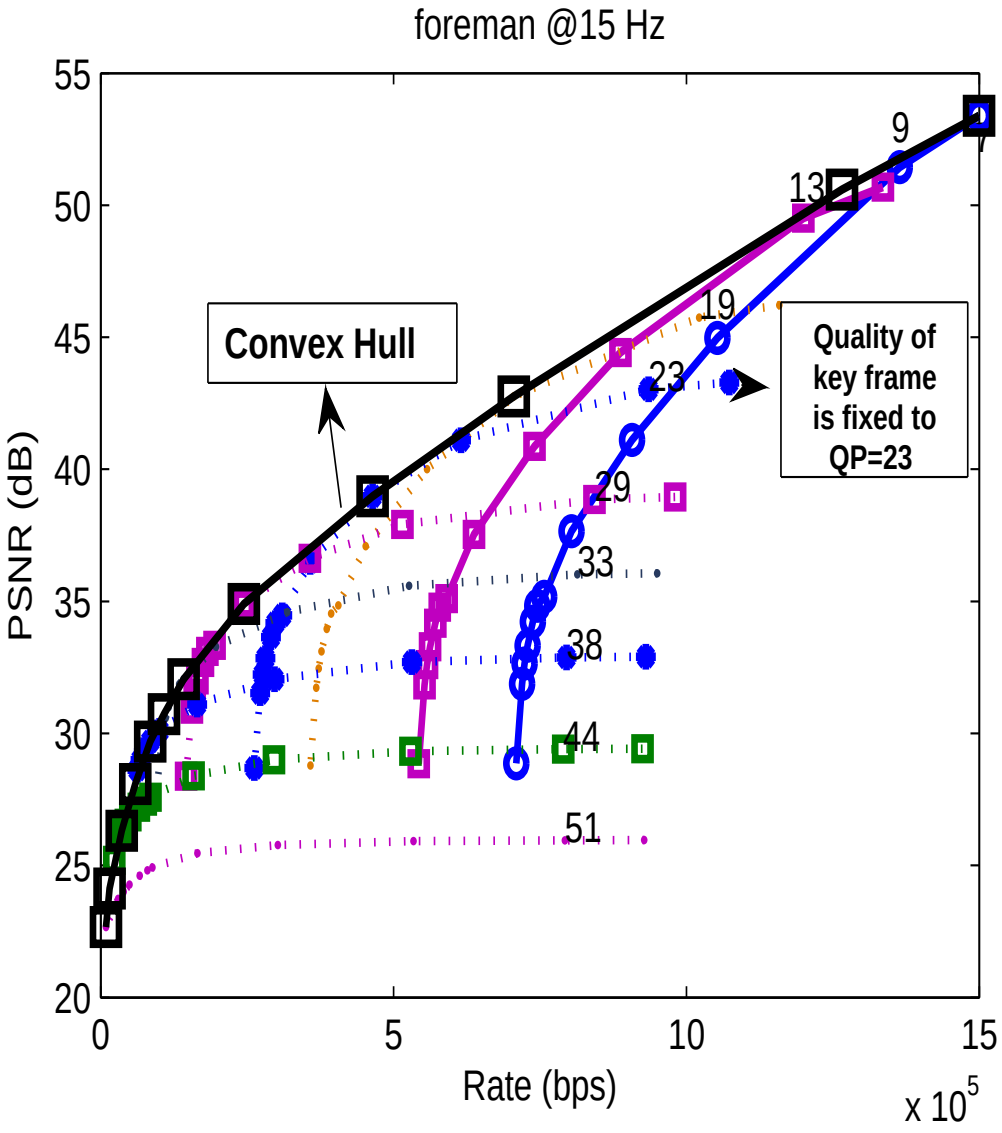


Figure 5.3: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Foreman*

In this work, the Qstep set for key frames is {1.375, 1.75, 2.25, 2.75, 4.5, 5.5, 7, 9, 11, 18, 22, 28, 40, 52, 72, 104, 224} corresponding to the QP set {7, 9, 11, 13, 17, 19, 21, 23, 25, 29, 31, 33, 36, 38, 41, 44, 47, 51}. The Qstep set for WZ frames is {0.03, 0.05, 0.195, 0.5, 1.47, 3.55, 4, 5, 7, 9, 12}. Each rate distortion (RD) point in these figures corresponds to a certain Qstep vector $QS = \{QS_K, QS_{WZ}\}$ for key frames and WZ frames. For each sequence, the convex hull of all these RD points

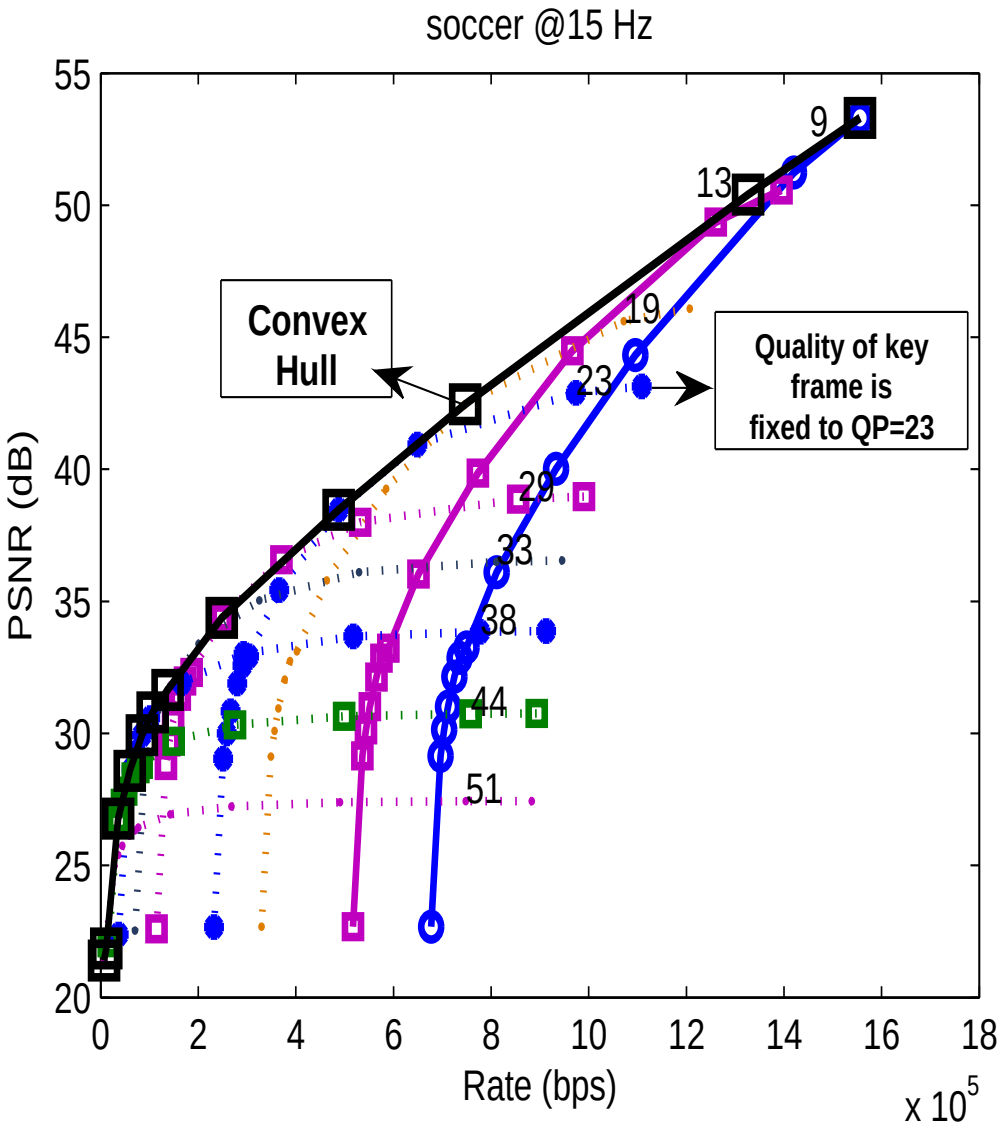


Figure 5.4: PSNR vs. rate of WZ codec at a given quality of key frames for different quality of WZ frames for *Soccer*

gives the optimum rate distortion curve. To find the best relationship between the Qstep of WZ and key frames, we study the set of QS vectors corresponding to RD points forming the convex hull of each sequence. Empirically, these sets of QS vectors for *Foreman* and *Soccer* which are relatively high motion sequences are the same. The ones for *Claire* and *Mother-daughter* which are relatively low motion sequences are also the same. So, we define two classes, low and high motion

activity.

As shown in Fig. 5.1- Fig. 5.4 at low rates, the convex hull is obtained by connecting the very first point of consecutive curves from the left side. These points belong to the cases where no bits are sent for WZ frames, and only side information generated by key frames is used to reconstruct them at the decoder. This indicates that the best bit allocation at low rates gives the total bit budget to key frames. It should be noted that the range of what is considered low rate is different based on the motion activity of the sequence.

5.2.1 High motion activity

MCFI methods become less successful where there is high motion activity. Therefore the error between a WZ frame and its corresponding side information grows and as a result more bits need to be sent. As shown in Fig. 5.1- 5.2, the QS vectors corresponding to RD points forming the convex hull for these relatively high motion sequences are $(1.75, 0.03)$, $(2.25, 0.05)$, $(7, 0.195)$, $(11, 0.5)$, $(40, 4)$, $(52, 5)$, $(72, 7)$, $(104, 12)$. For these QS vectors, QS_{WZ} is plotted vs. QS_K in Fig. 5.5. We use a curve fitting technique to estimate QS_{WZ} as a polynomial function of QS_K. We divide these points into two separate sets, S_1 = { $(1.75, 0.03)$, $(2.25, 0.05)$, $(7, 0.195)$, $(11, 0.5)$, $(40, 4)$, } and S_2 = {(40,4),(52,5),(72,7),(104,12)}, to have a more reliable estimation. A least squares fitting technique was used to find the best quadratic polynomial fit for each sample set, which for set S_1 is $f(x) = 0.00191\,x^2 + 0.002419\,x - 0.02457$, and for set S_2 is $f(x) = 9.5 \times 10^{-4}\,x^2 - 0.0124\,x + 3.01$. Fig. 5.5 (a) and (b) show data of sets S_1 and S_2 and their polynomial fits.

5.2.2 Low motion activity

As shown in Fig. 5.3- 5.4, the QS vectors corresponding to RD points forming the convex hull for the relatively low motion sequences are $(1.375, 0.03)$, $(1.75, 0.05)$, $(2.75, 0.195)$, $(4.5, 0.5)$, $(7, 1.47)$, $(11, 3.55)$, $(18, 10)$. The best polynomial fit of degree 2 for these points is $f(x) = 3.2 \times 10^{-2}x^2 + -3.17 \times 10^{-2}x + 1.8 \times 10^{-2}$. Fig. 5.5 (c) shows data of set S_3 and its polynomial fit.

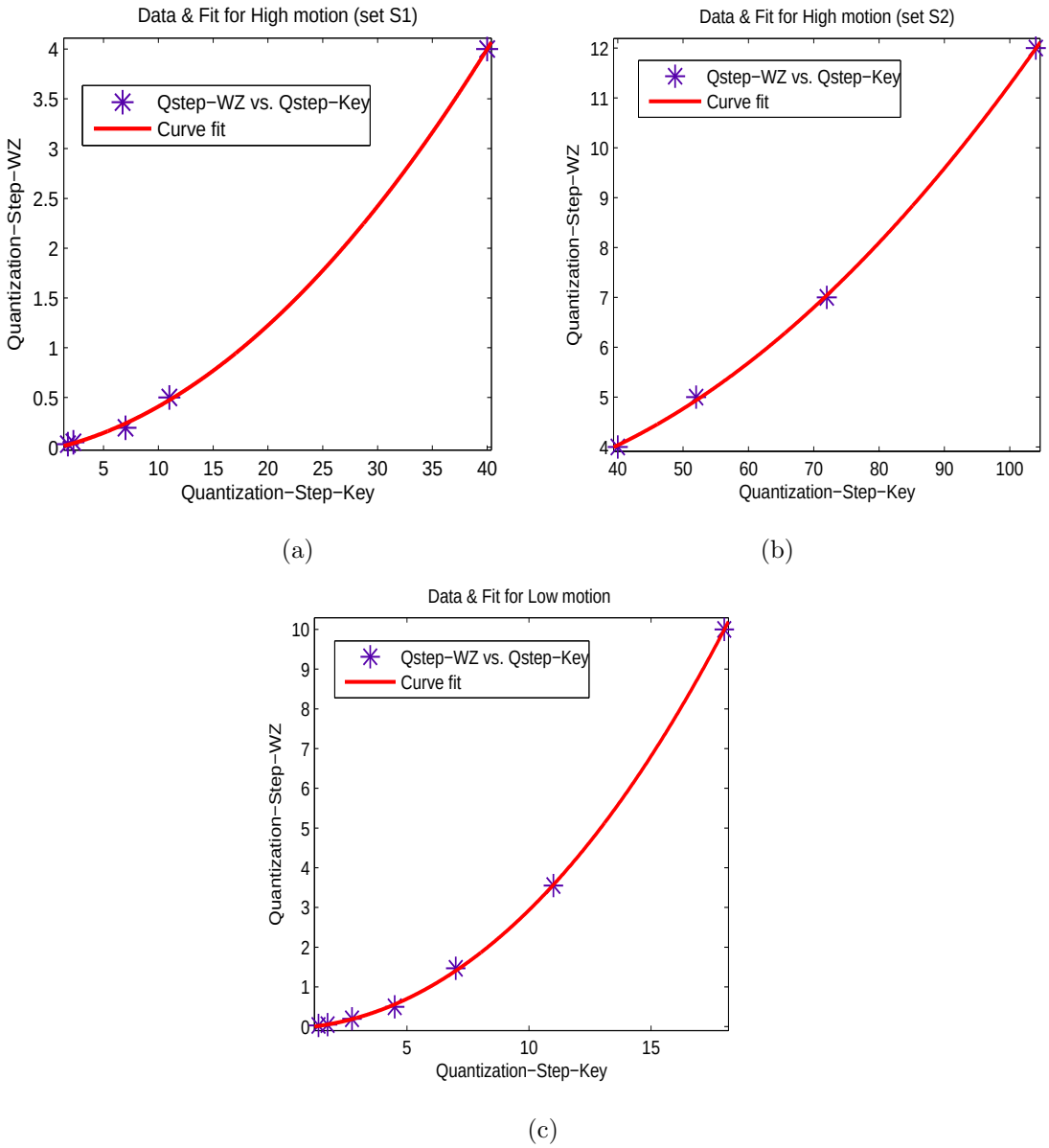


Figure 5.5: Polynomial fit to data points of high and low motion convex hull

5.3 Rate control

In the previous section, we found a relationship between the Qstep size of key and WZ frames in order to efficiently distribute the bit budget between key and WZ frames for low and high motion activity cases. We use this result to define an algorithm to first select Qstep sizes for key frames based on the target rate,

and then for WZ frames based on their motion activities and on the Qstep size of the corresponding key frames. The bit budget for each GOP (B_{GOP}) is calculated by $B_{GOP} = R_T \times \frac{N}{f}$ where R_T , N and f are the target rate, GOP size and frame rate, respectively. In this work, we consider a GOP of size 2 which is commonly used in the context of WZ video coding. Each GOP consists of a WZ frame and the next adjacent key frame.

5.3.1 Motion activity classification

To estimate the motion activity of a WZ frame, we estimate the side information at the encoder. The previous or next adjacent key frames or the average of both can be used at the encoder as a rough estimation of side information. From these three candidates, the one that minimizes the total absolute difference D between the candidate side information and the WZ frame is selected as the encoder side information. The candidate that minimizes D is called SI_e and its corresponding D is called D_{min}. D_{min} is calculated for each frame of our four sample sequences.

Table 5.1: Minimum, Maximum and Average of D_{min} over all frames of four sequences

D_{min}	$Minimum$	$Maximum$	$Average$
Soccer	1.54×10^5	7.63×10^5	3.71×10^5
Foreman	0.24×10^5	6.13×10^5	2.02×10^5
Mother-Daughter	0.16×10^5	0.76×10^5	0.37×10^5
Claire	0.11×10^5	0.46×10^5	0.22×10^5

Table 5.1 shows the minimum, maximum and average values of D_{min} over all frames for different sequences. To divide frames into low and high motion activity classes, a threshold value must be defined. We considered $Claire$ and $Mother-Daughter$ to be low motion sequences throughout, and so chose 7.6×10^4 which is the highest value of D_{min} over all frames of these two sequences to be the threshold value. A frame is classified as:

$$\begin{cases} Low\ motion & \text{if } D_{min} \leq 7.6 \times 10^4 \\ High\ motion & \text{otherwise} \end{cases} \tag{5.1}$$

Based on this threshold value, 100% of WZ frames of *Claire* and *Mother − Daughter* and 0% and 16% of WZ frames of *Soccer* and *Foreman* are considered to be low motion.

5.3.2 The relation between rate and Qstep size for key frames

For key frames, the source distortion is closely related to the quantization error which is controlled by the quantization step size. The relation between rate and quantization is usually derived based on a rate-distortion (R-D) model. In order to keep the encoder low complexity, in this work we use the simple R-D model [35]:

$$B_K = \frac{A}{QS_K} \tag{5.2}$$

where A is a constant, QS_K is the quantization step size and B_K is the number of coded bits for the frame. In practice, we start coding the first and third frames of the sequence which are key frames with some initial QP. A gets updated as

$$A = B_K \times QS_K \tag{5.3}$$

every time a key frame is coded. This is used to calculate QS_K for the next GOP. In this work, we set the initial QP based on the target average number of bits per pixel (*bpp*), calculated by

$$bpp = \frac{\frac{B_{GOP}}{N}}{W \times H} \tag{5.4}$$

where B_{GOP}, N, W and H are the bit budget for each GOP, GOP size, width and height of each frame, respectively. The initial QP is set as follows:

$$Initial\ QP = \begin{cases} 45 & \text{if } 0 < bpp \leq 0.3 \\ 35 & \text{if } 0.3 < bpp \leq 0.7 \\ 30 & \text{if } 0.7 < bpp \leq 1.5 \\ 25 & \text{if } bpp > 1.5 \end{cases} \tag{5.5}$$

As shown in Equation (5.5), threshold values are set to 0.3, 0.7 and 1.5. These values are selected based on R-D curves in Fig. 5.1- Fig. 5.4 to avoid being far away from the target rate in the beginning.

5.3.3 Choosing Qstep size for key frames

To choose the Qstep of key frames for each GOP, we need to find the relationship between the number of bits for each GOP and the corresponding QS_K. Then, given a target number bits, we can determine the Qstep. We use $\widehat{B}_{GOP}$ to denote the estimated number of bits for the next GOP. $\widehat{B}_{GOP} = \widehat{B}_K + \widehat{B}_{WZ}$ where $\widehat{B}_K$ and $\widehat{B}_{WZ}$ are the estimated numbers of bits for corresponding key and WZ frames. As explained before, the number of bits for a key frame can be estimated by $\widehat{B}_K = \frac{A}{QS_K}$ where A is calculated from the previously coded key frame. We define $c = \frac{B_{WZ}}{B_K}$ which gets updated every time a GOP is coded. This is used to estimate the number of bits for the WZ frame of the next GOP as:

$$\widehat{B}_{WZ} = c \times \widehat{B}_K \tag{5.6}$$

Therefore the number of bits for each GOP is estimated by:

$$\begin{aligned} \widehat{B}_{GOP} &= \widehat{B}_{WZ} + \widehat{B}_K \\ &= c \times \widehat{B}_K + \widehat{B}_K \\ &= (c+1) \times \widehat{B}_K \\ &= (c+1) \times \tfrac{A}{QS_K} \end{aligned} \tag{5.7}$$

by setting this equal to the bit budget of the GOP, we calculate QS_K as follows:

$$QS_K = (c+1) \times \frac{A}{B_{GOP}} \tag{5.8}$$

Since in the H.264 standard, only 51 different values can be used for Qstep, QS_K is set to the one closest to the calculated value by Equation (5.8). After a GOP is encoded, the actual number of bits used should be subtracted from the total available bit budget for this GOP and all future GOP to update the available bit budget for remaining GOPs. When the motion activity of the next GOP is very different from that of the current GOP, this would result in an abrupt change in Qstep and PSNR between GOPs which can be subjectively annoying. To limit this sharp change, our algorithm does not allow the QP_K difference between GOPs to exceed 2.

5.3.4 Choosing Qstep size for WZ frames

Once the QS_K is determined, QS_{WZ} which is the quantization step size of the WZ frame can be calculated. As a first step, D_{min} which estimates the motion activity of the WZ frame should be calculated as in Section 5.3.1. Then the WZ frame is classified as low or high motion using Equation (4). Based on the results of Section 5.2.1 and 5.2.2, QS_W is calculated as follows:

For a low motion activity WZ frame:

$$QS_{WZ} = \begin{cases} 3.2 \times 10^{-2} \times QS_K^2 - 3.17 \times 10^{-2} \times QS_K + 1.8 \times 10^{-2} & \text{if } QS_K \leq 18 \\ \infty & \text{otherwise} \end{cases} \tag{5.9}$$

$QS_{WZ} = \infty$ means that all bit budget goes for the key frame and no bits are sent for the WZ frame. For a high motion activity WZ frame:

$$QS_{WZ} = \begin{cases} 0.00191 \times QS_K^2 + 0.02419 \times QS_K - 0.02457 & \text{if } QS_K \leq 40 \\ 9.5 \times 10^{-4} \times QS_K^2 - 0.0124 \times QS_K + 3.01 & \text{if } 40 < QS_K \leq 104 \\ \infty & \text{otherwise} \end{cases} \tag{5.10}$$

5.4 Rate control simulation results

The test sequences are *Foreman*, $Hall-Monitor$, *Coastguard* and *Soccer* QCIF (176×144) at 15 fps. Fig. 5.6 (a)-(d) compare the rate-distortion perfor-

mance of the WZ video codec applying our proposed rate control algorithm to the one applying the method proposed in [33] (Discover) for the test sequences. As we can see, our proposed rate control algorithm which automatically selects the Quantization step size of key and WZ frames based on motion activity provides better or equal performance to Discover where the quantization parameters of key and WZ frames are selected offline. For $Hall - Monitor$, the performance gain is up to 0.8 dB, whereas for the other test sequences the gain is smaller or negligible. Note that the main purpose of our work is to accomplish on-line simple rate control to achieve a given target rate. The fact that our method also gives a slight performance improvement is an extra benefit.

Our success in achieving the target rate after coding each GOP is monitored by calculating $R_{GOP}(n)$ and shown in Fig. 5.7 (a)-(d). $R_{GOP}(n)$ is the actual rate achieved after coding the n^{th} GOP and is calculated by

$$R_{GOP}(n) = \frac{\sum_{i=1}^{2*n+1} B_F(i)}{2*n+1} \times f \tag{5.11}$$

where $B_F(i)$ is the number of coded bits for frame i, n is the GOP number and f is the frame rate which is 15 for our test sequences.

As we can see, our proposed method is capable of achieving the target rate quickly. For our test sequences the achieved rate is at most 10% different from the target rate after a maximum of 15 GOPs.

5.5 Subjective quality

In the previous section, we evaluated our method regarding rate-distortion performance and showed that the rate control goal is achieved at better or equal rate distortion performance to [33]. To avoid abrupt changes of PSNR between GOPs, we limited the QP difference between consecutive GOPs. The other concern is the quality difference between key and WZ frames. Since in our method we were aiming to minimize the distortion at a certain rate, we investigated the relationship between quantization parameters of key and WZ frames through the RD points located on the convex hull of different sequences. Table 5.2 shows the *average*

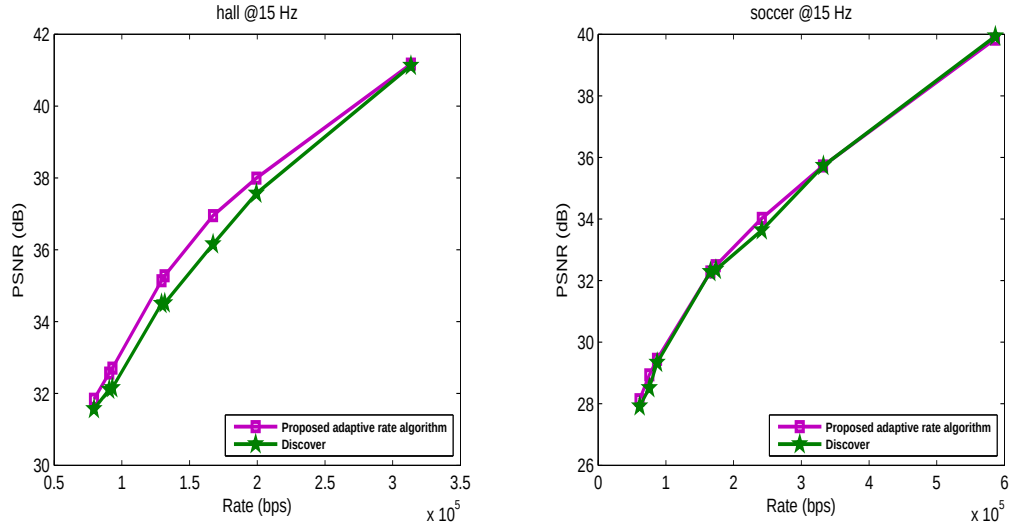


Figure 5.6: Rate-distortion performance of the WZ video codec applying our proposed rate control algorithm and the Discover method

PSNR for key and WZ frames at different RD points of the convex hull of two different sequences. In this table, P_K and P_{WZ} are the *average* PSNRs of key and WZ frames, and R_{WZ} and R_K are the *average* rates of WZ and key frames.

As we can see through Table 5.2, the *average* PSNR of key frames is usually higher than that of WZ frames. Our method gives moderately higher priority to

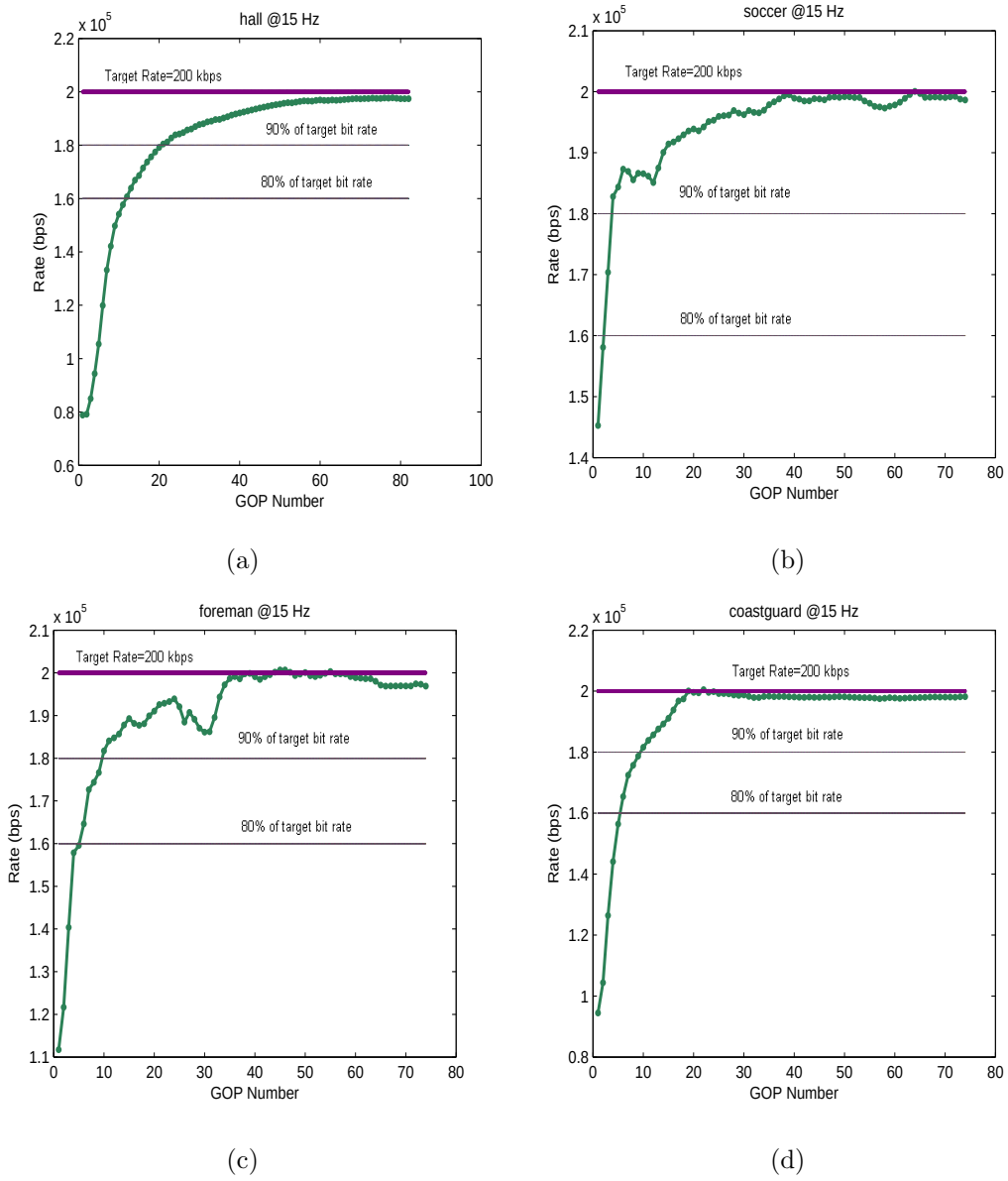


Figure 5.7: The achieved bit rate at each GOP for 15 fps sequences

key frames when distributing the bit budget between key and WZ frames as this improves the overall distortion. Giving more quality to key frames results in more PSNR fluctuation. The question may arise here of how much PSNR fluctuation is acceptable from the subjective quality point of view. We ran a limited subjective test to evaluate the subjective quality of our method compared with the method in [33].

Table 5.2: *Average* PSNR and Rate for key and WZ frames at different RD points of convex hull

(a)

	Claire @ 15 Hz			
(QS_{WZ}, QS_K)	P_K	P_{WZ}	R_{WZ}	R_K
$(0.030, 1.375)$	55.3	56.11	62.5×10^4	9.15×10^5
$(0.050, 1.750)$	55.6	53.79	30.1×10^4	7.76×10^5
$(0.195, 2.750)$	50.8	49.71	7.50×10^4	6.35×10^5
$(0.500, 4.500)$	48.2	45.84	1.65×10^4	5.12×10^5
$(1.470, 7.000)$	45.8	43.09	1.50×10^4	3.75×10^5
$(3.550, 11.00)$	43.2	41.11	1.35×10^4	3.18×10^5

(b)

	Foreman @ 15 Hz			
(QS_{WZ}, QS_K)	P_K	P_{WZ}	R_{WZ}	R_K
$(0.030, 1.750)$	52.9	53.91	69.1×10^4	19.0×10^5
$(0.050, 2.250)$	50.9	50.30	39.8×10^4	16.6×10^5
$(0.195, 5.500)$	43.9	42.02	8.40×10^4	14.1×10^5
$(0.500, 9.000)$	40.8	37.84	6.42×10^4	12.1×10^5
$(1.470, 18.00)$	36.0	33.97	3.87×10^4	9.44×10^5
$(3.000, 22.00)$	35.0	32.66	3.37×10^4	8.80×10^5

In [33], 8 different quantization tables are defined to encode WZ frames. For a given sequence, for each of these quantization tables, a quantization parameter to encode key frames is determined through offline exhaustive search to provide high coding gain at similar quality of key and WZ frames. To compare the subjective quality of our method against [33], we pick five of these RD points. We pick only five since as shown in Fig. 5.6 (a)-(d) some of them are very close to each other. These 5 RD points are chosen such that different bit rates from low to high are included. The rates of these five RD points will be the target rates for our rate allocation algorithm. In this way, for each sequence, we have five samples where the rate of our method is almost identical to that of [33]. So for each case, we compare the subjective quality of these two methods at equal rates.

5.5.1 Experimental design

The sample sequences are processed by our rate control algorithm and the method of [33] at 5 chosen rates. These 5 chosen rates are slightly different for each sequence. Table 5.3 shows the selected rates of each sequence for the subjective test. The outputs which are two versions of the decoded sequence at certain rates are recorded. At each rate, the recorded sequences are displayed in two windows side by side within a gray background. Different versions are placed randomly on the left or right. They start and stop at the same frame and are synchronized. The first video is a dummy video. The viewer is asked to decide which video has better quality. There are three options: right side, left side and the same quality. After displaying one video, the viewer is asked if he needs to replay to make a decision. It can be replayed as often as needed to make a decision. We include two replications (i.e. repetitions of identical conditions) to withdraw unreliable answers. We keep the result if the subject gives the same answer to both replications, and discard it otherwise [36].

Table 5.3: Five selected rates of different sequences for subjective test

Rate	r_1	r_2	r_3	r_4	r_5
Hall-Monitor	0.91×10^5	1.30×10^5	1.67×10^5	1.99×10^5	3.13×10^5
Foreman	0.85×10^5	1.64×10^5	2.31×10^5	3.31×10^5	5.12×10^5
Soccer	0.76×10^5	1.66×10^5	2.43×10^5	3.33×10^5	5.86×10^5
Coastguard	0.69×10^5	0.94×10^5	1.71×10^5	2.31×10^5	5.12×10^5

5.5.2 Subjective Quality results

Here also, the test sequences are $Foreman$, $Hall-Monitor$, $Coastguard$ and $Soccer$ QCIF (176×144) at 15 fps. Table 5.5 shows the *average* PSNR of Key frames, P_K, and the *average* PSNR of WZ frames, P_{WZ}, of both our method and [33]. In this table r_1, r_2,..., r_5 are the rates of all frames at 5 chosen RD points.

Fifteen observers participated in our experiment. A score of -1 is assigned if [33] is preferred, and $+1$ is assigned if our method is preferred. The score is zero if the observer chooses same quality. Table 5.4 shows the total score over

all participants for each sequence at each rate. Rates r_1,r_2,...,r_5 are the same as the ones in Table 5.3. The results in the table are surprising because one usually assumes that PSNR fluctuations are perceptually bad. These results suggest that a moderate amount of PSNR fluctuation which leads to *overall* reduced distortion can be slightly beneficial. As we can see, our method has been selected (positive values) as the one with better quality in 13 out of 16 cases where a preference is expressed. Results in this table show that our method which gives moderately higher priority to key frames provides a better perceived quality.

Table 5.4: Average score of subjective test over all participants for different sequences

	r_1	r_2	r_3	r_4	r_5
Hall-Monitor	2	3	0	0	-1
Foreman	7	1	3	1	-1
Soccer	3	5	4	1	2
Coastguard	-2	1	0	2	0

Table 5.5: *Average* PSNR for five different rates for key and WZ frames

(a) $Hall - Monitor$

		$r_1 = 91k$	$r_2 = 130k$	$r_3 = 167k$	$r_4 = 199k$	$r_5 = 313k$
P_K	Ours	33.59	36.24	38.23	39.57	43.04
	Discover	31.99	34.34	35.95	37.30	40.84
P_{WZ}	Ours	32.55	34.34	36.12	37.06	39.74
	Discover	32.31	34.06	36.37	37.86	41.42
PVAR	Ours	2.16	2.55	2.02	2.35	3.20
	Discover	0.14	0.11	0.13	0.15	0.12

(b) $Foreman$

		$r_1 = 85k$	$r_2 = 164k$	$r_3 = 231k$	$_4 = 331k$	$r_5 = 512k$
P_K	Ours	31.18	34.31	35.80	37.77	41.41
	Discover	29.96	32.64	33.92	36.06	39.25
P_{WZ}	Ours	29.01	31.80	33.65	35.42	38.67
	Discover	29.72	32.57	34.08	36.38	40.67
PVAR	Ours	5.34	5.22	4.45	4.58	4.72
	Discover	0.95	0.84	0.80	0.76	1.14

5.6 Conclusion

We investigated the relationship between quantization step sizes of key and WZ frames in a GOP of size 2 for WZ video coding in order to efficiently distribute the bit budget between key and WZ frames. The result was applied to propose a rate control algorithm which dynamically adjusts the quantization parameters to achieve a certain rate. In this method, the quantization step sizes of key and WZ frames was automatically adjusted based on the target rate and motion activity of the WZ frame. In our approach, GOPs are differentiated in selecting the Qstep of key and WZ frames based on their motion characteristics. Simulation results showed the proposed method achieves better (up to 0.8 dB) or equal rate distortion performance as [33] where there is no rate control. In [33] all quantization parameters are set offline for each individual sequence which is therefore applicable only for archival video. In contrast, our approach would be valid for low delay video coding and limited bandwidth or file size applications as all coding parameters are determined automatically on the fly with a bit rate constraint. The experimental results showed that the proposed method is successful at meeting the target rate after a few GOPs. Also, the subjective quality of our proposed method was compared to [33] and test results showed that our method provides slightly better perceptual quality even with slightly more PSNR fluctuation.

This chapter is in part a reprint of the material in the paper: G. Esmaili and P. Cosman, "Adaptive Rate Control for Wyner-Ziv Video Coding: Performance and Subjective Quality", *submitted to IEEE Transactions on Image Processing*, 2011.

Chapter 6

Conclusions

In today's video standards, the temporal correlation between adjacent frames is exploited by applying motion estimation and compensation techniques. Motion estimation which tries to find the best matching block in the reference frame for each block of the frame to be encoded requires intensively high computation. This makes the encoder 5 to 10 times more complex than the decoder. A "high-complexity" encoder with a "low-complexity" decoder is desirable for many video applications such as video streaming and video playback. In these cases, video is encoded once and decoded many times. For example, a movie which is encoded once and recorded on DVD will be decoded and played millions of times by different users. However, some emerging applications like video surveillance have very different requirements. In these applications, since we usually have so many encoders and just a few decoders, having low cost, low complexity encoder is essential. Wyner Ziv video coding which is based on the Slepian-Wolf and Wyner-Ziv theorems is a promising solution for such applications. In this approach, unlike predictive coding, the decoder is responsible to exploit similarities between adjacent frames. So, the complexity is shifted to the decoder. Existing Wyner-Ziv video coding methods use an encoder with the complexity close to intra coding but with the efficiency closer to inter coding than intra coding. In Chapter 2, the Transform Domain Wyner-Ziv video coding adopted in this dissertation was explained in detail.

In this dissertation, we proposed several techniques to enhance Tranform

Domain Wyner-Ziv video (TDWZ) coding in different ways. First we proposed a new method of correlation noise estimation to exploit the statistical dependency between source and side information. This approach which was based on block matching classification at the decoder was able to improve the coding gain without increasing the encoder complexity. Simulation results showed up to 2 dB improvement over "Conventional" and 1 dB improvement over the best proposed method (coefficient level) in [23].

Next, we proposed a technique to exploit temporal correlation between adjacent key frames by extending the idea of WZ coding to key frames. This method resulted in better performance by an advanced mode selection scheme for frequency bands of key frames followed by side information refinement. Then a hierarchical Wyner-Ziv coding approach including correlation noise classification and key frame coding mode selection was proposed. Simulation results showed that with the possible cost of additional buffering at the encoder, the proposed key frame encoding with side refinement combined with correlation noise classification results in up to 5 dB improvement over the Conventional method equipped with correlation noise classification. Experimental results showed that one can achieve up to 1 dB additional improvement by applying the hierarchical method at the cost of extra latency. All the proposed methods keep the encoder low in computational complexity.

Furthermore, we investigated the relationship between quantization step sizes of key and WZ frames in a GOP of size 2 for WZ video coding in order to efficiently distribute the bit budget between key and WZ frames. The result was applied to propose a rate control algorithm which dynamically adjusts the quantization parameters to achieve a certain rate. The adjustment is based on the target rate and motion activity of the WZ frame. In this approach, GOPs are differentiated in selecting the Qstep of key and WZ frames based on their motion characteristics. In this method, all coding parameters are determined automatically on the fly with a bit rate constraint which makes it valid for low delay video coding and limited bandwidth applications. Simulation results showed the proposed method achieves better (up to 0.8 dB) or equal rate distortion performance

compared to [33] where there is no rate control and all quantization parameters are set offline for each individual sequence which is therefore applicable only for archival video. Also, the subjective quality of our proposed method was evaluated and compared to a common method in the literature. Test results showed that our method provides slightly better perceptual quality even with slightly more PSNR fluctuation.

Bibliography

[1] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," *Proc. Asilomar Conference on Signals and Systems*, vol. 1, pp. 240–244, Nov. 2002.

[2] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Information Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.

[3] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. on Information Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[4] R. Puri and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles," *Proc. Allerton Conference on Communication, Control, and Computing*, Oct. 2002.

[5] R. Puri, A. Majumdar, and K. Ramchandran, "PRISM: A video coding paradigm with motion estimation at the decoder," *IEEE Trans. on Image Processing*, vol. 16, pp. 2436–2447, Oct. 2007.

[6] A. Aaron, S. Rane, and B. Girod, "Transform-domain Wyner-Ziv codec for video," *VCIP*, vol. 5308, pp. 520–528, January 2004.

[7] ——, "Wyner-Ziv video coding with hash-based motion compensation at the receiver," *Proc. IEEE ICIP*, vol. 5, pp. 3097–3100, Oct. 2004.

[8] A. Aaron and B. Girod, "Wyner-Ziv video coding with low encoder complexity," *Proc. Picture Coding Symposium*, Dec. 2004.

[9] A. Aaron, D. Varodayan, and B. Girod, "Wyner-Ziv residual coding of video," *Proc. Picture Coding Symposium*, April 2006.

[10] C. Brites, J. Ascenso, and F. Pereira, "Improving transform domain Wyner-Ziv video coding performance," *IEEE ICASSP*, vol. 2, pp. 525–528, May 2006.

[11] S. Argyropoulos, N. Thomosy, N. Boulgourisz, and M. Strintzis, "Adaptive frame interpolation for Wyner-Ziv video coding," *IEEE 9th Workshop on Multimedia Signal Processing*, pp. 159–162, Oct. 2007.

[12] S. Ye, M. Ouaret, F. Dufaux, and T. Ebrahimi, "Improved side information generation with iterative decoding and frame interpolation for distributed video coding," *Proc. ICIP*, pp. 2228–2231, Oct. 2008.

[13] W. A. R. J. Weerakkody, W. A. C. Fernando, J. L. Martinez, P. Cuenca, and F. Quiles, "An iterative refinement technique for side information generation in DVC," *IEEE Int. Conf. Multimedia Expo*, pp. 164–167, July 2007.

[14] R. Martins, C. Brites, J. Ascenso, and F. Pereira, "Refining side information for improved transform domain Wyner-Ziv video coding," *IEEE Trans. on Circ. and Sys. for Video Tech.*, vol. 19, pp. 1327–1341, Sep. 2009.

[15] J. Zhang, H. Li, Q. Liu, and C. W. Chen, "A transform domain classification based Wyner-Ziv video codec," *IEEE International Conference on Multimedia and Expo*, pp. 144–147, July 2007.

[16] L. Liu, D. He, A. Jagmohan, L. Lu, and E. Delp, "A low complexity iterative mode selection algorithm for Wyner-Ziv video compression," *Proc. IEEE ICIP*, pp. 1136–1139, Oct. 2008.

[17] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive distributed source coding using low-density parity-check codes," *Proc. Asilomar Conference on Signals and Systems*, pp. 1203–1207, Oct. 2005.

[18] G. Esmaili and P. Cosman, "Wyner-Ziv video coding with classified correlation noise estimation and key frame coding mode selection," *IEEE Transactions on Image Processing*, Dec. To be published.

[19] R. Westerlaken, R. Gunnewiek, and R. Lagendijk, "The role of the virtual channel in distributed source coding of video," *Proc. IEEE ICIP*, vol. 1, pp. 1–581–4, Oct. 2005.

[20] R. Westerlaken, S. Borchert, R. Gunnewiek, and R. Lagendijk, "Dependency channel modeling for a LDPC-based Wyner-Ziv video compression scheme," *Proc. IEEE ICIP*, pp. 277–280, Oct. 2006.

[21] L. Qing, X. He, and R. Lv, "Distributed video coding with dynamic virtual channel model estimation," *Int'l Symposium on Data, Privacy and E-Commerce*, pp. 170–173, 2007.

[22] C. Brites, J. Ascenso, and F. Pereira, "Studying temporal correlation noise modeling for pixel based Wyner-Ziv video coding," *IEEE International Conference on Image Processing*, pp. 273–276, Oct. 2006.

[23] C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding," *IEEE Trans. on Circ. and Sys. for Video Tech.*, vol. 18, pp. 1177–1190, Sep. 2008.

[24] X. Huang and S. Forchhammer, "Improved virtual channel noise model for transform domain Wyner-Ziv video coding," *ICASSP*, pp. 921–924, 2009.

[25] J. Ascenso, C. Brites, and F. Pereira, "Improving frame interpolation with spatial motion smoothing for pixel domain distributr," *5th EURASIP*, July 2005.

[26] A. B. B. Adikari, W. A. C. Fernando, H. K. Arachchi, and W. A. R. J. Weerakkody, "Low complex key frame encoding with high quality Wyner-Ziv coding," *IEEE International Conference on Multimedia and Expo*, pp. 605–609, Aug. 2006.

[27] G. Esmaili and P. Cosman, "Low complexity spatio-temporal key frame encoding for Wyner-Ziv video coding," *Data Compression Conference*, pp. 382–390, March 2009.

[28] ——, "Frequency band coding mode selection for key frames of Wyner-Ziv video coding," *11th IEEE International Symposium on Multimedia*, pp. 148–152, Dec. 2009.

[29] S. Sofke, F. Pereira, and E. Muller, "Dynamic quality control for transform domain wyner-ziv video coding," *EURASIP Journal on Image and Video Processing*, 2009.

[30] A. Roca, M. Morbee, J. Prades-Nebot, and E. J. Delp, "Rate control algorithm for pixel-domain Wyner-Ziv video coding," *Visual Communications and Image Processing*, vol. 6822, 2008.

[31] U. Cirac, M. Dalai, and R. Leonardi, "Adaptive key frame rate allocation for distributed video coding," *Picture coding symposium (PCS)*, November 2007.

[32] M. Jakubowski, J. Ascenso, and G. Pastuszak, "Constant bitrate control for a distributed video coding system," *SIGMAP*, pp. 131–138, 2008.

[33] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: Architecture, techniques and evaluation," *Picture Coding Symposium*, 2007.

[34] I. E. Richardson and I. E. G. Richardson, "H.264 and MPEG-4 video compression: Video coding for next generation multimedia," *WILEY*, pp. 191–193, 2003.

[35] J. Katto and M. Ohta, "Mathematical analysis of mpeg compression capability and its application to rate control," *Proc. IEEE ICIP*, pp. 555–559, 1995.

[36] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommuncation Union*, 1996.