

國立臺灣大學電機資訊學院電子工程學研究所



碩士論文

Graduate Institute of Electronics Engineering  
College of Electrical Engineering & Computer Science

National Taiwan University

Master Thesis

具時空間旁資訊分散式影像編碼

Distributed Video Codec with Spatiotemporal Side Information

李岳穎

Yueh-Ying Lee

指導教授：簡韶逸 博士

Advisor: Shao-Yi Chien, Ph.D.

中華民國壹佰零伍年參月

March 2016



# Distributed Video Codec with Spatialtemporal Side Information

By  
Yueh-Ying Lee

## THESIS

Submitted in partial fulfillment of the requirement  
for the degree of Master of Science in Electronics Engineering  
at National Taiwan University  
Taipei, Taiwan, R.O.C.

March. 2016

Approved by :

Shang-Chi Tai-Ling Wu Tsung-Hwa Tsai  
Chih-Ying Lee

Advised by :

Shang-Chi Tai-Ling Wu

Approved by Director :

Shen-Tzuan Lu

國立臺灣大學碩士學位論文  
口試委員會審定書  
具時空間旁資訊之分散式影像編碼  
Distributed Video Codec with  
Spatiotemporal Side Information

本論文係李岳穎君 (R02943006) 在國立臺灣大學電子工程學研究所完成之碩士學位論文，於民國一零五年三月二十二日承下列考試委員審查通過及口試及格，特此證明

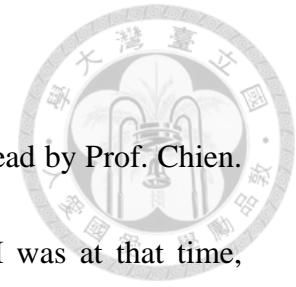
口試委員：

簡凱遠  
\_\_\_\_\_  
(指導教授)  
吳宗祥      蔡禮  
\_\_\_\_\_  
李俊哲      詹文璋  
\_\_\_\_\_

系主任、所長

劉深崎  
\_\_\_\_\_

# Acknowledgement



For about three years ago, I entered the Media & System IC Lab lead by Prof. Chien. Until today, I've always remembered how excited and aggressive I was at that time, contrary to the numerous adversities that I had met in the following few years. Lack of ken about the difficulty doing research, these hardship did discourage me. Sustaining support from peers and Prof. Chien helped me come through series of dilemma. People yearn for encouragements when depressed, and I'm so lucky to have the privilege to weather these predicaments with such warmth and assistance from lab. I sincerely appreciate all the aids received from anyone in my period as a master student.

So far, days of being a master student has almost come to an end. This experience is a treasure in my journey of life and never fade out, from which I learned how to properly equip myself with perseverance in face of adversity. Those halcyon days prattling with my most familiar seniors, Bear, Johnlin, and Dante, always cheer me up. Besides the yak, they also broadened my view by sharing their own experience and story. If not for their kindness, I would have been beaten by the failure. I can't thank them more.

In addition to specialty, duration as a graduate grants me the chance to introspect on myself and help me find what I am really seeking for. Always being a person pursuing eminence, I am just able to stand on the starting line after finishing this thesis.

Yueh-Ying

May, 2016



# 中文摘要



在這篇論文中，提出了一個具有時空間旁資訊分散式影像編碼的架構。這個提出的架構解決了以前分散式影像編碼在輸入的影片變動較大時影片壓縮效率不好的問題，這種變動很大的影片是在物聯網中由穿戴式相機或者是車用相機所拍到的影片無法避免的情形。實驗結果顯示，當和經典的 DISCOVER 分散式影像編碼比較時，平均具有 12.93% 的壓縮率改進，而在變動大的影片中這些壓縮率改進更加的明顯。除此之外，和 DISCOVER 分散式編碼相比，平均所需要的編碼時間只要其 92.3%。

這篇論文所提出的架構透過整合時間和空間上的預測系統來達到上述的壓縮率改進。這篇論文所提的架構的貢獻在於確立了在分散式影像編碼中使用空間預測系統時所需要使用的編碼工具以及相關的流程。在所提出的架構中採用了超解析度技術來產生空間的旁資訊預測，並且採用了支持向量機分類器來動態從時間或者空間預測結構中選擇。除此之外，在幀、塊、以及係數的層級上各採用了不同的編碼模式選擇以更進一步的提升壓縮表現。







# **Distributed Video Codec with Spatiotemporal Side Information**

*Yueh-Ying Lee*

*Advisor: Shao-Yi Chien*

*Graduate Institute of Electronics Engineering*

*National Taiwan University*

*Taipei, Taiwan, R.O.C.*

March 2016





# Contents

<b>Abstract</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Conventional Video Coding Systems . . . . .	2
1.2 Distributed Video Coding (DVC) . . . . .	4
1.3 Thesis Organization . . . . .	5
<b>2 Background Knowledge</b>	<b>7</b>
2.1 Distributed Source Coding (DSC) . . . . .	7
2.1.1 Slepian-Wolf Theorem . . . . .	7
2.1.2 Wyner-Ziv Theorem . . . . .	8
2.2 General DVC Framework . . . . .	10
2.2.1 Transform and Quantization . . . . .	11
2.2.2 Auxiliary Information . . . . .	11
2.2.3 Wyner-Ziv Coder . . . . .	12
2.2.4 Side Information Generation . . . . .	12
2.2.5 Correlation Noise Modeling . . . . .	13
2.2.6 Reconstruction and Post-processing . . . . .	15
<b>3 Motivation and Target Problem</b>	<b>17</b>
<b>4 Proposed Framework</b>	<b>21</b>
4.1 System Overview . . . . .	21

4.2	Coding Structure . . . . .	23
4.3	Proposed Encoder . . . . .	24
4.3.1	Frame Level Coding Structure Selection (CSS) . . . . .	24
4.3.2	Transform and Quantization . . . . .	25
4.3.3	Block Level Skip Mode . . . . .	26
4.3.4	Coefficient Level Coding . . . . .	27
4.4	Proposed Decoder . . . . .	28
<b>5</b>	<b>Experimental Results of the Proposed Framework</b>	<b>31</b>
5.1	Potential of Spatially-Predicted DVC . . . . .	32
5.1.1	Residual Coding . . . . .	32
5.1.2	Block Level Skip Mode . . . . .	33
5.1.3	Coefficient Level Coding . . . . .	34
5.2	CSS Classifier Training and Testing . . . . .	37
5.3	Overall Performance of the Proposed Framework . . . . .	41
5.3.1	RD Performance and BD Rate . . . . .	41
5.3.2	Running Time Evaluation . . . . .	45
<b>6</b>	<b>Conclusion</b>	<b>49</b>

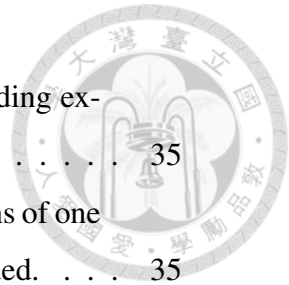




# List of Figures

1.1	Simplified encoder block diagram of HEVC [1]. . . . .	3
2.1	Distributed source coder . . . . .	8
2.2	Viable rate region of Slepian-Wolf theorem [2]. . . . .	9
2.3	Lossless DSC with side information Y. . . . .	9
2.4	Lossy DSC without side information at encoder. . . . .	10
2.5	General framework of DVC. . . . .	10
2.6	Algorithm of computing SI in [3]. . . . .	13
4.1	The proposed DVC framework. . . . .	22
4.2	All possible coding structure in the proposed framework. The number labelled on frames denote decoding order. The arrows indicate the prediction directions while decoding. The solid frames are key frames, and the dotted frames are WZ frames. . . . .	24
4.3	Block diagram of frame level coding structure selection. . . . .	25
4.4	Grouping of transform bands used for coefficient coding mode selection for temporally-predicted WZ frames. . . . .	28
5.1	BD rate of residual coding of one testing sequence of spatially-predicted DVC anchoring non-residual coded spatially-predicted DVC. . . . .	33
5.2	BD rate of different skip mode thresholds of one testing sequence anchoring non-skipped spatially-predicted DVC with residual coding. . . . .	34

5.3	Grouping of frequency bands used in coefficient level coding experiment for spatially-predicted DVC. . . . .	35
5.4	BD rate of different coefficients level coding configurations of one training sequence anchoring all groups being channel coded. . . . .	35
5.5	BD rate curve of testing sequence Foreman anchoring all groups being channel coded. . . . .	36
5.6	BD rate curve of testing sequence Coastguard anchoring all groups being channel coded. . . . .	36
5.7	BD rate curve of testing sequence Soccer anchoring all groups being channel coded. . . . .	37
5.8	BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework. . . . .	38
5.9	BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework. . . . .	38
5.10	BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework. . . . .	39
5.11	BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework. . . . .	39
5.12	BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework. . . . .	40
5.13	BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework. . . . .	40
5.14	RD curve of testing sequence Hall with common anchors. . . . .	42
5.15	RD curve of testing sequence Foreman with common anchors. . . . .	42
5.16	RD curve of testing sequence Coastguard with common anchors. . . . .	43
5.17	RD curve of testing sequence Soccer with common anchors. . . . .	43
5.18	RD curve of testing sequence Walk with common anchors. . . . .	44
5.19	RD curve of testing sequence Mountain with common anchors. . . . .	44





# List of Tables

4.1	Modified quantization table from [5]. . . . .	26
4.2	Coding modes and decision rules. . . . .	28
5.1	CSS classifier testing accuracy . . . . .	41
5.2	BD rate of the proposed framework with different anchors. . . . .	45
5.3	Encoding time comparison. . . . .	47
5.4	Decoding time comparison. . . . .	47







# Abstract

In this thesis, a DVC framework with spatiotemporal side information is proposed. The proposed framework addresses the shortage of poor compression performance for high-motion video sequences which is inevitable in Internet of Things(IoT) applications with wearable cameras and cameras in vehicles in previous DVC frameworks. Experimental results show that the average BD rate reduction of the proposed framework is 12.93% compared with DISCOVER DVC, and the coding gain is significant especially for high-motion sequences. Moreover, the average computing complexity is only 92.26% of DISCOVER DVC.

The proposed framework achieves compression performance gain by integrating both temporal and spatial prediction schemes into one framework. The major contribution lies in establishing coding flow and relevant coding tools for spatial prediction in DVC. Super-resolution technique is employed for spatially-predicted side information generation, and a support vector machine classifier is trained to adaptively select the coding structure between spatial and temporal prediction. In addition, coding mode selection at different granularities, including frame, block and coefficient levels, can further improve the coding performance.





# Chapter 1

## Introduction

Lately, Internet-of-Things (IoT) is viewed as the next generation of technology development. In IoT application, devices are thought to be everywhere, while at the same time, connect tightly. Most people agree with the layered IoT paradigm. IoT paradigm is comprised of several hierarchical layers, including sensing layer, network layer, logic layer, and application layer. The lower level of the hierarchy represents lower abstraction level of data at that layer. The bottom most layer, sensing layer, takes the responsibility to collect raw signal of the world, for example, video and audio signals. There will be large amount of data collected. As described above, sensing devices are assumed to be scattered everywhere, routing and management of limited bandwidth for data transfer becomes very important, which is tackled by the network layer. The amount of collected raw data is enormous, and thus beyond the order that human could handle. In that case, these collected raw data need to be analyzed and converted into higher level of abstraction. This process is done at logic layer. The semantics of the collected data would be extracted and the amount would decrease drastically. After the densified data is observed, human takes reaction to these data, which is the function of application layer.

The proposal of this thesis lies inbetween sensing layer and network layer. Among all kinds of data, video sequence occupies the transfer bandwidth most. In order

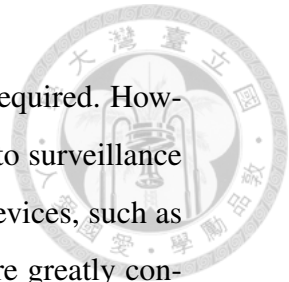
to prevent running out of limited bandwidth, video compression is required. However, under IoT application, video sensing devices are not limited to surveillance cameras anymore. It could also be cellphones, or some low-cost devices, such as raspberry Pi and so on. These terminal sensing device encoders are greatly constrained by computing resource or power consumption. Compared with conventional application, computing resource at decoder in IoT will be more sufficient than encoder. Under these constraints, conventional video codec may not be such suitable anymore, and the Distributed Video Codec (DVC) is proposed to address the problem.

## 1.1 Conventional Video Coding Systems

Conventional video codec, such as state-of-the-art HEVC, or H.265, is proposed to obtain good compression performance regardless of encoding complexity. Most of the application assumes plenty of computing resource available at encoder, for instance, video broadcasting. In these cases, compression performance casts much more importance than encoding complexity. Great compression rate could be obtained by making accurate prediction where the success of conventional video codec comes from.

Fig.1.1 shows the simplified block diagram of state-of-the-art video codec HEVC. As can be seen, input video signal is split into non-overlapped Coding Units (CUs). Every CU is predicted using either intra prediction or inter prediction. And the prediction residue is then transformed, quantized and entropy coded. Transform coding could help compact the residue energy, and thus fewer information requires encoding. Quantization scheme further reduces the information at the cost of quantization loss. Entropy coding represents information in an efficient way.

One feature of conventional video codec is decoder loop being integrated into encoder with an eye to both ensuring consistent behavior and exploring compres-



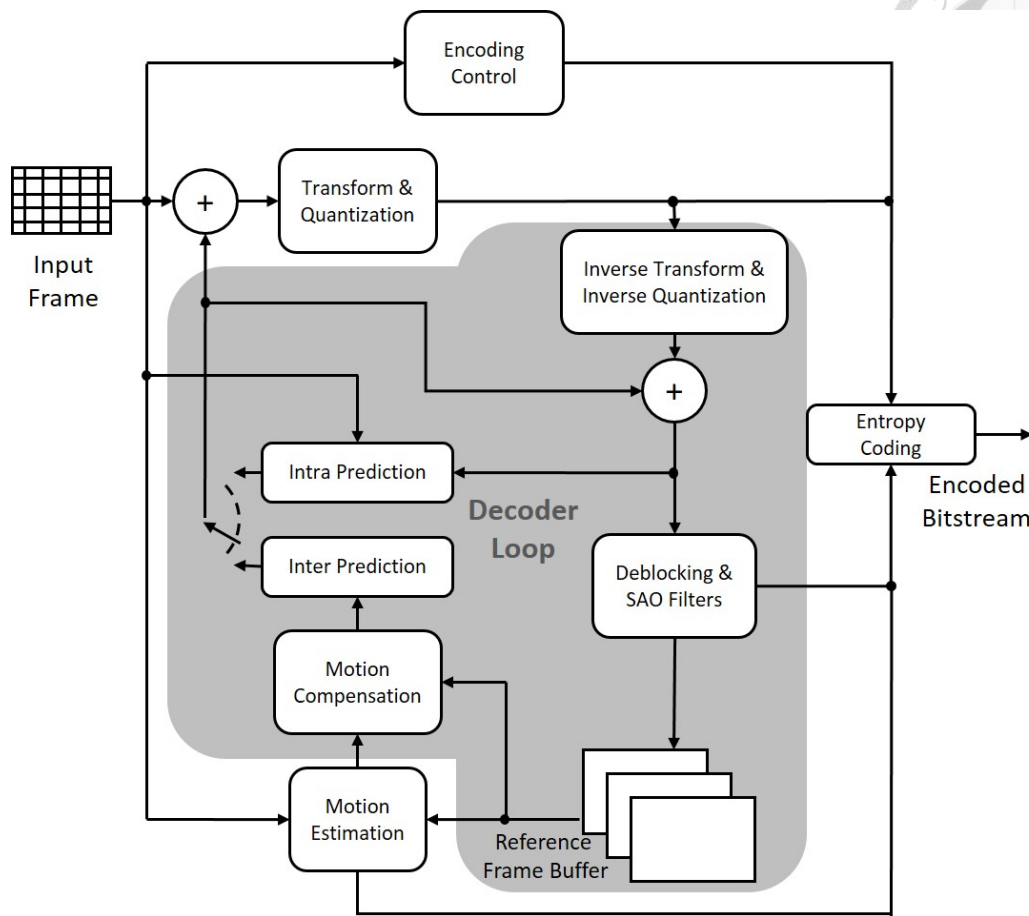
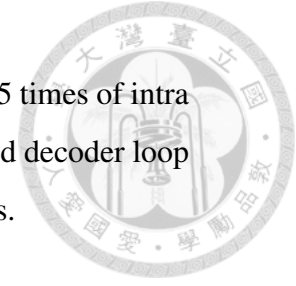


Figure 1.1: Simplified encoder block diagram of HEVC [1].

sion potential to the utmost. As mentioned before, performance of conventional video codec comes from accurate prediction. Intra prediction is conducted by using spatial information of neighbored CU, and there are 35 prediction methods for currently encoding CU in HEVC. Inter prediction is most effective for video encoding since video sequence bears high temporal correlation and redundancy. Inter prediction is mainly done by motion estimation of each prediction units, which utilizes temporal information as prediction of currently encoding CU. Motion estimation is conducted by block matching among all reference frames in the frame buffer and the resultant motion vectors is transmitted to decoder. Even though the compression is majorly obtained by inter prediction, it is the most complex part in conventional video codecs. [6] reports inter prediction loop comprises

more than 70% of the encoding computation in HEVC, more than 5 times of intra prediction. Both the complex prediction methods and encoder-sided decoder loop make conventional video codec not so suitable for IoT applications.



## 1.2 Distributed Video Coding (DVC)

In order to meet the application constraints of IoT, encoding complexity needs to be greatly reduced. In order to achieve this goal, distributed video codec is proposed with an eye to shifting encoding complexity to decoder. The underlying theorem of DVC is the Wyner-Ziv theorem [7]. Wyner-Ziv theorem is derived from Slepian-Wolf theorem [2] which states that for lossless compression scheme, it's possible to achieve Shannon bound under certain constraints with joint decoding. Wyner-Ziv further extend the Slepian-Wolf theorem to lossy compression scheme. The joint decoding process is done by making prediction at decoder only. The prediction at decoder is called Side Information (SI). SI is generated based on some frames or information that is called auxiliary information in this thesis transmitted from encoder without exhausting computing. The other frames are Wyner-Ziv encoded that is computed with very low complexity. For example, every two frame is intra encoded that is a lot less complex and transmitted to decoder, and decoder interpolates the SI between 2 intra coded frames. Then the SI is enhanced or corrected by Wyner-Ziv decoder. In this way, burden of inter prediction or motion estimation could be shifted to decoder.

In other words, conventional video codec could be viewed as SI exactly available at both encoder and decoder, and thus called symmetric coding scheme; while SI is only available at decoder in DVC, and thus called asymmetric coding scheme. With SI available at decoder only, DVC could remove decoding loop from encoder and shift computing burden to decoder. Although the encoding complexity problem is addressed, there is still compression performance gap compared to that of conventional video codec. To some extent, DVC is trade-off between complexity

and compression performance of conventional video codec.



### 1.3 Thesis Organization

The following of this thesis is organized as below. In chapter 2, background of DVC will be introduced, including underlying theorem and general DVC framework. Some related work will also be mentioned. In chapter 3, motivation and target problem of the proposed framework will be stated. Some literatures will be included to help locate the target problem. In chapter 4, the proposed framework will be introduced, including coding tools and coding structure of the proposed framework. In chapter 5, experimental results of the proposed framework will be shown. Coding tools are evaluated progressively and overall compression performance and encoding complexity of the framework are compared to some common anchors and frameworks. In chapter 6, conclusion is made based on experiment results.







## Chapter 2

# Background Knowledge

In this chapter, underlying theorem of DVC, Distributed Source Coding (DSC), will first be briefly introduced. Then general form of common DVC frameworks will be introduced. Besides, some detail about the most popular DVC framework, DISCOVER DVC on which the proposed framework made progress, will be explained, too.

### 2.1 Distributed Source Coding (DSC)

DVC is an implementation of distributed source coding (DSC). Slepian-Wolf theorem for lossless DSC is first proposed, and then Wyner-Ziv extend it for lossy DSC. DSC refers to the coding of two dependent random sequences, but with separate encoder. Each encoder sends a separate bitstream to a single decoder which may operate jointly on all incoming bitstreams and thus exploit the statistical dependencies. As shown in Fig.2.1.

#### 2.1.1 Slepian-Wolf Theorem

Consider two statistically i.i.d. random sequences  $X$  and  $Y$ . Assume with separate entropy encoders and decoders, one can achieve

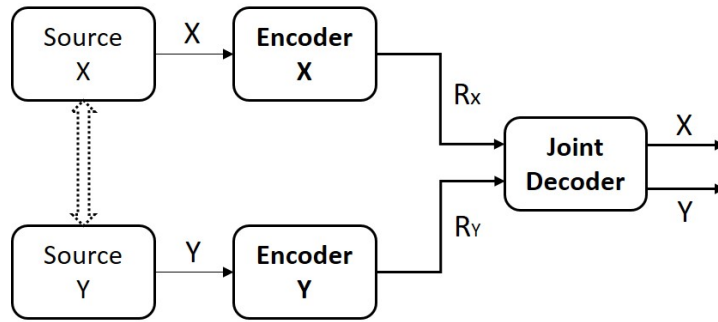


Figure 2.1: Distributed source coder

$$R_X \geq H(X) \quad (2.1)$$

$$R_Y \geq H(Y) \quad (2.2)$$

where  $H(X)$  and  $H(Y)$  are the entropies of  $X$  and  $Y$ , respectively.

Slepian-Wolf theorem suggests it is possible to do better with joint decoding but with separate encoding. Slepian-Wolf theorem establishes the rate region as

$$R_X + R_Y \geq H(X, Y) \quad (2.3)$$

$$R_X \geq H(X|Y), R_Y \geq H(Y|X) \quad (2.4)$$

And the viable rate region is shown in Fig.2.2.

Compression with decoder SI,  $Y$ , is a special case in DSC. Now the DSC in Fig.2.1 becomes the case shown in Fig.2.3.

Since coding  $Y$  at  $H(Y)$  is achievable, compression with decoder SI corresponds to the corner of the rate region in Fig.2.2, and hence  $R_X \geq H(X|Y)$ .

### 2.1.2 Wyner-Ziv Theorem

Wyner and Ziv extended Slepian-Wolf theorem to establish information-theoretic bounds for lossy compression with side information at decoder.  $X$  and  $Y$  represent samples of two i.i.d. random sequences, modeling source data and side information, respectively. And source values  $X$  are encoded without access to the side

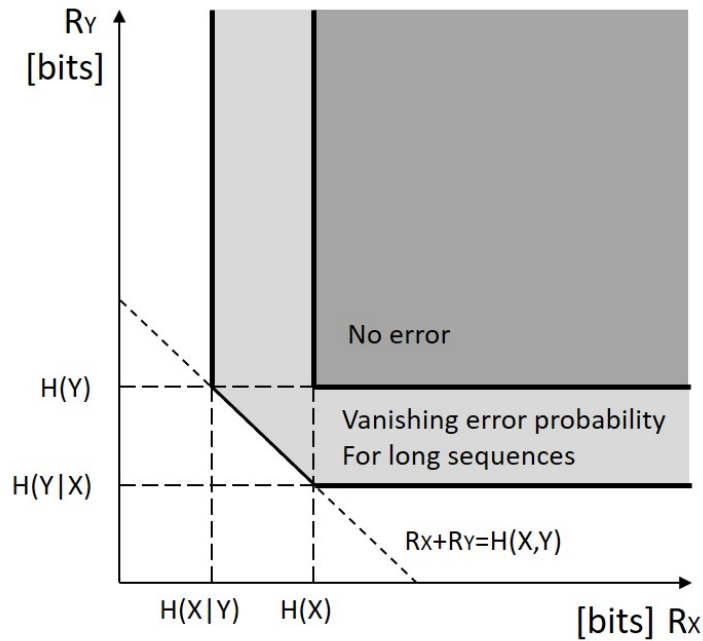


Figure 2.2: Viable rate region of Slepian-Wolf theorem [2].

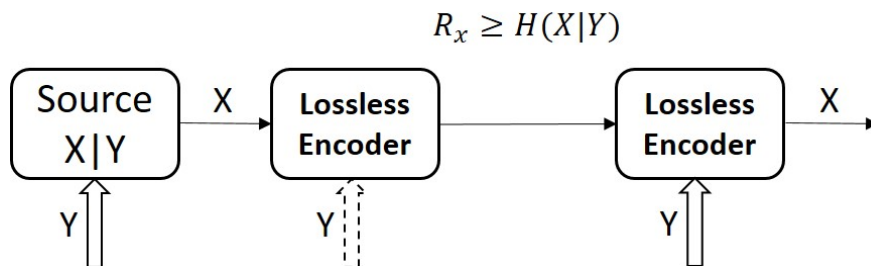


Figure 2.3: Lossless DSC with side information Y.

information Y. As shown in Fig.2.4. On the contrary, decoder has access to Y, and obtains a reconstruction  $\hat{X}$  of the source values X. Assume the acceptable distortion  $D = E[d(X, \hat{X})]$ . The rate-distortion function of Wyner-Ziv theorem given D is denoted by  $R_{X|Y}^{WZ}(D)$ .  $R_{X|Y}(D)$  denotes the minimum rate required if the side information were available at the encoder. Wyner and Ziv also proved that, unsurprisingly, a rate loss  $R_{X|Y}^{WZ}(D) - R_{X|Y}(D) \geq 0$  is incurred when the encoder does not have access to the side information.

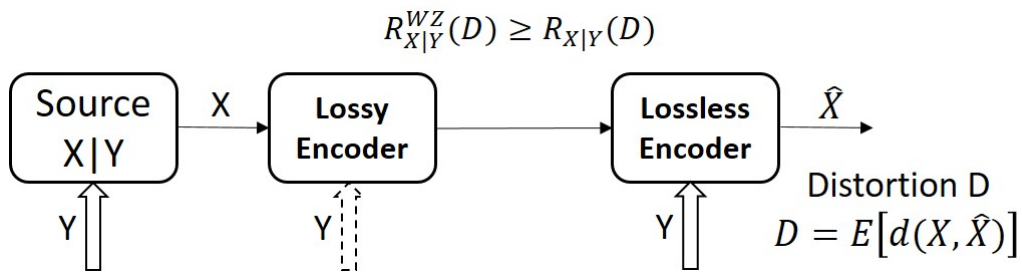


Figure 2.4: Lossy DSC without side information at encoder.

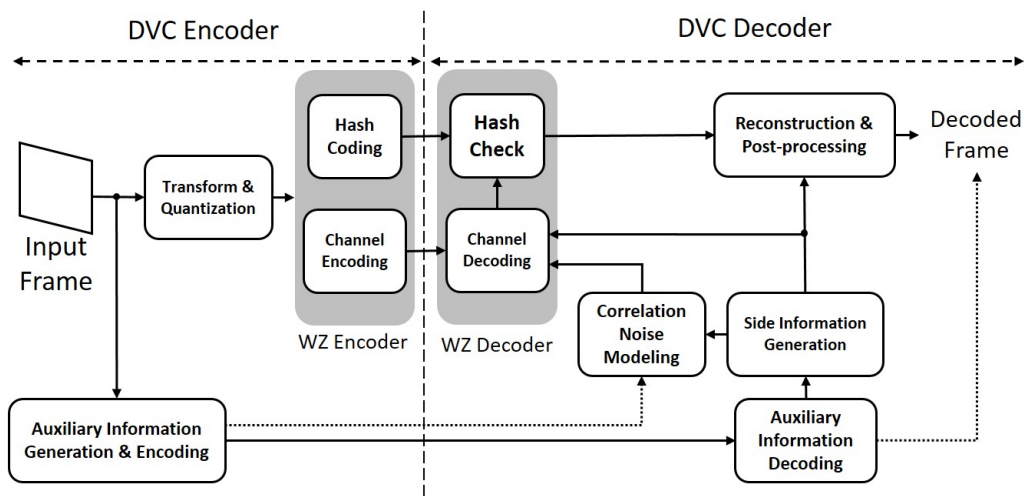
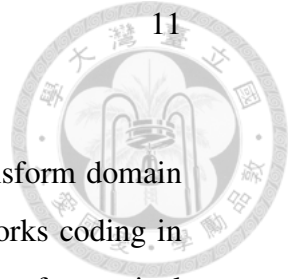


Figure 2.5: General framework of DVC.

## 2.2 General DVC Framework

Among DVC literatures, frameworks could be summarized in a general form as shown in Fig.2.5. These functional blocks will be introduced in the following sections. Conceptually, DVC shifts the prediction burden from encoder to decoder by transmitting some auxiliary information that does not require much computation but useful for generating SI. SI is viewed as a noisy version of original frame at decoder. Some bitstream encoded by the so-called Wyner-Ziv (WZ) encoder are transmitted as well. WZ encoded bitstream is capable of correcting erroneously computed SI. To summarize decoding flow of DVC, decoder would compute SI with received auxiliary information, and SI will then enhanced or corrected by WZ decoder.



### 2.2.1 Transform and Quantization

Most DVC frameworks in literatures adopt block-based DCT transform domain coding as in conventional video codecs. There are some frameworks coding in pixel domain. However, almost all the transform domain DVC outperforms pixel domain ones.

As for the quantization scheme, most frameworks utilize scalar quantization rather than vector quantization due to complexity issue. There are 2 kinds of methods to determine quantization levels. The first is simply utilizing predefined quantization levels according to different configurations, such as the classical framework from Stanford [8]. The other is to compute suitable quantization levels based on transformed coefficients, as the method in [9] from Berkeley. The advantage of the former method is extremely low quantization complexity because of table lookup operation; while the latter could better sustain video quality for variant video sequence for a given target rate.

### 2.2.2 Auxiliary Information

No matter what SI generation algorithm is used at decoder, some auxiliary information is required. Auxiliary information should be able to compute with low complexity. We may consider motion vectors as auxiliary information of conventional video codec but requiring exhausting computation. The auxiliary information functions as an approach to shift complexity between encoder and decoder. The finer the auxiliary information is, the less prediction complexity required at decoder. For example, in the classical framework [8], the auxiliary information is the intra-coded key frames, and SI is interpolated using temporally neighbored key frames. On the contrary, [9] conducts motion estimation depending on how much resource available at encoder. If there being abundant computing resource at encoder, construction of SI will be conducted at encoder, and prediction residue between original frame and SI would be fully known by WZ encoder, and thus results in better compression performance. The previous case is just the same as

conventional video codec. Otherwise, zero motion prediction will be used as a estimation of SI at encoder.



### 2.2.3 Wyner-Ziv Coder

Wyner-Ziv (WZ) coder is the core of DSC implementation, it majorly comprised of channel code and hash code. As mentioned before, DVC is viewing SI as a noisy version of original frame, and the error could be corrected by channel code commonly used in communication. Originally, channel code is the parity bits concatenated after the source bitstream, and the rate of the resultant bitstream will increase. However, only the parity bits are transmitted in DVC. As long as the parity bits required to correct SI is less than source bitstream, compression is obtained. Turbo code [10] and Low-Density-Parity-Check (LDPC) code [11] are widely used in DVC. Rate-adaptive LDPC code, called LDPCA [12], is proposed and outperforms turbo code. Therefore, most of DVC frameworks adopts LDPCA code as channel code.

It is possible that channel code converge to a wrong result. Therefore, in addition to channel code, hash code is also transmitted to guarantee the correctness of channel decoding. Cyclic-Redundancy-Check (CRC) code is widely adopted in DVC.

### 2.2.4 Side Information Generation

For frameworks similar to [8] which uses group of picture (GOP) coding structure, once the temporally adjacent key frame is decoded, SI of currently decoding WZ frame will be computed using motion compensated frame interpolation algorithms. As described in 2.2.3, WZ decoder view SI as a noisy version of original frame, and correct the error part, which means that good SI quality does not require much rectifying, and thus less rate is needed. Many algorithms are proposed to refine SI quality. The SI generation algorithm adopted in [3] is briefly introduced here and shown in Fig.2.6.

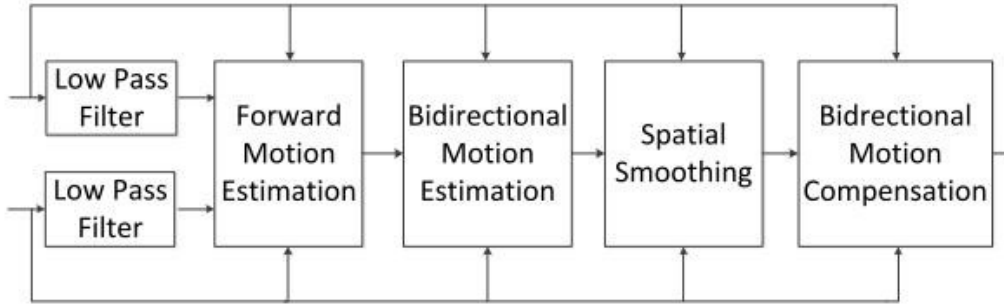


Figure 2.6: Algorithm of computing SI in [3].

Two temporally adjacent key frames are first low-pass filtered, and then single direction motion estimation is conducted. The resultant motion vectors are further refined using bidirectional motion estimation and spatial smoothing. Finally, motion compensation is conducted based on the refined motion vectors.

### 2.2.5 Correlation Noise Modeling

Because of the utilization of channel code, either soft or hard input for channel decoder is required. Literature reveals soft input for channel decoder performs better. As explained, DVC views SI,  $Y$ , as a noisy version of original frame  $X$ , and tries to correct the error using channel code. Therefore, it is required to come up with a way to model the residue  $Z$  between  $X$  and  $Y$  such that channel code could know where error probably takes place in SI. The model is called Correlation Noise (CN) model. The residue is formulated with  $Z = X - Y$ , and modeled with Laplacian distribution of zero mean as

$$f_Z(z) = \frac{\alpha}{2} e^{-\alpha|z|} \quad (2.5)$$

where  $\sigma^2 = 2/\alpha^2$

However, it is not possible to exactly know  $X$  at decoder of DVC framework. Besides, it is not practical to obtain exact  $Y$  at encoder, which is opposite to the goal of DVC. The more accurate the correlation noise model is, the less rate of channel

code requires to correct error in SI. Method of calculating CN model can be classified into 2 categories, off-line or on-line. Off-line CN model has problem adapting to different sequences or temporal variation within certain sequence. Therefore, recent DVC frameworks utilize on-line methods to compute CN model. On-line modeling could be done in 2 ways, either computing at encoder with coarse SI generation and then transmitting to decoder, or computing at decoder using available data only. The latter is more preferred with regards to both bitrate overhead and computing complexity.

In addition to method of computing CN model, at what granularity the model is obtained is also important, ranging from sequence level toward block level. [13] concludes that CN modeling at finer granularity, block level, adapt to contents variation better. In other words, every transformed block shares one CN model where spatial characteristics within the block is also taken into consideration. The most widely adopted on-line CN modeling [13] is conducted in the following way. In the beginning, a residual frame based on motion compensated frame is computed

$$R(x,y) = \frac{X_f(x + dx_f, y + dy_f) - X_b(x + dx_b, y + dy_b)}{2} \quad (2.6)$$

where  $X_f, X_b$  denote forward and backward unilateral motion compensation with best motion vector candidate  $(dx_f, dy_f)$  and  $(dx_b, dy_b)$  individually.  $(x, y)$  pair specify the pixel coordinate in R frame.

Then the variance of the residual frame is computed as

$$\hat{\sigma}^2 = E_R[R(x,y)^2] - (E_R[R(x,y)])^2 \quad (2.7)$$

$$\hat{\alpha}_R^2 = \sqrt{\frac{2}{\hat{\sigma}^2}} \quad (2.8)$$

Where  $\hat{\alpha}_R$  is the estimation of  $\alpha$  in the Laplacian model. Finally, variance of k-th block,  $R_k$ , is compared to residual frame variance to achieve content adaptation





as

$$\hat{\alpha}_{R_k} = \begin{cases} \hat{\alpha}_R, & \hat{\sigma}_{R_k}^2 \leq \hat{\sigma}_R^2 \\ \sqrt{\frac{2}{\hat{\sigma}_{R_k}^2}}, & \text{otherwise,} \end{cases} \quad (2.9)$$

## 2.2.6 Reconstruction and Post-processing

For channel decoded frames, inverse quantization and inverse DCT are applied. The optimal inverse quantization that minimizes mean squared error of source  $X$  proposed by [14] given SI  $Y$  is computed as

$$\hat{x}_{opt} = E[x | x \in [z_i, z_{i+1}, y]] = \frac{\int_{z_i}^{z_{i+1}} x f_{x|y}(x) dx}{\int_{z_i}^{z_{i+1}} f_{x|y}(x) dx} \quad (2.10)$$

where conditional probability function  $f_{x|y}$  is derived from CN model, and  $z_i$  is the quantization boundary. And the close form solution can be represented with

$$\hat{x}_{opt} = \begin{cases} z_i + \frac{1}{\alpha} + \frac{\Delta}{1 - e^{-\alpha\Delta}}, & y < z_i \\ y + \frac{(\gamma + \frac{1}{\alpha})e^{-\alpha\gamma} - (\delta + \frac{1}{\alpha})e^{-\alpha\delta}}{2 - (e^{-\alpha\gamma} + e^{-\alpha\delta})}, & y \in [z_i, z_{i+1}) \\ z_{i+1} - \frac{1}{\alpha} - \frac{\Delta}{1 - e^{-\alpha\Delta}}, & y \geq z_{i+1} \end{cases} \quad (2.11)$$

where  $\Delta = z_{i+1} - z_i$ ,  $\gamma = y - z_i$ ,  $\delta = z_{i+1} - y$

Once the inverse quantization and inverse transform are done, pixel domain reconstruction is obtained. However, block artifacts are likely to arise from block-based compression scheme. So deblocking filters could be applied to enhance decoded frame quality, such as filter used in HEVC [15] or other algorithms [16].





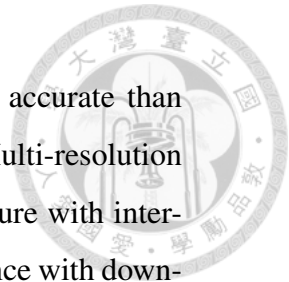
## Chapter 3

# Motivation and Target Problem

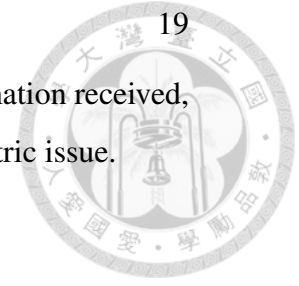
So far, DVC has made great progress since the first two DSC implementations [8] [9] were proposed. The most well-known and adopted framework is DISCOVER DVC framework [3]. The DISCOVER framework basically follow the architecture as in [8] but much more mature in almost all aspects. Full evaluation of DISCOVER DVC is available at [5]. Based on [3], coding techniques used in conventional video codecs have been integrated into DVC framework. For example, in both [17] and [4], residual coding and block level skip mode is used to reduce bit rate. In addition to block level skip mode, Chiu et al. further employs different source coding techniques at the transform coefficient level according to the prediction residue energy [4]. With the state-of-the-art works, great BD rate reduction of low-motion video sequences, for instance, surveillance video sequences captured with fixed cameras, can be successfully achieved. However, frameworks similar to [3] basically employ the motion-compensated SI generation scheme based on the temporal correlation between frames. They cannot work well when there is no high correlation between key frames and WZ frames, such as the case of videos captured by wearable cameras or vehicle cameras. Therefore, the proposed framework aims to meliorate contemporary DVC frameworks on this kind of high-motion or first-person-view videos that are inevitable in IoT applications.

Inspired by conventional video codec, spatial prediction is more accurate than temporal prediction when there is not much temporal correlation. Multi-resolution video coding could be a feasible solution. In [18], a coding structure with inter-layer prediction is proposed. It deals with high motion video sequence with down-sampled base layer as key frame and enhancement layer as WZ frame. It follows a fixed coding structure with conventionally inter-predicted encoding scheme among key frames. There's certainly some performance loss. Therefore, the proposed framework aims to adjust prediction structure based on input frame and integrate inter-layer prediction scheme into existing GOP-based DVC framework.

Furthermore, confirmed by [4] [17] [19] [20], coding mode selection at different granularities could increase adaptability and compression performance for variety of input sequences. [4] and [17] proposed encoder-driven coding mode selection; while [19] and [20] proposed decoder-driven coding mode selection. The latter two reveals better compression performance than the former two. Decoder-driven coding mode selection is done by first computing SI based on the received key frame, and explores correlation between SI and key frame in order to determine how to efficiently encode current WZ frame. It is not until encoder receive the determined coding mode from decoder that it starts to encode current WZ frame. Performance improvement of decoder-driven coding mode selection result from the symmetric behavior at WZ encoder and WZ decoder, which is the major issue to be addressed in all DVC frameworks. Decoder-driven coding mode selection, though it results in better compression performance, has the risk of encoder idling due to the latency from SI generation at decoder when more complex algorithm is adopted to compute SI. Therefore, the proposed framework still adopts encoder-driven coding mode selection. Besides, DVC compression performance greatly depend on how much encoder is aware of SI, just the symmetric issue mentioned above. Most DVC frameworks use key frames as auxiliary information as DISCOVER does. Enlightened by [9] and [21], a classifier at encoder could help



encoder to gain idea on how SI might be based on auxiliary information received, and is adopted in the proposed framework to deal with the symmetric issue.







# Chapter 4

## Proposed Framework

### 4.1 System Overview

The proposed DVC framework is shown in Fig.4.1. It is developed based on [4], and the newly added blocks are shown as dotted blocks. The coding structure selection (CSS) first decides whether the input frame is suitable to be predicted temporally or spatially with a support vector machine (SVM) classifier. It aims to find suitable auxiliary information for decoder to generate high quality SI. As can be seen in Fig.4.1, no matter which kind of prediction is adopted, prediction residue is encoded, just as residual coding in conventional video codec. Auxiliary information is selected to be intra coded frames using conventional video codec, and is called key frame in this thesis. Besides, only GOP size 2 is considered in this thesis since the target problem is high-motion input sequence where larger GOP size only deteriorates compression performance.

If it is decided to employ temporal prediction, the residue to be encoded is

$$R_{temp} = f_t - REC_{intra}(f_{t-1}) \quad (4.1)$$

where  $f_t$  is the input frame at time t,  $REC_{intra}(\cdot)$  is the operation of reconstruction (decoding) after intra coding, and the following coding encoding flow is the same as [4]. Auxiliary information in temporal prediction is temporally prior key frame,

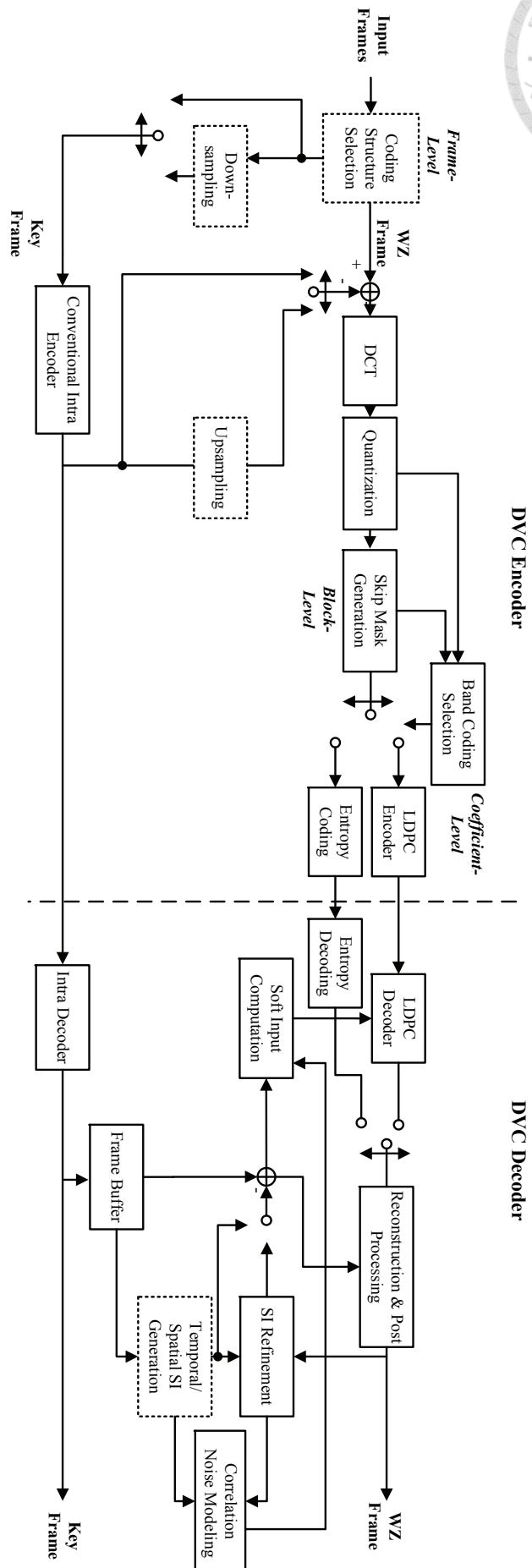


Figure 4.1: The proposed DVC framework.



$REC_{intra}(f_{t-1})$ .

On the other hand, if spatial prediction is employed, the residue to be encoded is

$$R_{spat} = f_t - UP(REC_{intra}(DOWN(f_t))) \quad (4.2)$$

where  $UP(\cdot)$  is the upsampling operation and  $DOWN(\cdot)$  is the downsampling with anti-aliasing operation. Auxiliary information in spatial prediction is down-sampled key frame of currently encoding frame,  $REC_{intra}(DOWN(f_t))$ .

Both temporal and spatial coding ow will be explained in the following sections. CSS selects the coding mode for each frame, which is called frame-level prediction selection in this thesis. After the frame-level prediction mode is determined, block-level skip mode and coefcient-level coding mode selection is conducted, similar to what is done in [4].

## 4.2 Coding Structure

Since the spatial prediction is integrated into the proposed framework, the coding structure will be different from the group of picture (GOP)-based coding structure used in [3] [4] [8]. As mentioned before, frames coded by the conventional intra encoder are called key frames, while frames coded by the WZ encoder are called WZ frames. In Fig.4.2, all possible coding structures of the proposed framework are shown. For a low-motion sequence, such as a surveillance video sequence, a GOP-based temporally-predicted coding structure with interleaved intra key frames, as shown in Fig.4.2(a), is adopted. For a high-motion sequence, however, a spatially-predicted coding structure with downsampled low-resolution frames as key frames and full-resolution frames as WZ frames, as shown in Fig.4.2(b), is adopted. For a general hybrid video sequence, a coding structure with both spatial prediction and temporal prediction is employed. Fig.4.2(c) shows one possible coding structure.

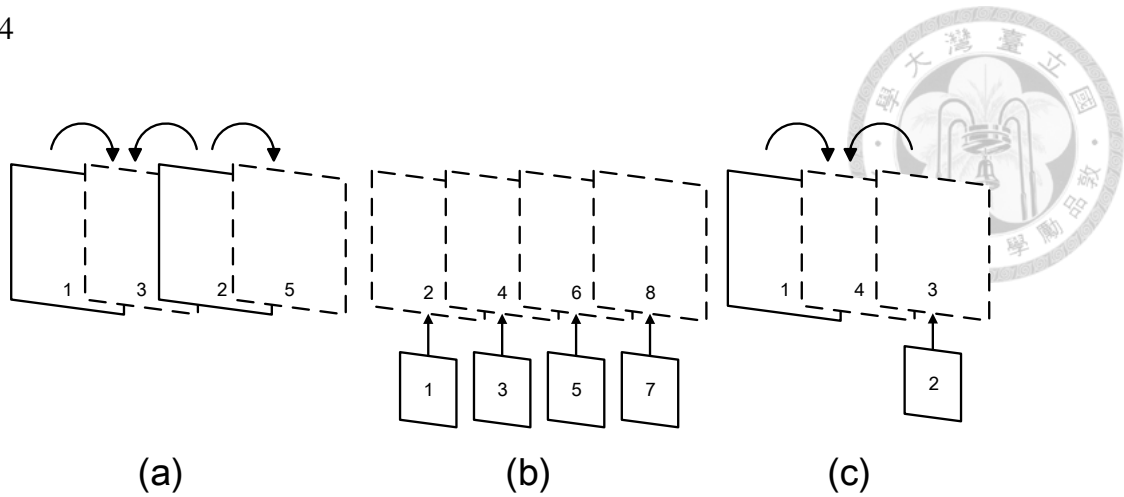


Figure 4.2: All possible coding structure in the proposed framework. The number labelled on frames denote decoding order. The arrows indicate the prediction directions while decoding. The solid frames are key frames, and the dotted frames are WZ frames.

## 4.3 Proposed Encoder

In this section, the proposed encoder with coding mode selection at different granularities will be introduced, from the frame level, the block level, toward the coefficient level.

### 4.3.1 Frame Level Coding Structure Selection (CSS)

The coding structure selection is developed to determine how a frame should be predicted, either spatially or temporally. It is the core block of the proposed system to integrate the previous temporal prediction scheme [4] and the proposed spatial prediction scheme into one framework to deal with all kinds of video-motion conditions. The block diagram of CSS is shown in Fig.4.3. First, the residue between consecutive original frames is obtained. Then the features of the residue are extracted, including statistical properties of the block sum of absolute difference (SAD) and the transformed band statistics, where mean, standard deviation, interquartile, minimum, maximum, and variance are included. There are total 4

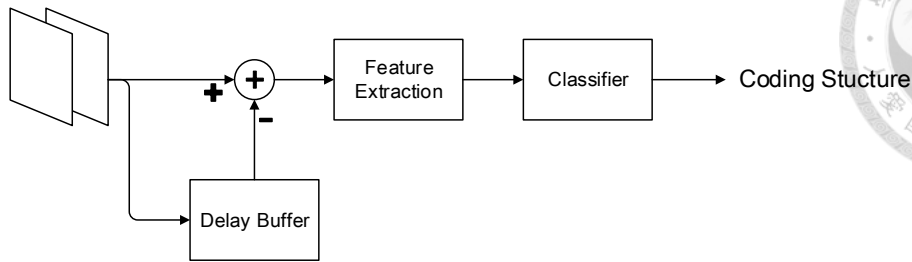



Figure 4.3: Block diagram of frame level coding structure selection.

different block sizes utilized to extract block SAD features, including 4x4, 8x8, 16x16, and 32x32. Feature dimension extracted from block-based statistics is 24 (4x6). As for the transformed bands statistics, since 4x4 DCT is used, the feature dimension based on transform bands will be 96 (16x6). Therefore, overall feature dimension used to test the input frame is 120. Finally, an offline-trained linear SVM classifier is used. Note that the simplest linear SVM model is employed since we believe that there do exist some underlying patterns of the temporal residue that allow the encoder to easily determine which kind of auxiliary information allows decoder to generate reliable SI. Besides, linear SVM classifier takes only a few time testing. Compared with [9] using a classifier at the coefficient level, the proposed framework utilizes classifier at the frame level, which mitigates the difficulty of accurate prediction of decoder behaviour.

### 4.3.2 Transform and Quantization

As soon as the prediction residue is obtained, 4x4 DCT is applied as in most DVC frameworks. The proposed framework adopts predefined quantization levels. Number of bits used to represent quantization outputs under different DVC configuration is listed in Table.4.1.

Given  $M$  bits to represent certain frequency band,  $b_i$ , the quantization step used for that band,  $Q_i$ , is computed with uniform quantization followed by thresholding.



4	3	0	0
3	0	0	0
0	0	0	0
0	0	0	0

 $Q_1$ 

5	3	0	0
3	0	0	0
0	0	0	0
0	0	0	0

 $Q_2$ 

5	3	2	0
3	2	0	0
2	0	0	0
0	0	0	0

 $Q_3$ 

5	4	3	2
4	3	2	0
3	2	0	0
2	0	0	0

 $Q_4$ 
  

5	4	3	2
4	3	2	2
3	2	2	0
2	2	0	0

 $Q_5$ 

6	4	3	3
4	3	3	2
3	3	2	0
3	3	0	0

 $Q_6$ 

6	5	4	3
5	4	3	2
4	3	2	0
3	2	0	0

 $Q_7$ 

7	6	5	4
6	5	4	3
5	4	3	0
4	3	0	0

 $Q_8$ 

Table 4.1: Modified quantization table from [5].

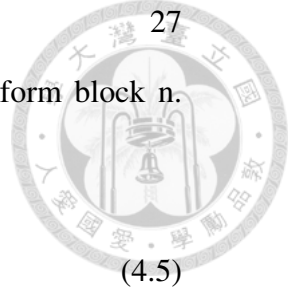
$$Q_i = \max \left( \frac{2|\max(b_i)|}{2^M - 1}, Q_{i,\min} \right), \quad (4.3)$$

Where  $Q_{i,\min}$  is the predefined minimum quantization step allowed at certain band.  $Q_{i,\min}$  aims to avoid too small quantization step to be used. Since the proposed framework adopts residual coding, range of  $2|\max(b_i)|$  is no longer as large as that in [3]. Therefore, some bands might result in very small quantization step, and almost no quantization is done. Besides, the smaller quantization steps are, the more accurate SI is required; otherwise, great amount of rate will be necessary to correct the SI at decoder and deteriorate overall compression performance drastically.

### 4.3.3 Block Level Skip Mode

After the quantization is done, in both cases when a frame is selected to be predicted spatially or temporally, a block-level skip mask is computed as the following way

$$Dist(n) = \sum_{u=0}^3 \sum_{v=0}^3 q_n^2(u, v), \quad (4.4)$$



where  $q_n(u, v)$  denotes the  $(u, v)$ -th quantized coefficient in transform block  $n$ . Skip mask is generated using a simple threshold  $\tau$  as follow

$$Mask(n) = \begin{cases} 1, & \text{if } Dist(n) < \tau, \\ 0, & \text{otherwise,} \end{cases} \quad (4.5)$$

The skip threshold is empirically set to 20 for temporally-predicted frames and 1 for spatially-predicted frames.  $\tau = 1$  means that any blocks containing non-zero quantized coefficients will not be skipped. Blocks with non-zero coefficients might involve object edges where the SI is less possible to be consistent with the original WZ frame. In addition, it also results in non-zero quantized coefficients when the key frames are encoded at low rate. In this case, blocks should not be skipped neither because the SI generated with a heavily distorted key frame is less reliable. After the skip mask is computed, it is compressed using run-length coding with pre-trained Huffman table.

#### 4.3.4 Coefficient Level Coding

For the coefficients within non-skipped blocks, the temporally-predicted residue is coded with both entropy coding and channel coding [4], where channel coding and entropy coding are employed for bands with high and low residue energy, respectively. First, bands are grouped as shown in Fig.4.4. Then sum absolute difference of each group is calculated as group energy. Then all the group energies are used to determine the way of coefficient level coding as described in Table.4.2.

As can be seen, if the lowest-banded residue energy,  $E_0$ , is smaller than a predefined threshold,  $\tau_0$ , transformed-quantized coefficients tend to be entropy coded. The entropy coding adopted is Context-based Adaptive Variable-Length (CAVLC) code. On the contrary, when all the group energies are high, encoder would consider SI being very unreliable and tend to correcting the SI rather than compensating SI with prediction residue.

On the other hand, the spatially-predicted residue is coded with entropy coding



0	0	1	2
0	1	2	2
1	2	2	2
2	2	2	2

Figure 4.4: Grouping of transform bands used for coefficient coding mode selection for temporally-predicted WZ frames.

Coding mode	Decision rule
Channel Coding	$E_0 > \tau_0, E_1 > \tau_1, E_2 > \tau_2$
Hybrid Mode 1	$E_0 > \tau_0, E_1 \leq \tau_1$
Hybrid Mode 2	$E_0 > \tau_0, E_1 > \tau_1, E_2 \leq \tau_2$
Entropy Coding	$E_0 \leq \tau_0$

Table 4.2: Coding modes and decision rules.

only. As mentioned above, channel coding is usually more effective when the residue between the SI and the WZ frame is larger since it has the ability to correct errors. When the frame is determined to be encoded/decoded with the spatially-predicted SI, it is assured that the spatially-predicted SI is good enough, just as entropy coding mode of temporally predicted WZ frame. Therefore, channel code is not used in this case.

## 4.4 Proposed Decoder

As shown in Fig.4.1, if the temporally-predicted side information is selected, the decoder architecture is similar to the ones in [3] [4]. The decoding scheme is the same as introduced in chapter 2. The modified decoder block in Fig.4.1 is the temporal/spatial SI generation. When the spatial prediction is selected, a single-image-based super-resolution algorithm is adopted for SI generation, where SR-CNN [22], one of the state-of-the-art works, is employed for the evaluation of

the proposed framework due to its extraordinary performance and fast computing speed.

Note that other super-resolution algorithms can also be used. As mentioned in Section 4.3.4, the WZ frames are reconstructed using SI and the entropy decoded residue. The other thing worth mentioning is that if the case of Fig.4.2(c) takes place, the temporally-predicted WZ frame will not be decoded until the next full-resolution spatially-predicted WZ frame is decoded. The number labelled in Fig.4.2 denotes the decoding order.







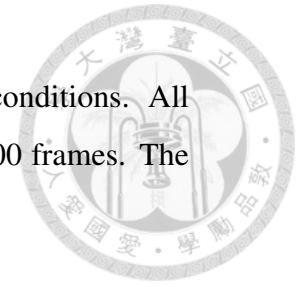
## Chapter 5

# Experimental Results of the Proposed Framework

In this chapter, full evaluation of the proposed framework is presented. First, the potential of spatially-predicted DVC is explored. Coding tools that are shown in Fig.4.1 are evaluated progressively and compared with different anchors to verify improvement being achieved. After the configuration of the spatially-predicted DVC is established with regards to the coding tools, the performance is compared to that of temporally-predicted DVC [4]. Then, labels of CSS could be obtained and evaluation of CSS classifier is conducted. Finally, overall compression performance and encoding complexity of the proposed framework are listed and compared with some common DVC anchors.

Because of learning-based frame-level solution, testing and training set are required. However, since there is no public data set for DVC, we collect several video sequences of IoT applications, including surveillance sequences and first-person view sequences as training set for the SVM classifier of CSS. As for the testing set, sequences that are widely used for evaluating DVC are adopted in convenience of performance comparison and verification, including Hall, Foreman, Coastguard, and Soccer listed in [5]. In addition to these commonly used sequences, we also add the other two sequences recognized as high-motion videos

to verify the proposed framework under different video-motion conditions. All these sequences are in CIF (352x288) resolution and consist of 300 frames. The H.264 reference software JM is used as key frame coder.



## 5.1 Potential of Spatially-Predicted DVC

In this section, performance of the spatially-predicted coding flow shown in Fig.4.1 is explored step by step. The configuration of the spatial prediction is derived from observation of training sequences, and is used to test over the target high-motion sequences among testing set. Therefore, only one testing result will be shown to verify the configuration. The evaluation begins with verification whether residual coding benefits compression performance or not. Secondly, evaluation of different thresholds used in block level skip mode is presented to determine suitable threshold. Finally, coefficient level evaluation is presented to establish how the prediction residue should be coded.

### 5.1.1 Residual Coding

Effect of residual coding of spatial prediction is evaluated anchoring non-residual code flow of spatially-predicted DVC. Besides, all the transformed-quantized coefficients are all channel coded. The quantization levels of non-residual coding flow are the same as [3], and that of residual coding flow follows the way used in [4]. In addition, in order to prevent error in the CN model which influence decoding rate a lot, both flows are evaluated with the exact CN model that is off-line fitted. The coding structure of both flow is the same as shown in Fig.4.2(b). BD rate [23] is used to evaluate whether compression performance gain or loss is obtained using residual coding for spatial prediction. One testing result is shown in Fig.5.1.

As can be seen, negative BD rate is obtained in almost all frames, which means compression gain is obtained. Therefore, residual coding is adopted for spatially-

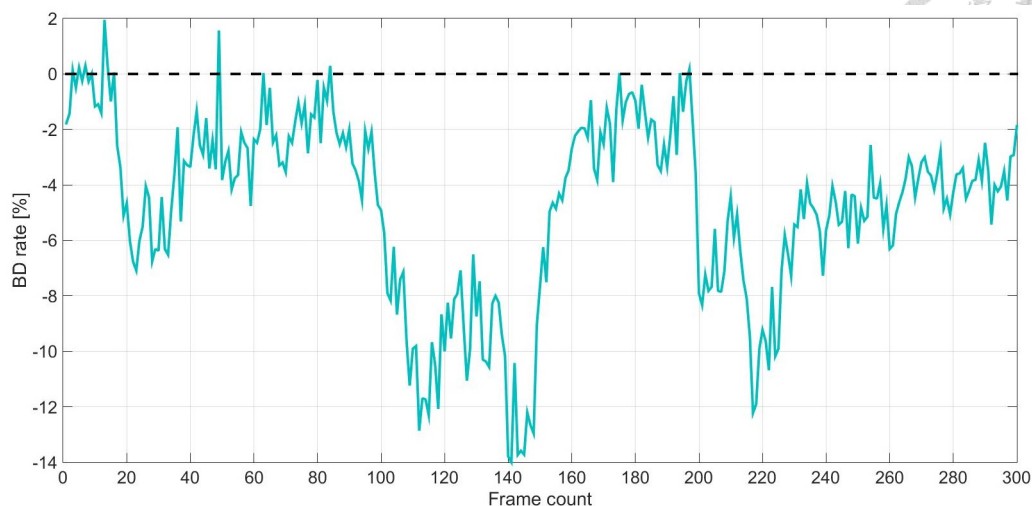


Figure 5.1: BD rate of residual coding of one testing sequence of spatially-predicted DVC anchoring non-residual coded spatially-predicted DVC.

predicted frames.

### 5.1.2 Block Level Skip Mode

Since the residual coding is adopted for spatially-predicted WZ frames, prediction residue is then transformed and quantized. Then block-based skip mode is applied. As introduced in 4.3.3, square sum of the block is computed and compared to a predefined threshold. Experiments are conducted to determine the threshold. With the idea that prediction residue only takes place at edges or boundaries, only low thresholds ranging from  $1^2$  to  $5^2$  are considered. Note that for coefficients within non-skipped blocks are all channel coded here since we have not established coefficient level coding method here. The anchor used in this evaluation is spatially-predicted WZ frames with residual coding only, and without use of skip mode. One of the result among testing set is shown in Fig.5.2.

As can be seen in Fig.5.2, thresholds of  $1^2$  to  $4^2$  behaves similarly and have similar BD rate reduction; nevertheless, overly skipped ratio with threshold  $5^2$  reveals unstable performance according to sequence contents. Due to the fact that encoders in DVC frameworks are less possible to know the prediction made at de-

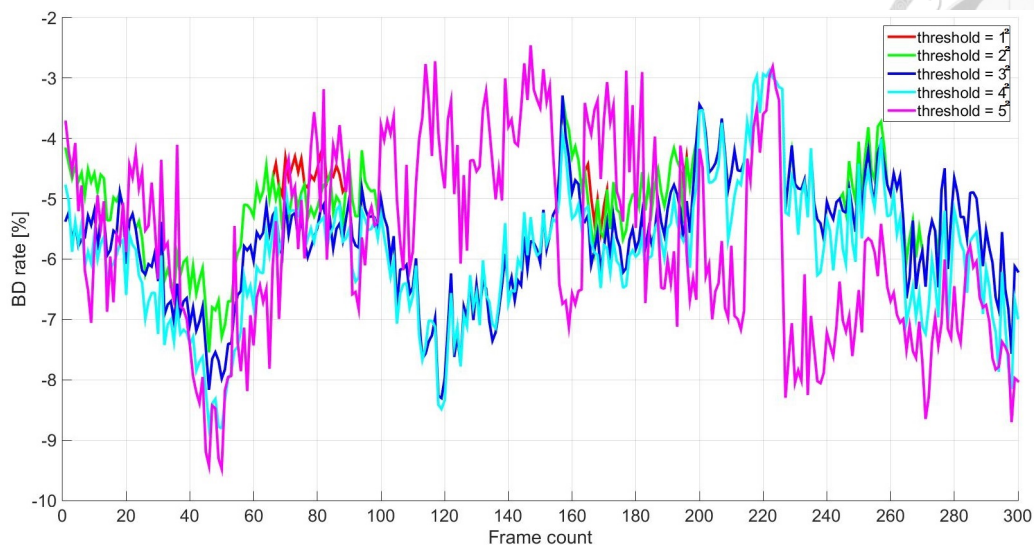


Figure 5.2: BD rate of different skip mode thresholds of one testing sequence anchoring non-skipped spatially-predicted DVC with residual coding.

coder exactly, it is more reasonable to make conservative choice of skip threshold. Based on this inspiration, the threshold is set to 1, the tightest one. This configuration makes sense because that after spatially-predicted residue is quantized, non-zero coefficients only take place at edges or objects boundaries where decoder is less likely to compute accurately based on the downsampled key frames received, and should not be skipped.

### 5.1.3 Coefficient Level Coding

For non-skipped block coefficients, different ways of coding are evaluated in this subsection. In order to explore potential RD space to the utmost, as experiment condition used above, CN modeling is done off-line. Off-line fitted CN model stands for the best accuracy of the noise model. Any online modelling technique, such as [?], is approximation of off-line model using decoder available information only. Besides, the transformed-quantized residue coefficients are grouped at finer resolution compared to Fig.4.4 as shown in Fig.5.3. Every band group is coded using either channel coding or entropy coding and different combinations

0	1	2	3
1	2	3	4
2	3	4	5
3	4	5	6

Figure 5.3: Grouping of frequency bands used in coefficient level coding experiment for spatially-predicted DVC.

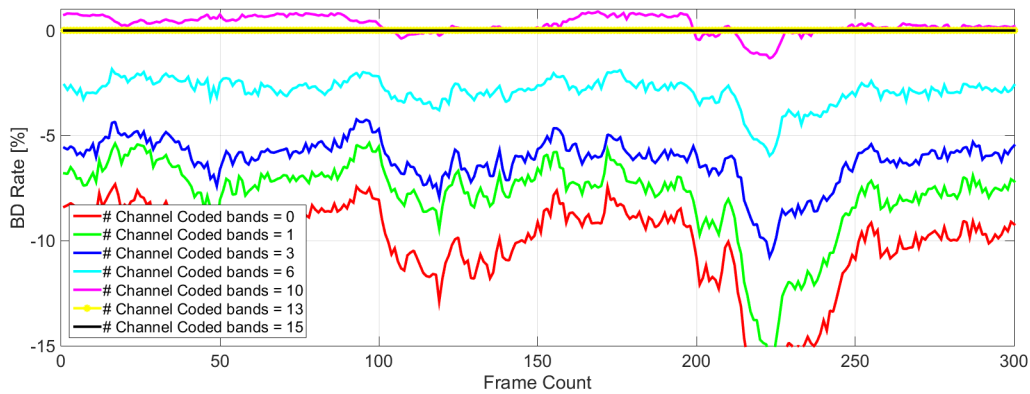


Figure 5.4: BD rate of different coefficients level coding configurations of one training sequence anchoring all groups being channel coded.

are tested. Fig.5.4 shows the BD rate of one sequence in the training set with the anchor of all coefficients coded with channel coding, with all frames being inter-layer predicted, and with skip threshold established as in section 5.1.2. The best BD rate reduction of spatial prediction is obtained when all groups are entropy coded, and there is no exception among sequences tested. Therefore, for spatially-predicted WZ frames, the prediction residues are all entropy coded. Fig.5.5-5.7 peek at performance of the configuration with same anchor described above of 3 widely used testing sequences.

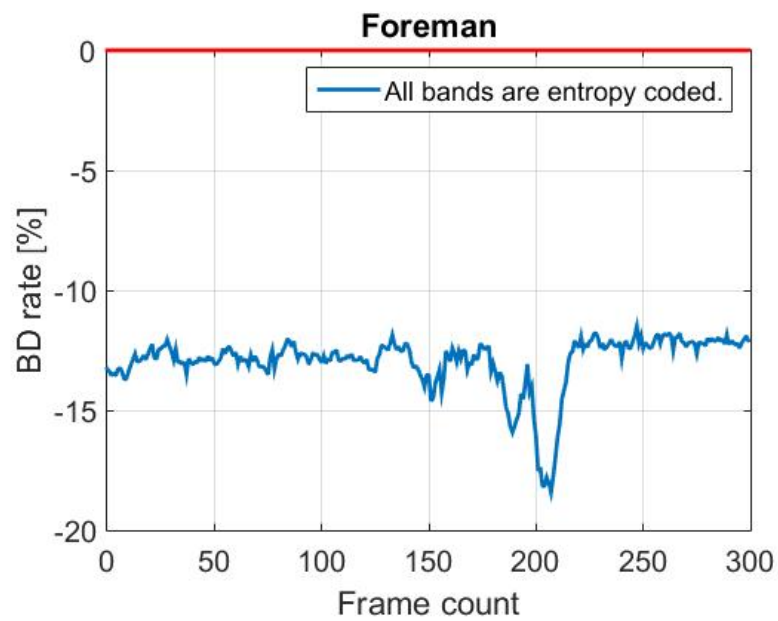


Figure 5.5: BD rate curve of testing sequence Foreman anchoring all groups being channel coded.

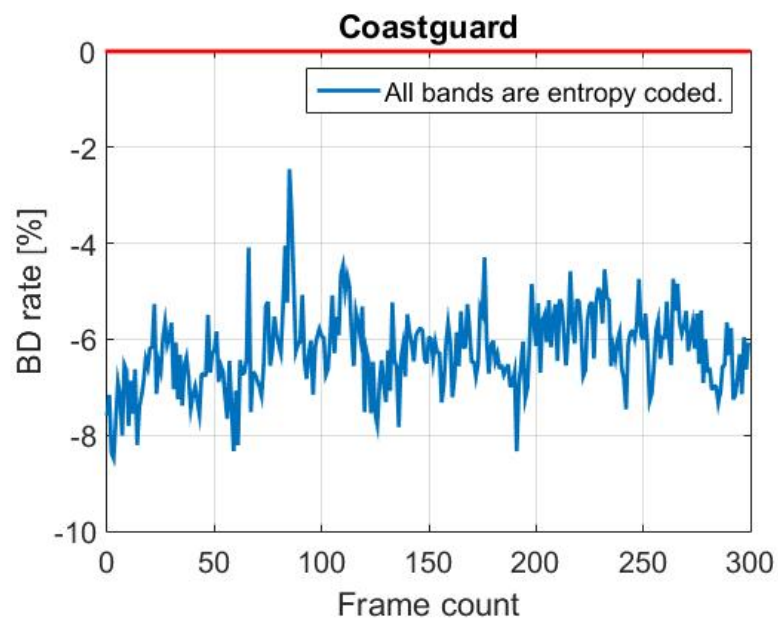


Figure 5.6: BD rate curve of testing sequence Coastguard anchoring all groups being channel coded.

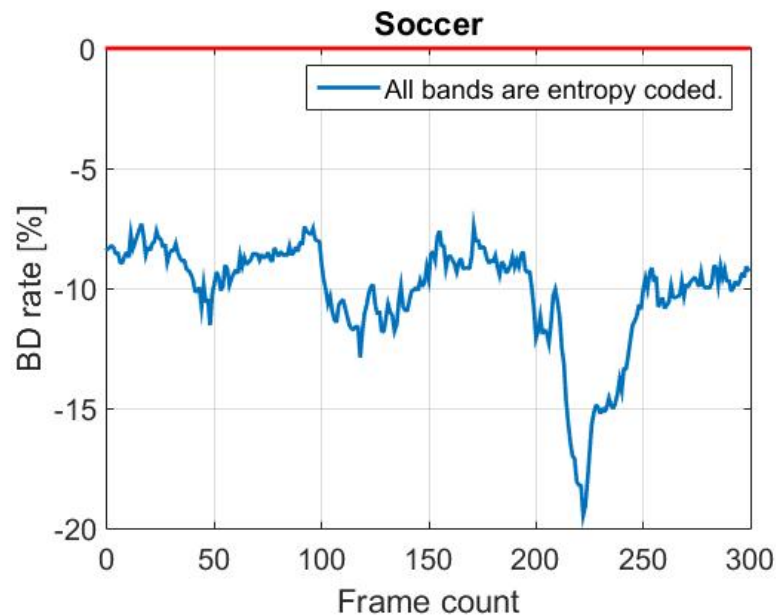


Figure 5.7: BD rate curve of testing sequence Soccer anchoring all groups being channel coded.

## 5.2 CSS Classifier Training and Testing

So far, configuration of spatial prediction has been established. Therefore, labels on each frame to be predicted either temporally or spatially could be set. Labels are generated as the following way. First, the RD performance with GOP size 2 from [4] is generated. It requires mentioning that GOP size 2 in [4] consists of 1 key frame and 1 WZ frame, and the rate differs significantly. Since we aim to label every frame in order to compare with spatial prediction framework, 2 frames share the same RD performance that is average of the key frame and WZ frame. Second, BD rates of spatial prediction are computed with anchors generated in the first step. The BD rate curves of one testing sequence are shown in Fig.5.8-5.13. Frames with negative BD rate are labeled as spatial prediction, while frames with positive BD rate are labeled as temporal prediction. With these labels, a linear support-vector-machine (SVM) classifier is trained. During training process, classification error is weighted on BD rate to prevent great compression performance loss. Finally, the CSS classifier is tested with the testing set labels. Testing

accuracy is listed in Table.5.1.

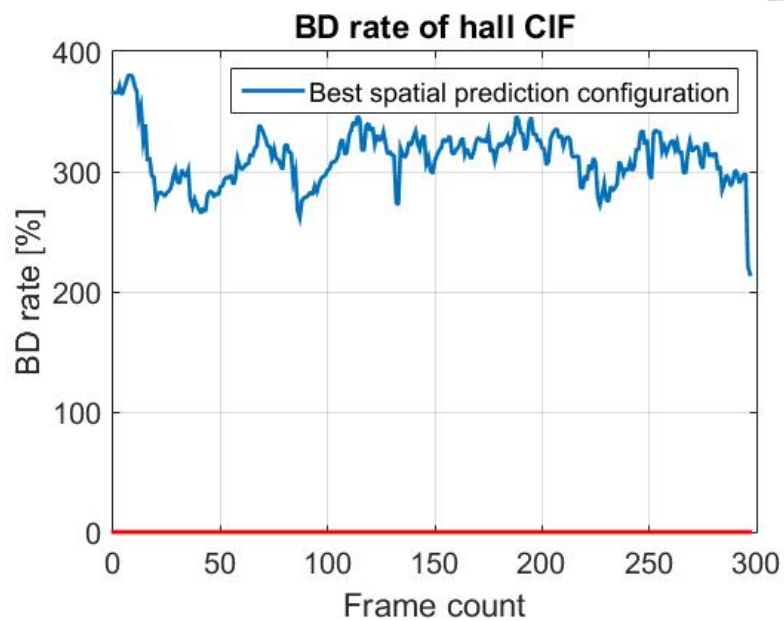


Figure 5.8: BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework.

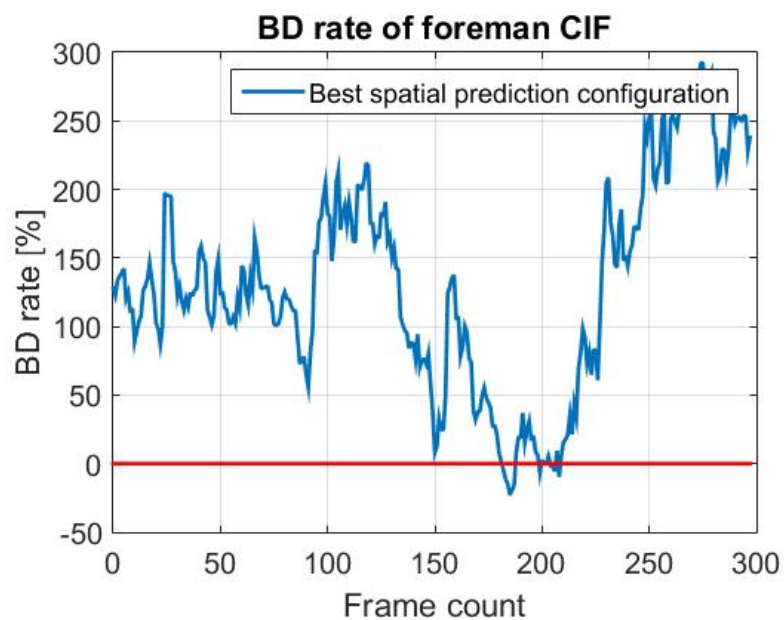


Figure 5.9: BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework.



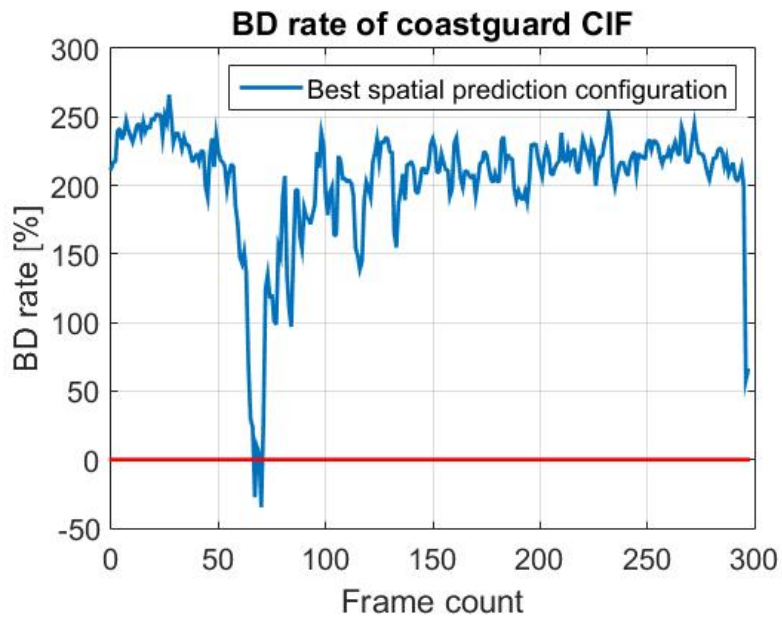


Figure 5.10: BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework.

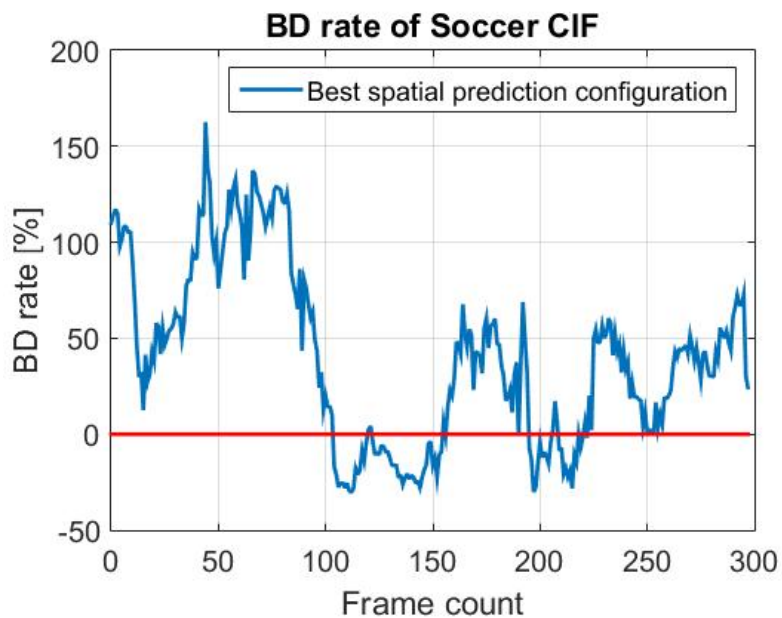


Figure 5.11: BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework.

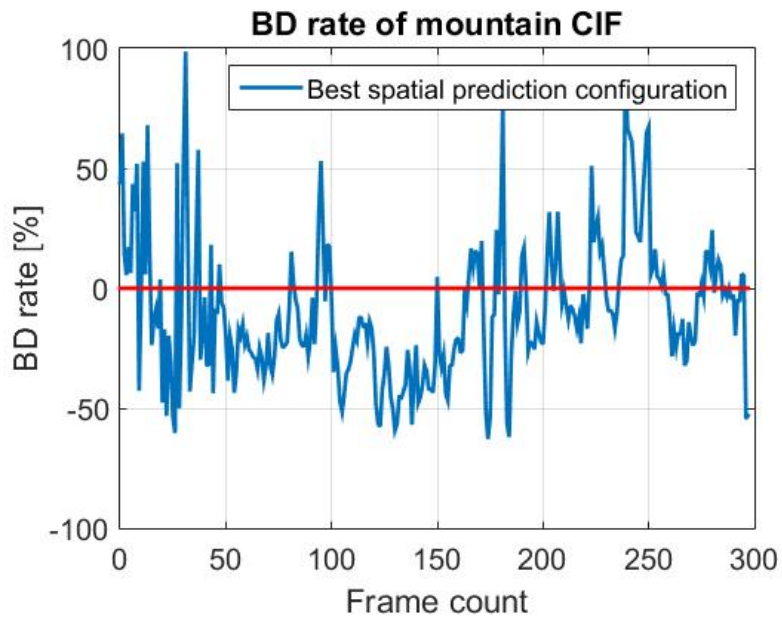


Figure 5.12: BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework.

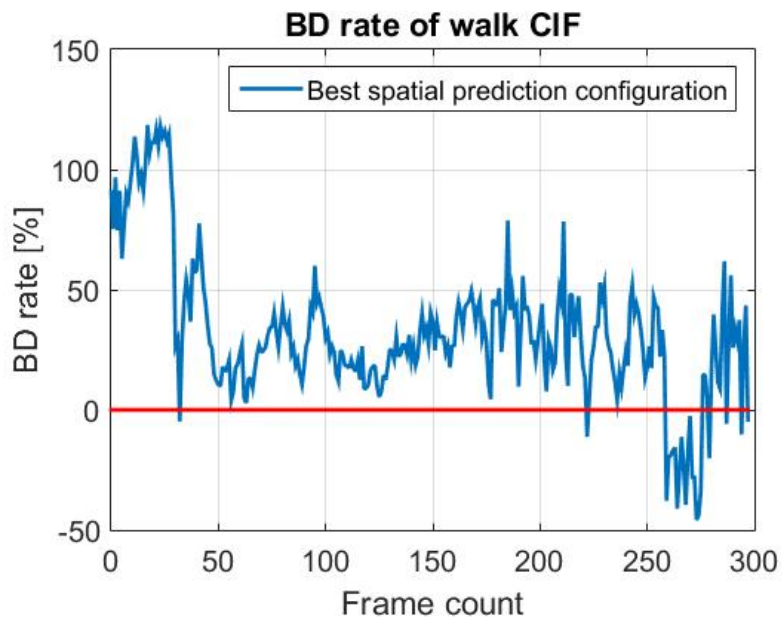


Figure 5.13: BD rate of spatially-predicted DVC anchoring [4] that is foundation of the proposed framework.



Table 5.1: CSS classifier testing accuracy

	CSS accuracy
Hall	100 %
Foreman	97 %
Coastguard	99.33 %
Soccer	96.33 %
Mountain	92.33 %
Walk	95.33 %
Average	96.72 %

### 5.3 Overall Performance of the Proposed Framework

In this section, overall performance of the proposed framework is evaluated using the testing set and compared to some common anchors used in DVC evaluation. Furthermore, encoding complexity of the proposed framework is also compared to [3] [4] that are foundations of the proposed framework.

#### 5.3.1 RD Performance and BD Rate

Fig.5.14 - Fig.5.19 shows the RD performance of anchors and the proposed framework using testing sequences. Recall that the proposed framework targets to deal with high-motion video contents. Such kinds of sequences could be located by viewing performance gap between intra coding and zero motion inter coding of conventional video codec anchors. As long as inter coding outperforms intra coding a lot, the sequence mainly consists of stationary contents; on the other hand, while intra coding outperforms inter coding, the sequence is comprised of extremely high-motion contents. The latter case is just as shown in Fig.5.17 - Fig.5.19 where distinguishable performance gain is obtained with the proposed framework.

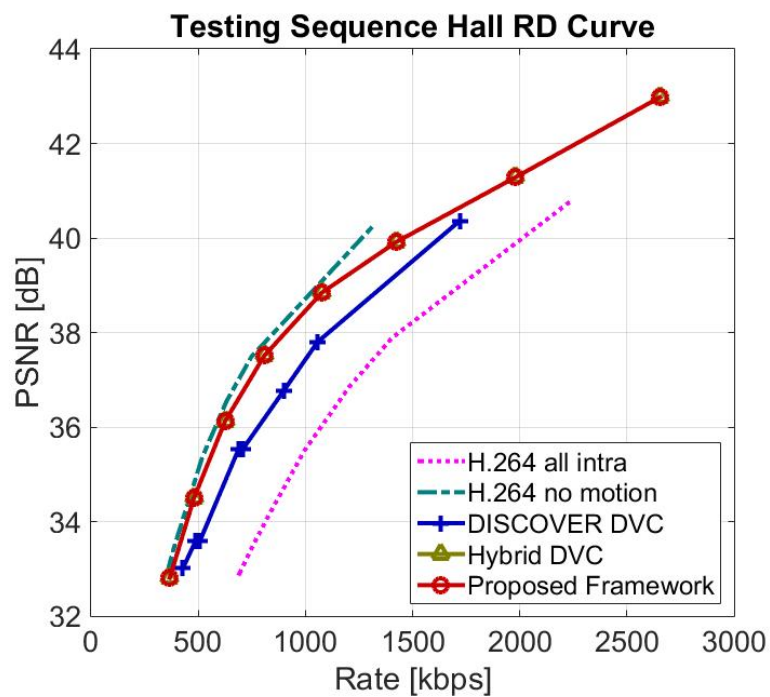


Figure 5.14: RD curve of testing sequence Hall with common anchors.

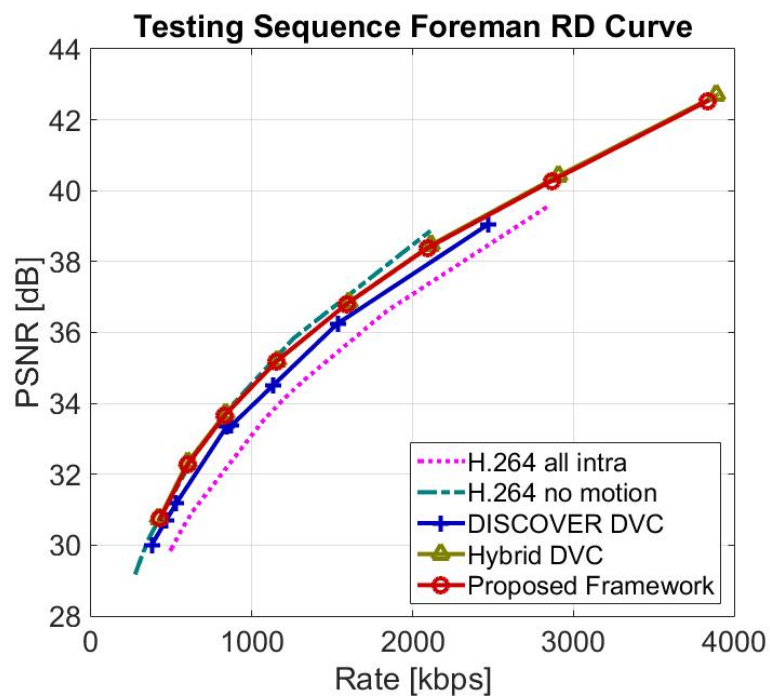


Figure 5.15: RD curve of testing sequence Foreman with common anchors.

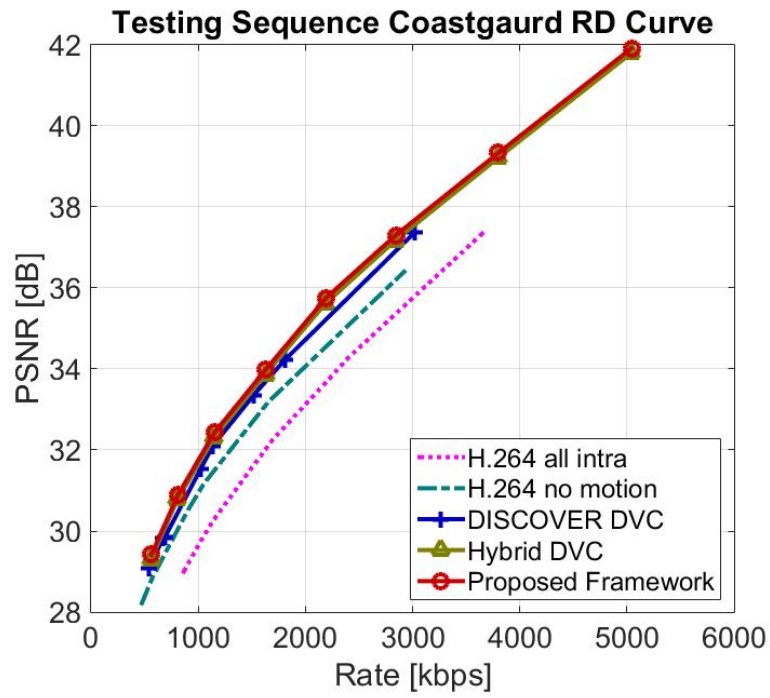


Figure 5.16: RD curve of testing sequence Coastguard with common anchors.

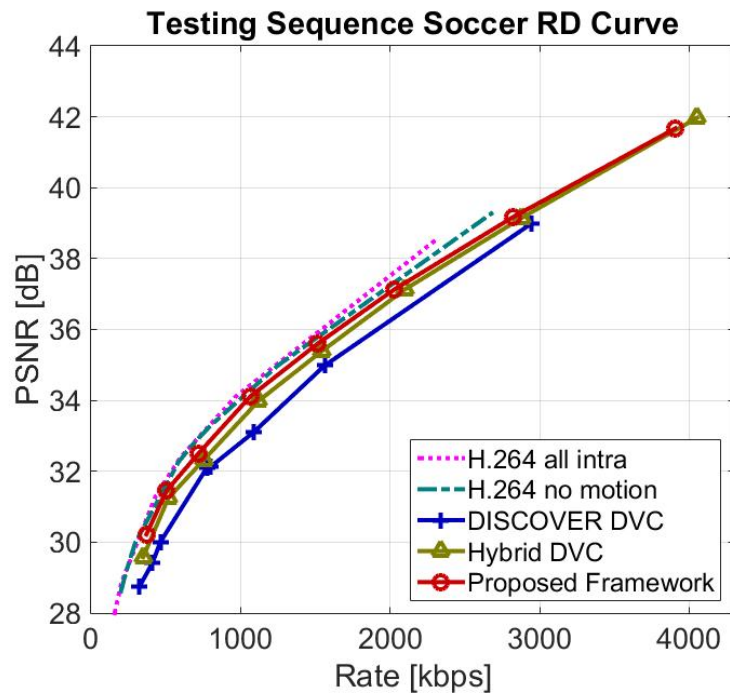


Figure 5.17: RD curve of testing sequence Soccer with common anchors.

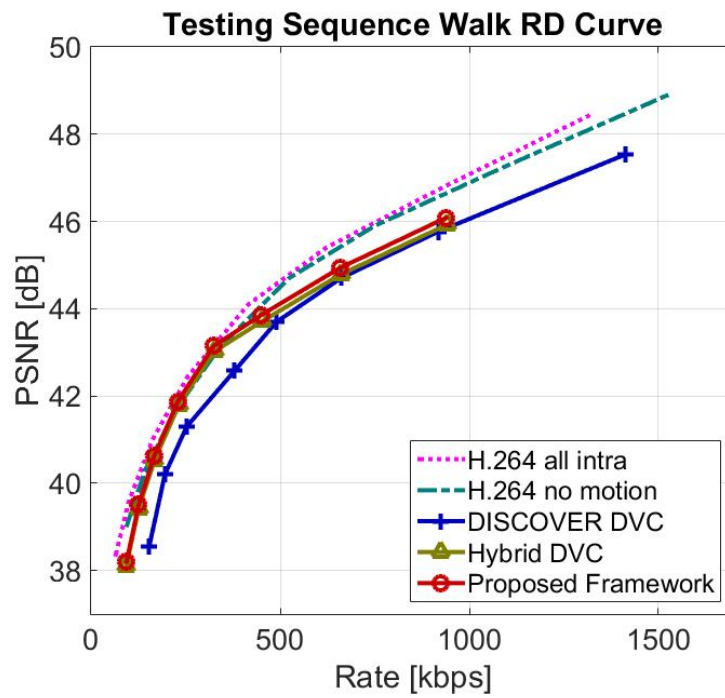


Figure 5.18: RD curve of testing sequence Walk with common anchors.

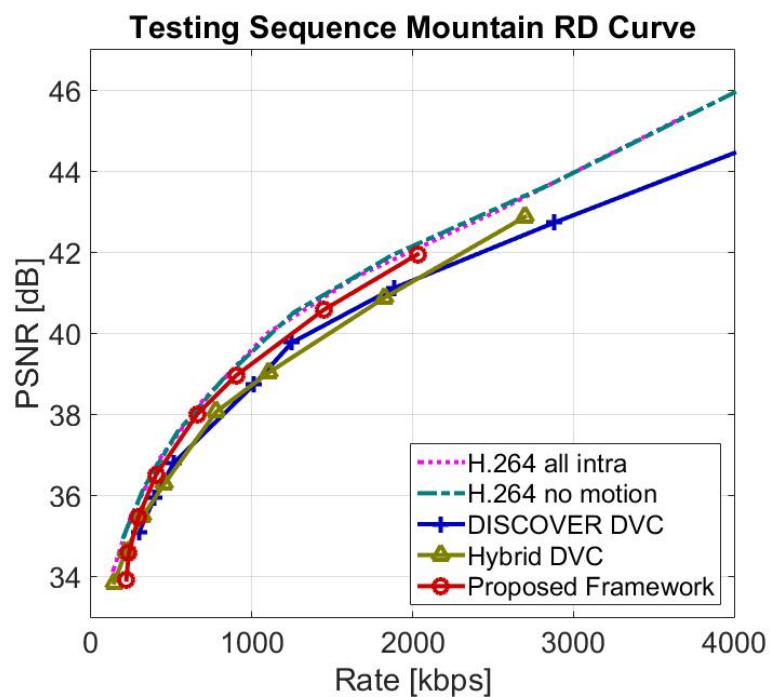


Figure 5.19: RD curve of testing sequence Mountain with common anchors.

In addition to the RD performance, BD rates with different anchors are shown in Table.5.2. As observed in RD curves, there is almost no performance loss with low-motion contents and observable improvements on high-motion sequences anchoring [4].

Table 5.2: BD rate of the proposed framework with different anchors.

	BD rate w/ anchor [4]	BD rate w/ anchor [3]
Hall	0 %	- 17.14 %
Foreman	+ 0.49 %	- 8.53 %
Coastguard	- 1.93 %	- 6.46 %
Soccer	- 5.45 %	- 17.4 %
Mountain	- 12.07 %	- 9.9 %
Walk	- 3.82 %	- 18.17 %
Average	- 3.8 %	- 12.93 %

### 5.3.2 Running Time Evaluation

The running time of the proposed framework with different testing sequences are provided in Table.5.3,5.4, where the total running time for encoding and decoding the whole sequence are shown respectively. The computational overhead of CSS is revealed Table.5.3 in the testing sequence Hall since all frames are classified to be temporally predicted and is only about 1.1% of [3] and 0.3% of [4]. It also shows that when more frames are spatially-predicted, less encoding running time is required. Profiling on the running time tells that the reduction comes from that encoding two low-resolution key frames is faster than encoding one full-resolution key frame, and the acceleration could compensate the overhead of CSS in most cases. The average encoding time among the testing set is 92.26% of the DISCOVER DVC and 91.6% of the hybrid DVC.

Table.5.4 shows the decoding running time using the highest quality configuration, namely largest quantization output levels as  $Q_8$  described in Table.4.1. As can be seen, decoding time required is significantly longer than that of encoding,

which DVC features. It requires mentioning the drastic reduction in decoding time between DISCOVER DVC and Hybrid DVC comes from utilization of entropy coding in hybrid DVC. In hybrid DVC, low motion sequences are more entropy coded than channel coded. Channel decoding, as far as the LDPCA code is concerned, is quite time-consuming. To obtain the minimum rate required by LDPCA, the channel decoding must start from the lowest rate of parity, followed by a large burden of belief propagation, and finally test correctness of convergence using CRC code. If it converges to a wrong result, further parity bits will be asked, and again conduct the process introduced above. This circumstance is extremely obvious in testing sequence Hall. There are still some channel coding in hybrid DVC for coefficient bands with large residue energy, which take place more often in high motion video sequences. This case is more likely to be encoded with inter-layer prediction and thus entropy coded in the proposed framework. The larger the motion is, the scenario happens more often. In other words, greater decoding time reduction could be obtained, comparing the proposed framework with hybrid DVC, when more frames label to be inter-prediction. It is suggested to refer to Fig.5.8-5.13. The exception lies in decoding testing sequence Mountain, where decoding time increment is obtained. This is because, when encoding with hybrid DVC, more bands that should have been channel coded are being entropy coded. This is a case that hybrid DVC fails and the failure could be obtained in RD curve Fig.5.19, in which performance loss of hybrid DVC when compared with DISCOVER.

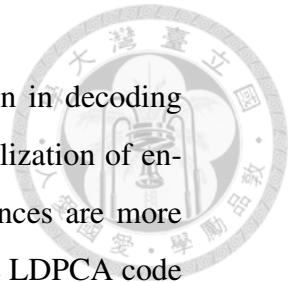






Table 5.3: Encoding time comparison.

	Proposed framework [sec]	Hybrid DVC [4] [sec]	DISCOVER DVC [3] [sec]
Hall	44.25	44.13	43.75
Foreman	48.17	48.79	48.53
Coastguard	57.87	57.28	56.92
Soccer	43.28	50.13	49.71
Mountain	23.4	41.21	40.83
Walk	36.21	34.85	34.72
Average	42.2	46.07	45.74
	-	91.6%	92.26%

Table 5.4: Decoding time comparison.

	Proposed framework [sec]	Hybrid DVC [4] [sec]	DISCOVER DVC [3] [sec]
Hall	53	53	4142
Foreman	3982	4647	10559
Coastguard	4726	5255	12524
Soccer	4467	6923	20301
Mountain	351	347	3151
Walk	1992	4936	32138
Average	2595	3693	13802
	-	70.26%	18.8%





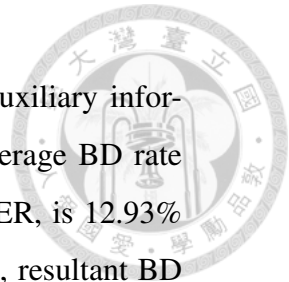
## Chapter 6

# Conclusion

In this thesis, a DVC framework with spatial prediction and temporal prediction is proposed. The major contribution of the proposed framework is improving DVC compression performance by integrating spatial prediction coding flow with coding tools at different granularities into previous DVC framework. These coding tools include frame level prediction selection, block level skip mode determination, and coefficient level coding mode selection. All of the previously proposed DVC frameworks consist of temporal prediction only, and will result in larger compression performance gap compared with conventional coding anchors when the input sequence contains high-motion content or abrupt changes. The failure is aroused from lack of temporal correlation on which motion compensated frame interpolation algorithms rely. The proposed framework successfully addresses this problem by utilizing single-image super-resolution algorithm to compute SI based on downsampled key frames in the spatial prediction scheme.

The most important issue in DVC framework design is to what extent encoder knows SI computed at decoder without complex computing. By analyzing the success of previous DVC frameworks with low motion or stationary contents, gain is obtained majorly from skipped block at encoder and matched good quality SI at decoder. In other words, encoder knows where SI is good enough and does not require correcting. The proposed framework deals with the issue using

frame level coding structure selection by transmitting different auxiliary information that allow decoder to generate better SI. The resultant average BD rate reduction anchoring the most popular DVC framework, DISCOVER, is 12.93% over the testing set but with only 92.3% of running time. Besides, resultant BD rate reduction anchoring hybrid DVC, foundation of the proposed framework, is 3.8% where gain is obtained from high-motion input sequences. To conclude, the proposed framework improves on the shortage of previous DVC frameworks and successfully solves the target problem.





## Reference

- [1] G. J. Sullivan, J. R. Ohm, Woo-Jin Han, and T. Wiegand, “Overview of the high efficiency video coding (hevc) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.
- [2] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources,” *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, Jul 1973.
- [3] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, “The DISCOVER codec: Architecture, techniques and evaluation,” in *Proc. Picture Coding Symposium (PCS’07)*, Nov. 2007.
- [4] C.-C. Chiu, S.-Y. Chien, C.-H. Lee, V S. Somayazulu, and Y.-K. Chen, “Hybrid distributed video coding with frame level coding mode selection,” in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1561–1564.
- [5] “<http://www.img.lx.it.pt/~discover/home.html>,” .
- [6] F. Saab, I. H. Elhajj, A. Kayssi, and A. Chehab, “Profiling of hevc encoder,” *Electronics Letters*, vol. 50, no. 15, pp. 1061–1063, July 2014.
- [7] A. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, Jan 1976.



- [8] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.
- [9] R. Puri, A. Majumdar, and K. Ramchandran, "PRISM: A video coding paradigm with motion estimation at the decoder," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2436–2448, Oct. 2007.
- [10] R. J. McEliece, D. J. C. MacKay, and Jung-Fu Cheng, "Turbo decoding as an instance of pearl's belief propagation algorithm," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 2, pp. 140–152, Feb 1998.
- [11] Y. Kou, S. Lin, and M. P.C. Fossorier, "Low-density parity-check codes based on finite geometries: A rediscovery and new results," *IEEE Trans. Inf. Theor.*, vol. 47, no. 7, pp. 2711–2736, Sept. 2006.
- [12] David Varodayan, Anne Aaron, and Bernd Girod, "Rate-adaptive codes for distributed source coding," *Signal Process.*, vol. 86, no. 11, pp. 3123–3130, Nov. 2006.
- [13] C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain wyner-ziv video coding," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 18, no. 9, pp. 1177–1190, Sept. 2008.
- [14] D. Kubasov, J. Nayak, and C. Guillemot, "Optimal reconstruction in wyner-ziv video coding with multiple side information," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, Oct 2007, pp. 183–186.
- [15] A. Norkin, G. Bjontegaard, A. Fuldseth, M. Narroschke, M. Ikeda, K. Andersson, Minhua Zhou, and G. Van der Auwera, "Hvc deblocking filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1746–1754, Dec 2012.



- [16] R. Martins, C. Brites, J. Ascenso, and F. Pereira, “Adaptive deblocking filter for transform domain Wyner-Ziv video coding,” *IET Image Process.*, vol. 3, no. 6, pp. 315–328, Dec. 2009.
- [17] T. Sheng, G. Hua, H. Guo, J. Zhou, and C. W. Chen, “Rate allocation for transform domain Wyner-Ziv video coding without feedback,” in *Proc. 16th ACM International Conference on Multimedia*, 2008, pp. 701–704.
- [18] B. Macchiavello, D. Mukherjee, and R. L. de Queiroz, “Iterative side-information generation in a mixed resolution wyner-ziv framework,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1409–1423, Oct 2009.
- [19] J. Slowack, S. Mys, J. korupa, P. Lambert, R. Van de Walle, N. Deligiannis, and A. Munteanu, “Bitplane intra coding with decoder-side mode decision in distributed video coding,” in *2010 IEEE International Conference on Image Processing*, Sept 2010, pp. 3733–3736.
- [20] Stefaan Mys, Jürgen Slowack, Jozef Škorupa, Nikos Deligiannis, Peter Lambert, Adrian Munteanu, and Rik Van de Walle, “Decoder-driven mode decision in a block-based distributed video codec,” *Multimedia Tools and Applications*, vol. 58, no. 1, pp. 239–266, 2012.
- [21] G. Correa, P. Assuncao, L. Agostini, and L. A. da Silva Cruz, “Four-step algorithm for early termination in hevc inter-frame prediction based on decision trees,” in *Visual Communications and Image Processing Conference, 2014 IEEE*, Dec 2014, pp. 65–68.
- [22] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, “Learning a deep convolutional network for image super-resolution,” in *Computer Vision–ECCV 2014*, pp. 184–199. Springer, 2014.
- [23] “Bjontegaard delta bitrate, video coding expert group (vceg) document vceg-m33,” .