Error Resilience and Concealment Techniques for High Efficiency Video Coding

João Filipe Monteiro Carreira

A Doctoral Thesis

Submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy of Loughborough University in London

*Dedicated to my wife, **Bianca** and my parents, **António** and **Maria***

*Anyone who has never
made a mistake has never
tried anything new.*

(Albert Einstein)

Acknowledgments

Firstly, I would like to express my gratitude to my advisers, Prof. Pedro António Amado Assunção, Prof. Sérgio Manuel Maciel de Faria, Dr Erhan Ekmekcioglu and Prof. Ahmet Kondoz for the tireless support of my PhD studies and related research, for their motivation, patience, and immense knowledge. I am thankful for their guidance and availability that were essential for the proper conduct of this research work. I would also like to acknowledge Dr. Varuna de Silva for the initial support on this research work.

I would like to acknowledge Fundação para a Ciência e a Tecnologia (FCT), which provided the financial support for this PhD, under the grant SFRH/BD/86368/2012. I am also grateful to the Instituto de Telecomunicações and to the Institute for Digital Technologies of Loughborough University London, which provided the computer resources and laboratory facilities with excellent conditions for my research work.

I would like to thank my research colleagues, both in Portugal and the United Kingdom, for the friendship and the wonderful work environment, for sharing new ideas, and also for the moments outside the laboratory. I also thank all my friends, that accompanied me during this journey, for the distracting and social moments.

I must thank all communities around the world that contribute for the development of many free and open-source tools used in this research work, namely the GNU/Linux, KDE software, Ubuntu and Gentoo communities. I also acknowledge the help of Luís Lucas and João Santos in the Calyp software, Remoterun scripts and general computer management, which improved our developing capabilities and ease the experimental evaluations.

A special thanks to my parents, António and Maria, and my sister Marta to whom I owe all that I have become. Also, I must acknowledge my uncles and cousins living in the United Kingdom for the kind welcome into their home, which significantly ease my research experience abroad. Last but not least, I also would like to express a special gratitude to Bianca Pires for her unconditional support and love during this journey.

Abstract

This thesis investigates the problem of robust coding and error concealment in High Efficiency Video Coding (HEVC). After a review of the current state of the art, a simulation study about error robustness, revealed that the HEVC has weak protection against network losses with significant impact on video quality degradation. Based on this evidence, the first contribution of this work is a new method to reduce the temporal dependencies between motion vectors, by improving the decoded video quality without compromising the compression efficiency.

The second contribution of this thesis is a two-stage approach for reducing the mismatch of temporal predictions in case of video streams received with errors or lost data. At the encoding stage, the reference pictures are dynamically distributed based on a constrained Lagrangian rate-distortion optimization to reduce the number of predictions from a single reference. At the streaming stage, a prioritization algorithm, based on spatial dependencies, selects a reduced set of motion vectors to be transmitted, as side information, to reduce mismatched motion predictions at the decoder.

The problem of error concealment-aware video coding is also investigated to enhance the overall error robustness. A new approach based on scalable coding and optimally error concealment selection is proposed, where the optimal error concealment modes are found by simulating transmission losses, followed by a saliency-weighted optimisation. Moreover, recovery residual information is encoded using a rate-controlled enhancement layer. Both are transmitted to the decoder to be used in case of data loss.

Finally, an adaptive error resilience scheme is proposed to dynamically predict the video stream that achieves the highest decoded quality for a particular loss case. A neural network selects among the various video streams, encoded with different levels of compression efficiency and error protection, based on information from the video signal, the coded stream and the transmission network.

Overall, the new robust video coding methods investigated in this thesis yield consistent quality gains in comparison with other existing methods and also the ones implemented in the HEVC reference software. Furthermore, the trade-off between coding efficiency and error robustness is also better in the proposed methods.

Contents

List of Figures

List of Tables

List of Abbreviations

AMVP	Advanced Motion Vector Prediction 24
BLA	Broken Link Access 14
BMA	Boundary Matching Algorithms 53
CABAC	Context Adaptive Binary Arithmetic Coding 11
CNN	Convolutional Neural Network 38
CRA	Clean Random Access 14
CTU	Coding Tree Unit 10
CU	Coding Unit 10
DBF	Deblocking Filter 30
DCT	Discrete Cosine Transform 29
DPB	Decoded Picture Buffer 26
DST	Discrete Sine Transform 29
EC	Error Concealment 3
ECU	Error Concealment Units 112
EL	Enhancement Layer 49
FC	Frame-Copy 53
FEC	Forward Error Correction 44
FMO	Flexible Macroblock Ordering 11
FSIM	Feature Similarity Index 122
GOP	Group of Pictures 42

H.264/AVC	H.264 - Advanced Video Coding 1
HEVC	High Efficiency Video Coding 1
HM	HEVC Test Model 25
HMVE	Hybrid Motion Vector Extrapolation 56
IDR	Instantaneous Decoding Refresh 14
IEC	International Electrotechnical Commission 1
IP	Internet Protocol 17
ISO	International Organization for Standardization 1
ITU-T	International Telecommunication Union - Telecommunication 1
JCT-VC	Joint Collaborative Team on Video Coding 10
LD	Low-Delay 63
LTE-Advanced	Long Term Evolution - Advanced 49
MB	Macroblock 20
MC	Motion-Copy 53
MDC	Multiple Description Coding 50
MMCO	Memory Management Control Operation 26
MMT	MPEG Media Transport 2
MOS	Mean Opinion Score 2
MPEG	Moving Picture Experts Group 1
MPEG-2 TS	MPEG-2 Transport Stream 2
MPEG-DASH	MPEG Dynamic Adaptive Streaming over HTTP 2
MSE	Mean Square Error 36
MTU	Maximum Transfer Unit 17
MV	Motion Vector 25
MVE	Motion Vector Extrapolation 55
NAL	Network Abstraction Layer 11
NN	Neural Network 5

PLR	Packet Loss Ratio 47
PMVE	Pixel-based Motion Vector Extrapolation 55
POC	Picture Order Count 11
PPS	Picture Paramenter Set 12
PSNR	Peak Signal-to-Noise Ratio 40
PU	Prediction Unit 20
QoE	Quality of Experience 2
QP	Quantized Parameter 13
RA	Random-Access 63
RADL	Random Access Decodable Leading 15
RAP	Random-Access Point 13
RASL	Random Access Skipped Leading 14
R-D	Rate-Distortion 5
ROI	Region of Interest 19
ROPE	Recursive Optimal Per-pixel Estimation 36
RPS	Reference Picture Selection 42
RPSet	Reference Picture Set 26
RTP	Real-time Transport Protocol 2
SAO	Sample Adaptive Offset 11
SEI	Supplemental Enhancement Information 3
SI	Spatial Information 61
SPS	Sequence Paramenter Set 12
SSIM	Structural Similarity Index Metric 38
STSA	Step-wise Temporal Sub-layer Access 16
TI	Temporal Information 61
TMVP	Temporal Motion Vector Predictor 28
TSA	Temporal Sub-layer Access 16
TU	Transform Unit 20
UEP	Unequal Error Protection 44
UHD	Ultra-HD 1

VCEG	Video Coding Experts Group 1
VCL	Video Coding Layer 3
VPS	Video Paramenter Set 11
WPP	Wavefront Parallel Processing 17
WPSNR	Weighted Peak Signal-to-Noise Ratio 122

Chapter 1

Introduction

1.1 Context and motivation

In recent years new advances in video compression and networking technologies have enabled the development of many multimedia applications over the Internet, such as video conferencing, video streaming, video publishing, among others. The increasing diversity of services, the demand for high quality multimedia and the introduction of Ultra-HD (UHD) formats (*e.g.*, $4k$ or $8k$ resolution and beyond) are pushing forward existing technologies and opening new research directions. Moreover, the emerging new video formats with increased resolutions have created stronger needs for higher coding efficiency, beyond the previous standards, i.e., the H.264 - Advanced Video Coding (H.264/AVC) [1, 2]. Such need has been driven not only by higher resolutions but also by 360-degree, stereoscopic and multi-view video formats, which increase the number of viewpoints that need to be encoded and delivered simultaneously. Moreover, the ever increasing traffic generated by mobile multimedia applications [3] and different video services with growing popularity (e.g., Netflix and Facebook) impose new challenges to the existing networks. To cope with such demanding requirements a new coding standard was developed by the International Telecommunication Union - Telecommunication (ITU-T) Video Coding Experts Group (VCEG) and the International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) Moving Picture Experts Group (MPEG) standardisation organisations, referred to as H.265 - High Efficiency Video Coding (HEVC) [4, 5].

The HEVC is currently the main video coding framework to extend current services towards higher resolution formats delivered with better compression efficiency. Similar to its predecessor, HEVC is a block-based hybrid video codec using predictive, transform and entropy coding. The coding flexibility and efficiency of HEVC results from the use of new coding tools and block structures, which enables new block partitions [6], improved prediction modes [7, 8] and new high-level features [9], such as explicit reference picture management and new parameter sets. However, new coding tools of HEVC not only bring high compression efficiency [10], but also some other disadvantages, such as increased complexity [11] and inherent reduction of error robustness [12]. Earlier studies about the error robustness characteristics of HEVC have shown that, in general, the use of high efficiency coding tools results in poor error resilience performance (score *slightly annoying* in the Mean Opinion Score (MOS) scale), in comparison with previous standards [13, 14]. Thus, when HEVC streams incur in transmission losses, this leads to significant degradation of both objective and subjective quality. This is mostly due to the strong decoding dependencies imposed by the highly complex prediction modes of HEVC, which assume error-free transmission [15]. A possible approach to encode robust video streams is to increase the error resilience of HEVC by including error-aware mechanisms in the coding process. As a result, the effect of error propagation, due to prediction mismatch at the decoding side, will be limited.

The widespread use of HEVC in broadcast and multimedia delivery services is based on transport technologies such as Real-time Transport Protocol (RTP) [16], MPEG-2 Transport Stream (MPEG-2 TS) [17], or more recent standards as MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH) [18]. Recent developments also enable combinations of different technologies, e.g, HTTP-based streaming and MPEG-2 TS, to increase the flexibility of over-the-top video services [19]. For content delivery over heterogeneous networks, the MPEG Media Transport (MMT) [20] is another possible technology to be used with the HEVC standard. Besides the transport technologies, the overall error robustness of video transmission may also benefit from the increasing flexibility of multimedia systems to cope with heterogeneous networks, by using auxiliary streams that allow the video decoder to increase the quality of the recovered lost slices [21, 22]. Moreover, error resilience techniques may also take advantage of the multi-stream based approaches, where each stream allows for different levels of error robustness [18, 23].

In summary, increasing the error robustness of the highly compressed video streams is a relevant research challenge, whose solutions will be useful to increase the adaptability for transmission over diverse network conditions and also cope with communication errors, providing increased levels of Quality of Experience (QoE) to the end users at all times. In general, increased error robustness in video transmission can be jointly

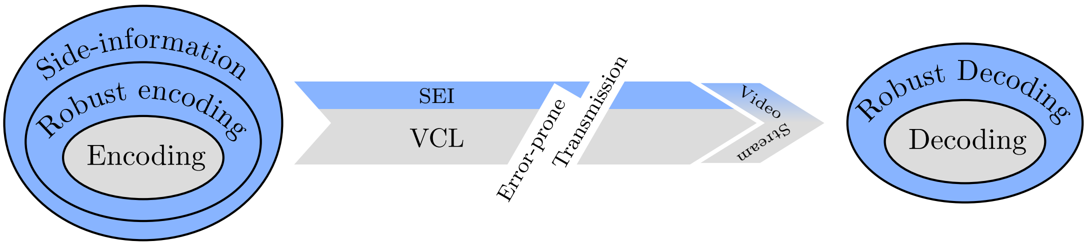


Figure 1.1: Generic robust video transmission system.

achieved at three different stages: (i) using robust video encoding, by selecting robust coding modes, (ii) after encoding, by adding robust features to the compressed streams, as side-information, and (iii) using robust video decoding techniques, e.g., efficient Error Concealment (EC). This is exemplified in Figure 1.1 where at the encoder different layers are used to increase the error robustness, which results in a video stream composed by a Video Coding Layer (VCL) and Supplemental Enhancement Information (SEI) layers. At the decoder two processing layers are used to achieve robust video decoding. The HEVC reference codec is continuously being improved and extended by multiple proponents from around the world, which also address the problem of error resilience and robust delivered transmission. Although different EC methods have been developed during the past years for H.264/AVC [24] and more recently for HEVC [25], a combined and optimised approach involving the encoder and decoder has never been fully exploited. This would improve the users QoE, by exploiting shared information between the encoder and the decoder-sides.

1.2 Main objectives and contributions

The research objectives of this thesis are defined within the scope of robust error resilience for efficient error concealment in highly efficiency video decoders. Thus, this work investigates novel techniques to increase the video quality of HEVC bitstreams delivered through error-prone networks. The following objectives were pursued in this work:

1. **Study and evaluation of error robustness in highly compressed video:** The objective of this research study is to evaluate the robustness against network losses of highly efficiently coded video. The study and its conclusions aim at finding relevant research directions to develop new methods for increasing the error robustness of coded streams. The results and conclusions of this study were published in conference paper (C3) and in journal paper (J1) (see Appendix B).

2. **Robust video coding:** At the encoder-side, novel techniques to extend the existing error robustness of highly predictive video coding techniques are investigated in order to increase the received video quality over error-prone networks. This research topic resulted in a new method to reduce the error propagation in motion vector coding algorithms, which was published in conference paper (C1). Moreover, a novel two-stage approach was devised to improve the error robustness of highly efficiently coded video at the encoder and streaming levels, resulting in two conference publications (C2 and C4) and a journal publication (J1).

3. **Robust video decoding:** error concealment techniques are used to increase the error robustness of video decoding and improve the perceived video quality. The paradigm of EC is evaluated at the encoder-side, as part of the encoding process, in order to optimise the reconstruction process at the decoder-side. This research resulted in a novel error concealment-aware technique to optimise the error concealment process at the encoder-side, which was proposed for intra-frame only in a conference paper (C5) and for generic robust video transmission in the journal (J2).

1.3 Outline of this thesis

This thesis is organised as follows. The current chapter introduces the context, motivation and objectives of this thesis, as well as the original contributions.

The first two chapters review the state-of-the-art concepts and previous work related to this research. Chapter 2 presents an overview of the HEVC standard, focusing on the concepts explored in this work, e.g., high-level features and temporal predictions. Chapter 3 reviews the most important work dealing with robust video transmission also providing relevant background knowledge to this work.

Chapter 4 firstly presents a study about the error robustness of high efficiently coded video, using the HEVC as the underlying framework for various coding configurations. The study presents the relevant tools that mostly affect the error robustness. Secondly, a new method is proposed to decrease the vulnerability of temporal motion prediction and decrease the error propagation, leading to higher video quality under error-prone conditions.

Chapter 5 proposes a novel error resilience approach for high efficient video codecs. A new method to distribute the use of reference frames is proposed to reduce the error propagation and improve the EC performance. Moreover, to reduce the effect of mismatched motion predictions (shown in Chapter 4), a method based on selective motion vector redundancies is proposed. This method is able to improve the decoded video quality using a small overhead.

Chapter 6 addresses the problem of robust EC of high efficiently coded video streams. Different EC algorithms were taken into consideration in order to develop a concealment-aware coding scheme. The proposed scheme is based on a scalable coding approach where the best EC methods to be used at the decoder are optimally determined at the encoder and signalled to the decoder. A generalised saliency-weighted Rate-Distortion (R-D) optimisation is used and the residue between coded frames and their EC substitutes is encoded using a rate-controlled enhancement layer.

Chapter 7 proposes a novel adaptive error robustness method, which combines the different techniques proposed in this thesis. At the encoder-side various video streams are generated with different levels of compression efficiency and error robustness. Then, a pre-training Neural Network (NN) is used to predict which video stream is expected to achieve the highest decoded video quality for a particular scenario. The selection is based on different input parameters, which include information from the video signal, the coded stream and the transmission network.

Finally, Chapter 8 concludes this dissertation and presents some suggestions for future work. Appendix A illustrates the original test signals used in simulations throughout the thesis. This includes a partial set of the HEVC test materials, as well as, other external test sequences. Appendix B presents a summary of the published papers, describing the contributions of the research work of this thesis.

Chapter 2

The High Efficiency Video Coding standard

2.1 Introduction

This chapter presents an overview of the HEVC [4] standard, focusing on the most relevant features of the encoding algorithm, considering the scope of the work developed in this thesis. In short, the HEVC standard essentially aims to support all existing applications of H.264/AVC [1], while extending two keys aspects: increased video resolution with higher compression ratio and also more efficient use of parallel processing architectures. In order to increase the coding flexibility and higher video resolutions, the HEVC standard adopts a new block partition structure, as well as, new prediction modes. In this chapter the new data structures and coding techniques, specifically introduced to increase the coding efficiency, are described. Particular emphasis is given to those which may have a significant impact on the error robustness characteristics of HEVC streams.

The chapter is organised as follows. Section 2.2 introduces the functional blocks of a generic hybrid video coding system and Section 2.3 provides an overview of the HEVC standard. Then, Section 2.4 describes the high-level syntax elements, which are used to carry relevant information in coded streams. Section 2.5 describes the data structures used to divide the video frames into smaller partitions and Section 2.6 describes particular aspects of the main predictions techniques, namely intra and inter coding. Finally, a brief description of the transform and quantisation, entropy coding and in-loop filters is provided in Sections 2.7, 2.8 and 2.9, respectively. Section 2.10 concludes the chapter.

2.2 Generic hybrid video coding

A digital video signal is a temporal succession of images, also referred to as pictures or frames, which usually presents a significant amount of temporal and spatial correlations. In general, video coding and distribution systems follow through a generic chain of processing and delivery stages, comprising pre-processing, encoding, storage or transmission, decoding and post-processing, as shown in Figure 2.1. The pre-processing stage is used to adapt the raw video signal, either analogue or digital, to an appropriate format, such as RGB or YUV [26]. Then, the video data is compressed by the respective encoding modules, mostly reducing its redundancy and visual irrelevancy. The encoding process is responsible for representing the raw video in a highly compressed format, while maintaining a good trade-off between video quality and bitrate through high efficient algorithms and coding tools. In the last stages of the chain, the compressed format is decoded and presented by displaying the signal content after adaptation to the consumer equipment through post-processing, when necessary.

Since the early days of video compressing, hybrid video coding has been the basic structure for all video coding standards and recommendations of ITU-T VCEG and ISO/IEC MPEG, e.g., from H.261 to HEVC. Figure 2.1 also shows the block diagram of a generic hybrid video coding system [27]. While the generic algorithmic structure has not been changed, the specific techniques represented by the building blocks have been refined and the applicable coding configuration has become more and more flexible over years, through the use of an increasing variety of coding tools. This is the reason why the coding scheme is called hybrid, as it combines quiet different prediction modes between pictures with transform and entropy coding [28].

In general, there are two kinds of redundancy in video signal: spatial and temporal. On the one hand, the spatial redundancy is exploited through intra frame prediction, where the underlying mechanism is to find the best matching block for the current one, by searching the neighbouring region. This is possible due to the existence of high correlation among neighbouring pixels within a frame. As a result, when intra frame prediction is used, a block of pixels is reconstructed from the neighbouring blocks previously processed. On the other hand, inter frame prediction is designed to exploit the temporal redundancy that exists among consecutive frames, such as those temporally adjacent, e.g., the past frame or the future frame in regard to the current one. Temporal redundancy is normally referred to as the correlation between two adjacent frames since they usually have quiet similar visual content, such as foreground objects and background. Thus, the purpose of inter-frame prediction is to search for

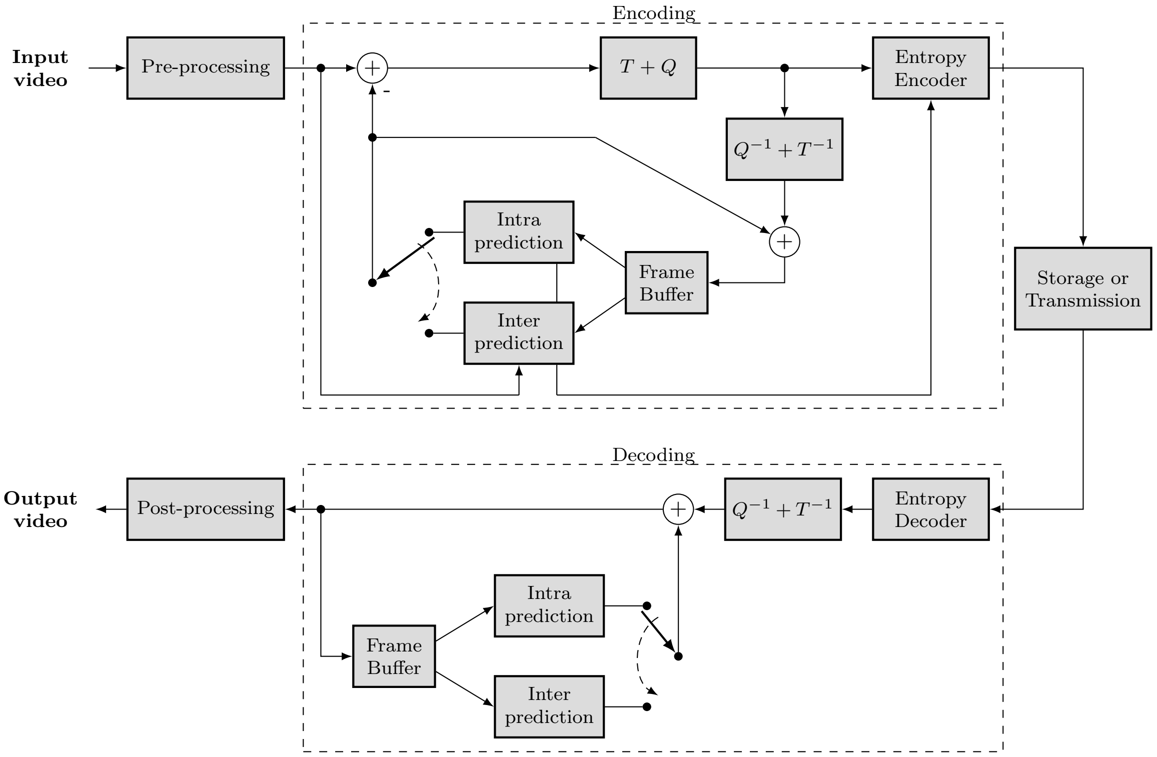


Figure 2.1: Block diagram of a generic video coding system, including the inner blocks of a hybrid encoding and decoding scheme (T: transform; Q: quantisation).

the best matching block for the current one in a set of reference frames.

After finding the best matching candidate for the current block, either by intra- or inter-frame prediction, only the difference between the original block and the predicted one needs to be encoded. As shown in Figure 2.1, these differences (also known as residual signal) are further processed and transformed (T) into coefficients in the frequency domain. Such a transformation decorrelates the residue signal and concentrates the signal energy in a few coefficients, mostly at low frequencies. Subsequently, the transformed coefficients are quantised (Q) to reduce the representation accuracy of the signal, and consequently to reduce the number of bits required for coding. Finally, the quantised transform coefficients together with the prediction information are further compressed by the entropy encoder, forming an output stream suitable for transmission or storage, as shown in Figure 2.1. On the other side of the communication chain, the decoding stage performs the inverse operations on the stream to reconstruct the video signal. In the next section, the aforementioned blocks are described in the context of the HEVC standard.

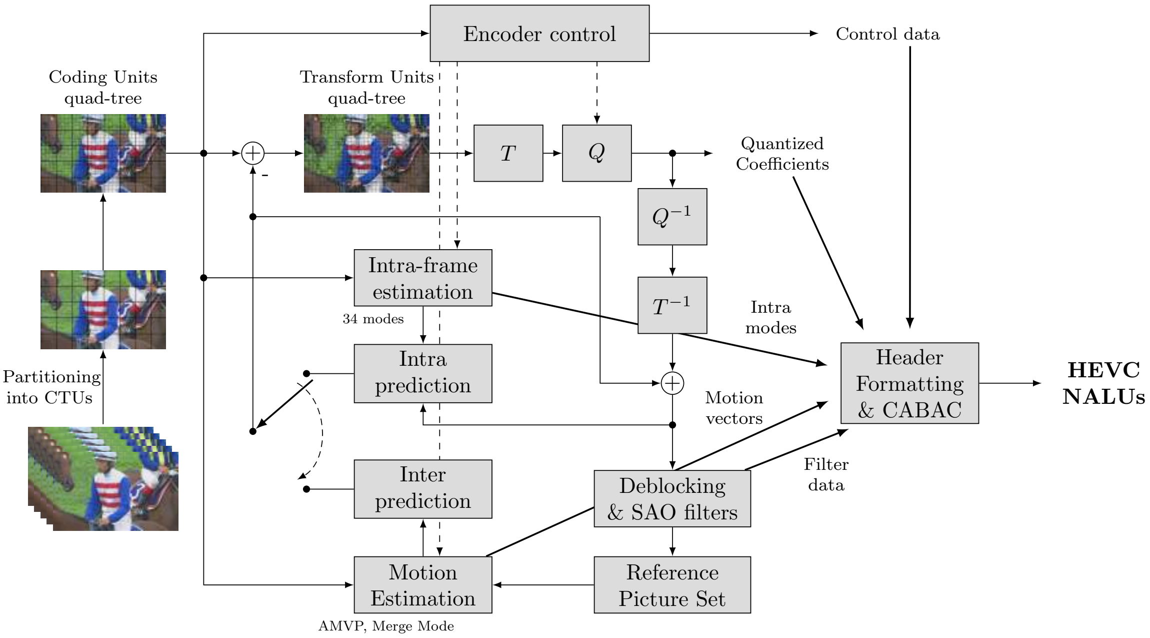


Figure 2.2: Picture partitioning and main functional blocks of an HEVC encoder.

2.3 Overview the standard features

Similar to the previous coding standards developed by Joint Collaborative Team on Video Coding (JCT-VC), in HEVC [5, 29, 30] only the coding syntax and semantics are standardised. The constrains and functionalities are defined through profiles and levels, while the semantic meaning of syntax elements define the decoding process, such that every decoder conforming to the standard generates the same output when given a conforming stream. Such delimitation of scope in the standard concede maximum freedom for developers and manufactures to develop proprietary optimisations, allowing to cope with various applications requirements (*e.g.*, balancing the compression ratio and the coding complexity). Figure 2.2 illustrates the block diagram of the HEVC encoder. It follows the same overall hybrid architecture as its predecessors starting from H.261. The encoder uses both intra- and inter-picture prediction to exploit spatial and temporal redundancies, respectively. Subsequently, the prediction residue is transformed by linear spatial transform and the resulting coefficients are then quantised and entropy coded, further reducing the bitstream redundancy.

However, there are some novelties introduced in the HEVC standard. As depicted in Figure 2.2, the encoding process follows a block-based approach, therefore each picture is partitioned into larger blocks, referred to as Coding Tree Units (CTUs), with a maximum size of 64 × 64 pixels. Subsequently, a dynamic quad-tree partitioning is used to divide the picture into Coding Units (CUs). These units are introduced as the basic processing units for prediction, while for the transform and quantisation operations, transform units are used. In order to generate the prediction residue, the HEVC may either use one of 34 intra-prediction modes or take advantage of symmetric and asymmetric partitions for motion estimation and inter-prediction. The motion estimation is performed in the pictures listed in the reference picture set, which was introduced in the HEVC to increase the robustness to errors and data loss of reference picture management. To improve the overall quality of the reconstructed pictures and smooth the coding artefacts, HEVC not only uses a deblocking filter but also a Sample Adaptive Offset (SAO) filter. Finally, the control data, quantised coefficients, prediction modes, motion vectors and filter information are compressed using the Context Adaptive Binary Arithmetic Coding (CABAC). The compressed bitstream is then allocated to various Network Abstraction Layer (NAL) units.

2.4 High-level syntax

This section describes the high-level components of the HEVC standard that provide support for signalling and stream partitioning, which are relevant elements in the context of robust video communications. HEVC inherits several high-level features from H.264/AVC, e.g., the NAL, parameter sets, the use of Picture Order Count (POC) and SEI messages for auxiliary data. However, some high-level features defined in H.264/AVC were not included, such as Flexible Macroblock Ordering (FMO), redundant slices, arbitrary slice order, data partitioning and switching slices. Furthermore, several new high-level features are introduced, such as the Video Paramenter Set (VPS), tiles and wavefront tools for parallel processing, dependent slices, for reduced delay, and a new reference picture management concept. Moreover, new picture types were introduced to increase the random-access flexibility and temporal sub-layer switching.

2.4.1 Network abstraction layer

Similar to H.264/AVC [1], the HEVC standard defines an adaptation layer to cope with different transport protocols. The NAL is used for stream partitioning and synchronization, which is useful to deal with lossy network environments [31]. Using this concept the video stream is composed of different NAL units. The NAL concept provides the ability to map the VCL data that represents coded video slices (or frames) on various transport layers, such as RTP, MPEG-2 Systems and MPEG DASH. The NAL units are classified into two categories: (i) VCL and (ii) non-VCL, as they carry coded

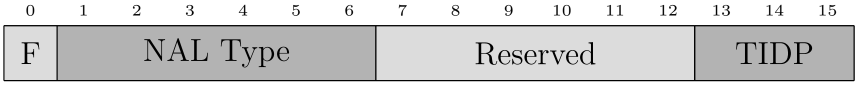


Figure 2.3: NAL unit header format in HEVC.

video or associated data, respectively.

The HEVC standard defines a two bytes header to describe the content of each NAL unit, which are transmitted using a fixed coding format length, i.e., without variable length codes. This allow extra error robustness but increases the bit overhead. Figure 2.3 shows the structure of the HEVC NAL unit header. The first bit *F*, referred to as *forbiden_bit*, is always zero in order to ensure that a start code can be detected across different protocol standards. This is followed by the NAL type, which identifies its content, enabling the decoder to either use it if required for a particular decoding purpose or discard it. Another syntax element introduced by the HEVC standard is the *temporal_id_plus_one* (TIDP), which defines the temporal layer where the NAL unit belongs. This allows implicit temporal scalability (layers ranging from 0 to 6) by immediately discarding the NAL unit if it belongs to a layer that is higher than the pre-defined one. In the header, six bits are reserved for future extensions, preventing the creation of extra NAL units for scalable and multiview extensions, as done in H.264/AVC.

2.4.2 Parameters set

The encoder uses a set of signalling packets as a robust method to share data with the decoder, referred to as parameter set units. This information was introduced in previous standards to deal with vulnerabilities in the transport of signalling information between the encoder and decoder. In this approach, such set of parameters is multiplexed in the bitstream and repeated as many times as required by the application, e.g., streaming, storage among others. In HEVC, this structure is inherited from H.264/AVC [31] with modifications to cope with new coding tools. The HEVC includes three sets of parameters, Sequence Paramenter Set (SPS), Picture Paramenter Set (PPS) and the newly introduced VPS [9].

The new set of parameters VPS contains information related to the different layers of the video signal, avoiding duplications like in the H.264/AVC and its extensions. VPS aims to provide signalling for every layer, guaranteeing an extendible standard, capable of supporting multiple layers. Moreover, VPS also carries information for session negotiation (*e.g.* profile, level and tier). The SPS contains information related

Table 2.1: Brief description of picture types in HEVC standard.

Random-access points pictures		
IDR	Instantaneous decoding refresh	All leading pictures can be decoded
CRA	Clean random access	Pictures in the display order can be decoded
BLA	Broken link access	Leading pictures cannot be decoded
Leading pictures		
RADL	Random access decodable picture	
RASL	Random access skipped picture	
Temporal sub-layer access pictures		
TSA	Temporal sub-layer access	Allows switching to a given temporal layer or higher
STSA	Step-wise temporal sub-layer access	Allows switching to a given layer only

to the whole sequence (*i.e.*, should affect all coded slices) and related to the decoder operation point, as well as flags to control optional tools and scalability. The PPS is responsible to carry information that may change for every picture, such as, initial Quantized Parameter (QP), flags for picture related tools and tilling information.

2.4.3 Picture types

The HEVC standard supports several pictures types, as show in Table 2.1, which are classified in the following categories: (i) random-access points, (ii) leading pictures and (iii) temporal sub-layer access pictures [9, 30]. The use of different types of pictures allows the HEVC stream to be more flexible, support a wider range of applications and achieve robust video communications. They are important to perform random-access, which is a critical feature for channel switching, seek operations and dynamic streaming services. Moreover, introducing random-access points into the bitstream increases its error robustness as they provide anchor points for a clean decoding after data loss or synchronisation problems.

Random-Access Point

The Random-Access Point (RAP) are pictures used to provide access points in the bitstream with no dependencies, where it is possible to start the decoding process. The standard defines that a conforming bitstream must start with a RAP picture.

These frames must belong to temporal sub-layer 0 and should be coded using intra coding techniques, i.e., must not use previously coded frames as reference. One should note that there may be frames compressed with intra prediction only, but not be marked as random-access points. It is always possible to start decoding from a RAP frame onwards, and to output any subsequent pictures in the display order, even if all pictures that precede the RAP in the decoding order are discarded from the bitstream.

The Instantaneous Decoding Refresh (IDR) frame is an intra-coded picture that completely refreshes the decoding process, i.e., cleans the decoder picture buffer preventing predictions to previously coded frames, starting a new conforming video stream. This means that neither the IDR nor any other subsequent frame in the decoding order have prior dependencies. IDR frames are allowed to have leading pictures, i.e., frames that follow the RAP picture in the decoding order but precede it in the display order. However, they must be decodable, thus creating a true random-access point where all subsequent frames in the decoding order can be decoded and displayed.

Introducing RAP reduces the coding efficiency since no prior reference frames can be used. In order to reduce this impact, the HEVC standard introduces a new RAP picture, referred to as Clean Random Access (CRA) picture. These frames do not refresh the decoding process, allowing leading pictures to depend on frames that precede the CRA picture in the decoding order. This enables a more efficient prediction structure, while maintaining an access point to begin the decoding process. However, it results in leading pictures which may not be decodable, which are explained in the next sub-section. The introduction of CRA frames in HEVC was able to improve the coding efficiency by up to 6% [32].

The HEVC supports bitstream splicing, which means taking a particular bitstream from a RAP (both IDR and CRA frames) and inserting it into another bitstream at a random-access point. In case of bitstream splicing, starting from an IDR frame, the leading pictures do not have dependencies from frames prior the RAP, thus all frames are decodable. However, whenever this is performed for a bitstream starting from a CRA frame, the leading associated frames might not be correctly decoded, because some of their reference pictures are not present in the combined bitstream. To make the splicing operation straightforward, the NAL units containing the CRA picture are changed to the type Broken Link Access (BLA). Therefore, the decoder is aware that the leading frames should be discarded, namely the Random Access Skipped Leading (RASL) pictures.

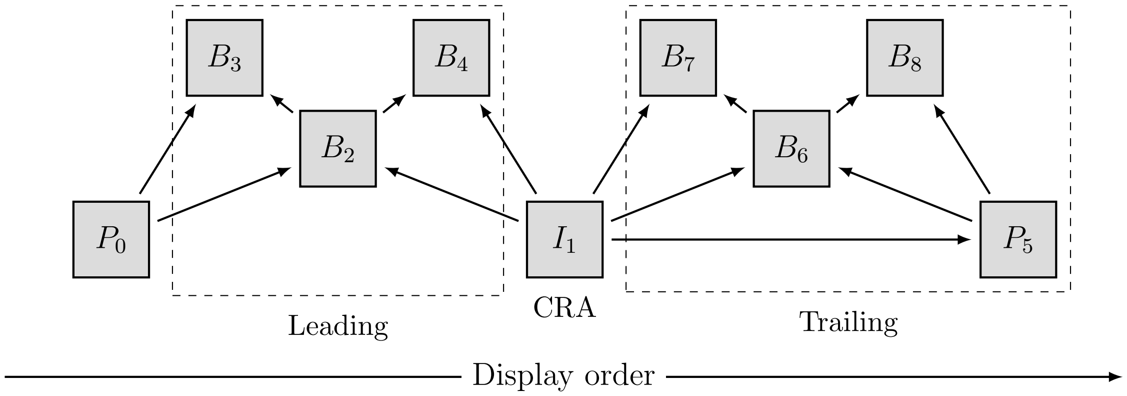


Figure 2.4: Leading and trailing pictures associated with the random-access point I_1 (the letters and indexes identify the coding type and coding order, respectively).

Leading and trailing pictures

Leading pictures correspond to frames that follow a particular RAP frame in the decoding order, but precede it in the display order. A trailing picture is a frame that follows a particular RAP frame in both decoding and display order. Figure 2.4 shows examples of leading and trailing pictures. In the figure, frames are illustrated in the display order and the number associated with the frame type (i.e., I, P and B) represents the decoding order. In this example frame I_1 corresponds to a CRA picture. Leading and trailing pictures are considered to be associated with the closest previous RAP picture in decoding order, such as frame I_1 in Figure 2.4. Also, all leading frames of a RAP frame must precede in decoding order all trailing frames that are associated with the same RAP. This means that the following order is imposed by the HEVC standard: 1) RAP picture, 2) associated leading pictures and 3) associated trailing pictures. Due to the introduction of more flexible random-access points using CRA pictures, it is important to mark the leading and trailing pictures, so the decoder can be aware of which frames can be correctly decodable whenever it starts decoding the bitstream.

There are two types of leading pictures, as shown in Table 2.1. The Random Access Decodable Leading (RADL) pictures, which can be correctly decoded, since they only depend on the associated RAP picture and do not have any prior dependencies to trailing pictures associated with the previous RAP frame. On the contrary, the RASL pictures may have dependencies to prior trailing pictures. When a random-access is performed at the associated RAP frame, these frames cannot be correctly decoded, therefore they have to be skipped. In the example of Figure 2.4 the leading pictures correspond to RASL pictures.

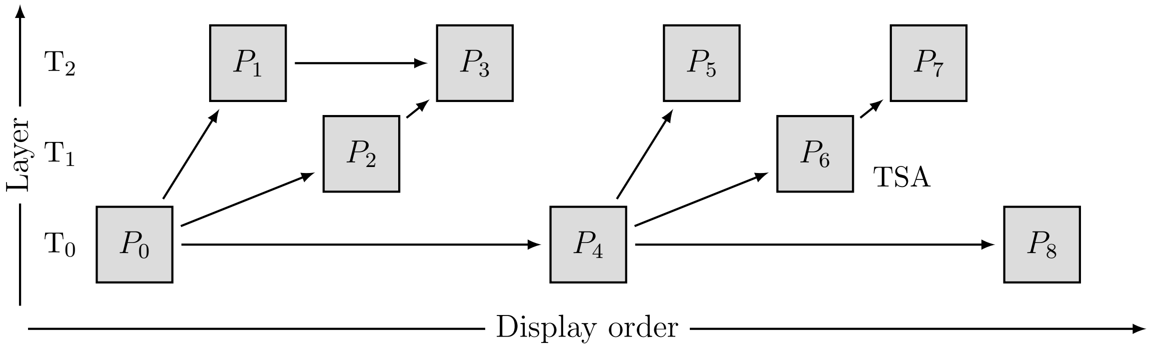


Figure 2.5: Temporal layers access pictures example with three layers, T_0 to T_2 (the letters and indexes identify the coding type and coding order, respectively).

Temporal sub-layer switching pictures

A Temporal Sub-layer Access (TSA) picture is a trailing picture that marks a temporal layer switching point. When decoding a sub-set of temporal layers, if a TSA picture is found in the temporal layer just above the maximum temporal layer currently being decoded, it is possible to start decoding any number of additional temporal layers. For example, frame P_6 in Figure 2.5 is considered a TSA picture since it only depends on a frame belonging to temporal layer 0 for prediction, as well as, any subsequent frame predicted from the TSA frame P_6. In this example, if the decoder is only decoding the temporal layer 0, after decoding P_6 it can also decode layer 1, or decode all the available three layers. Although these frames allow higher flexibility in terms of layer switching, they severely constrain the prediction of frames following the TSA picture.

To reduce the constrains in frame prediction due to layer switching, HEVC introduces the Step-wise Temporal Sub-layer Access (STSA) picture. These pictures have similar purpose as the TSA frames, but they only guarantee that frames in the same temporal layer as the STSA frame are able to be correctly reconstructed. This is guaranteed by not using frames preceding the STSA as reference in its temporal layer. One example of a STSA frame is also shown in Figure 2.5 in frame P_2. This frame can be used to switch to layer 1, as it does not have dependencies to any prior frame in the same layer (P_2 only depends on P_0). However, it cannot be classified as TSA picture because P_3 has dependencies to a prior frame in the decoding order (P_1). Summarising, STSA frames can be used to switch to a particular layer, as TSA frames can be used to switch to any layer.

2.4.4 Picture partitioning

The high-level segmentation of a picture in HEVC is achieved by using four different approaches associated to different data structures: regular slices, dependent slices, tiles and Wavefront Parallel Processing (WPP). Picture partitioning normally serves one or more of the following three purposes:

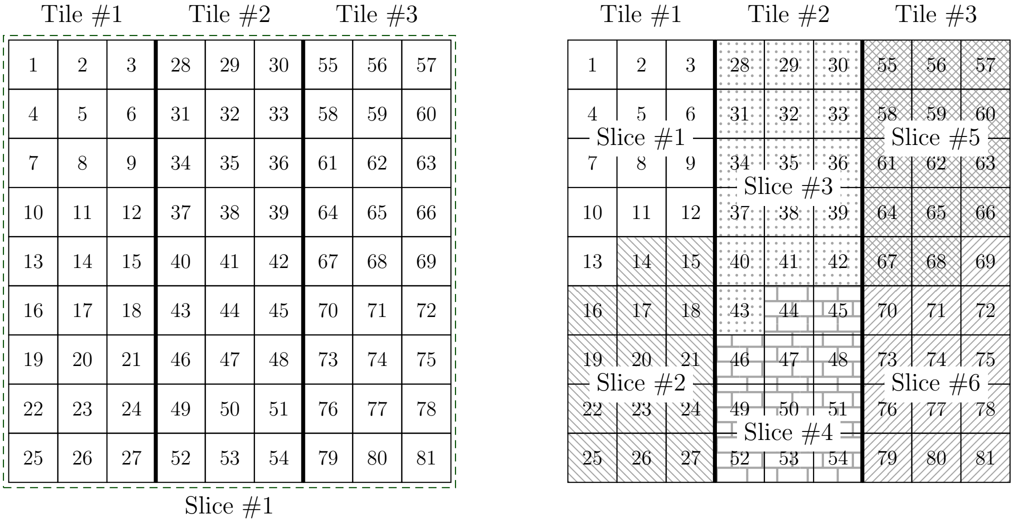
Error robustness: partitioning the picture into smaller self-contained units in order to increase error robustness, allowing to re-synchronize both the parsing and decoding processes in case of data losses.

Network adaptation: adapt to the network constraint of Maximum Transfer Unit (MTU) size, found for example in Internet Protocol (IP) networks. Such packetisation scheme restricts the maximum number of payload bits within a slice regardless the size of the coded frame. To keep each slice within this limit and minimise the packetisation overhead, a variable number of coding units is used for each slice.

Parallel processing: partitioning the coded frame into processing data units, which can be encoded in parallel. This is achieved by dividing the coding units such that they can be encoded and decoded independently of each others.

Slices

The HEVC standard preserves the slice structure previously defined in the H.264/AVC standard. A slice may comprise either the entire frame or a section of it, and all the associated data (*i.e.*, entropy symbols, prediction and residue information) can be independently decoded. Each slice is transported in a different NAL unit. As some dependencies across slice boundaries are disabled, each slice can be independently reconstructed, regardless whether previous slices were lost or incorrectly decoded. Slice partitioning is the only tool that can be used for parallel processing in a virtually identical form as in H.264/AVC. As each slice is independent from each other, it is straightforward to process multiple slices in parallel, without inter-process communication (except for inter-frame prediction). Although this is a simple approach, it incurs in substantial coding overhead due to the higher number of slice headers, and reduction of causal neighbours for prediction (due to lack of predictions across slice boundaries). Another disadvantage of using slices for parallel processing is that they are also used to meet the MTU size constrains. Therefore, it might be impossible to meet both requirements using only one technique.



(a) (b)

Figure 2.6: Examples of the interaction between tiles and slices.

Another partitioning tool in HEVC standard is the dependent slices. These types of slices allow the partitioning of a frame at the coding unit boundaries without breaking intra-frame predictions. Moreover, as they have a smaller slice header they are more efficient in terms of signalling overhead than regular slices. Dependent slices are normally used to reduce the end-to-end delay by allowing part of the slice to be transmitted while the rest of its data is still being processed.

Tiles

A new data partition has been included in the HEVC standard, referred to as Tiles [33]. Tiles enable the partition of a coded frame into a smaller group of CTUs using horizontal and/or vertical boundaries. Figure 2.6 shows two examples of slice partitioning using tiles. In both cases the frame is partitioned into three tiles by using two vertical boundaries. By only using boundaries to define tiles, signalling is achieved with a small overhead. These boundaries are defined in the PPS NAL units.

Although the HEVC standard enables the use of both slices and tiles at the same time, some constrains are imposed in order to reduce the implementation complexity. One of two cases must be used: (i) all CTUs within a tile belong to the same slice, or (ii) all CTUs within a slice belong to the same tile. These cases are illustrated in Figure 2.6 (a) and (b), respectively.

The use of tiles provide several advantages. Specifically, they improve frame partitioning for parallel processing [34] when compared to slices, by reducing the required

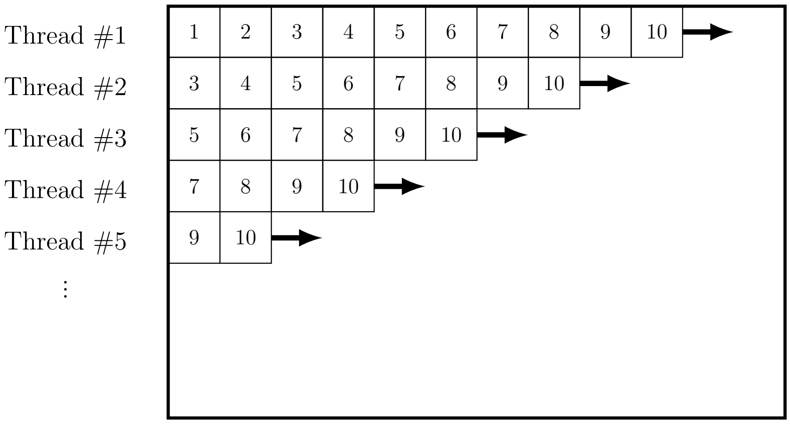


Figure 2.7: Wavefront parallel processing (the numbers indicate the processing instant of each CTU).

overhead. Moreover, they also improve the MTU size matching and reduce the line buffer memory (by dividing the frame as in Figure 2.6). Finally, as tiles allow flexible partitioning, they enable the definition of Regions of Interest (ROIs) for asymmetric video coding.

Wavefront Parallel Processing

As mentioned before, the use of either slices or tiles to process each frame using parallel threads requires the use of different entropy coding contexts for each slice/tile, in order to make them independently decodable. As a consequence, the compression efficiency decreases. In order to overcome this issue, the concept of WPP was introduced, which enables coding using efficient parallel processing [35].

In WPP, each slice (or tile) is divided into single rows of CTUs which can be processed in parallel, but entropy coding and prediction are enabled across CTUs of different rows, in order to avoid loss in coding efficiency. To re-use the same context in the whole slice (or tile) each row requires a delay of two CTUs in order to ensure that all necessary dependencies have already been encoded/decoded. Figure 2.7 illustrates an example of WPP using at least five different threads. Each square corresponds to a CTU and the numbers correspond to the time instance when a CTU starts being processed. Using this staggered start, which appears like a wavefront, as represented in Figure 2.7, the parallelisation may use as many threads as the number of CTUs rows available in the video frame. However, although the number of threads does not affect the coding efficiency, the required inter-process communication substantially increases.

2.5 Data structures

Despite the fact that the HEVC standard follows a traditional hybrid video coding architecture, it introduces significant changes in the data structures when compared to the previous standards (e.g., H.264/AVC). In H.264/AVC each frame is divided into units so-called Macroblocks (MBs), which are composed by a 16 × 16 block of luma samples and two 8 × 8 blocks of chroma samples, assuming video data in YUV 4:2:0 chroma sub-sampling. The MB is considered the basic processing unit in the H.264/AVC, and the respective coding mode defines whether the corresponding pixels are coded as intra or intra prediction. Each MB could also be partitioned into sub-blocks, which should follow the same prediction mode.

The basic data structures in HEVC are quite different from H.264/AVC. Each picture is divided into blocks named as CTU [36]. The maximum CTU size is an encoder configuration parameter that is signalled to the decoder, and should be one of the following: 64 × 64, 32 × 32 and 16 × 16 pixels. The use of data structures with larger size in HEVC was adopted to achieve higher coding efficiency especially in high resolution video. Each CTU can be recursively partitioned into four smaller units, referred to as CUs, until it reaches the minimum unit size, which cannot be smaller than 8×8 pixels. Similar to the maximum CTU size, the minimum size is also defined at the encoder. For instance, in case of homogeneous regions, large CUs can be used to represent such regions by using a smaller number of symbols than in the case of using several small units. Figure 2.8 illustrates an example of a CTU partitioning into CUs. The partitioning illustrated in the figure leads to the quad-tree structure represented on the left side with the corresponding coding order shown on the right side. Data structures in HEVC provide a significant improvement in comparison with previous standards, by eliminating the distinction between MB and sub-MB and using only the CU concept. This allows for the multilevel hierarchical quadtree structure to be specified in a simple and elegant way, with size-independent syntax representation.

The CU is the basic processing unit in HEVC. It is further partitioned into two types of units, specifically Prediction Unit (PU) and Transform Unit (TU), which are data blocks used for prediction and transform operations, respectively. The decision whether to use intra-frame or inter-frame prediction modes is taken at the CU level. However, depending on the partitioning of the CU into PUs, different intra or inter coding modes can be used for each PU.

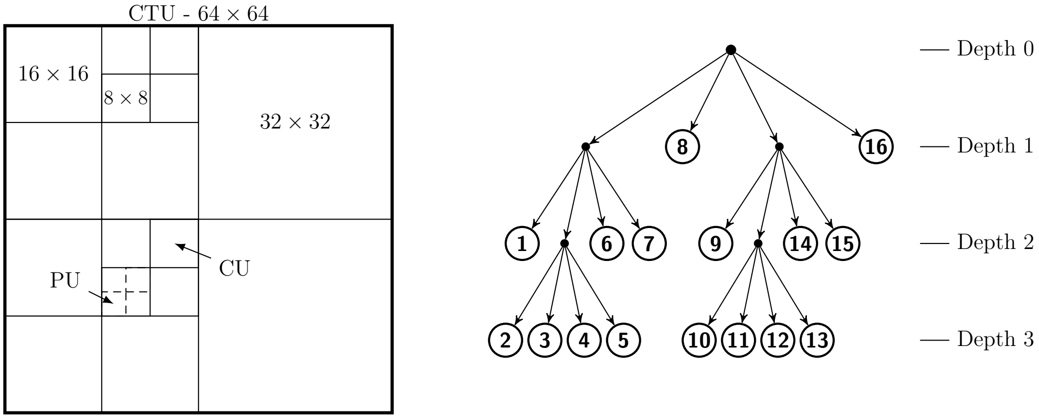


Figure 2.8: Representation of a CTU (64 × 64) and its partitioning into CUs and PUs (left); corresponding quad-tree with coding order (right).

2.6 Prediction modes

Prediction techniques play an important role in image and video coding standards, due to their ability to reduce the signal redundancy based on the previously encoded pixels. These techniques include the directional intra prediction for efficiently reduce spatial redundancy, and inter-frame motion compensation to remove the temporal redundancies. This section describes the prediction techniques used in the HEVC standard that comprises different algorithms.

2.6.1 Intra-frame prediction

Intra-frame prediction is used to efficiently reduce the spatial redundancy within video frames. Intra predicted blocks are obtained using the previously encoded pixels around the current PU. When the intra prediction mode is chosen for a CU, the size of the PU needs to be the same as the CU, except when the smallest CU size is selected and further division into four PUs is allowed. For example, when the smaller CU size is 8×8, the encoder may decide to divide it into four PUs of 4×4 pixels, each one having its own intra prediction mode. This is signalled using a binary flag. The use of small PUs is useful for regions with many details that require fine-granularity prediction.

In HEVC there are 35 intra-frame prediction modes: Planar, DC and 33 angular modes, as shown on the left of Figure 2.9 [7, 37]. For each PU, the encoder chooses among all the available intra prediction modes the best mode to be used by the decoder, based on the previously reconstructed pixels also available for similar prediction at the decoder. The reconstructed reference pixels used in the prediction process belong to the neighbouring blocks located at left-down, left, top-left, top and top-right positions,

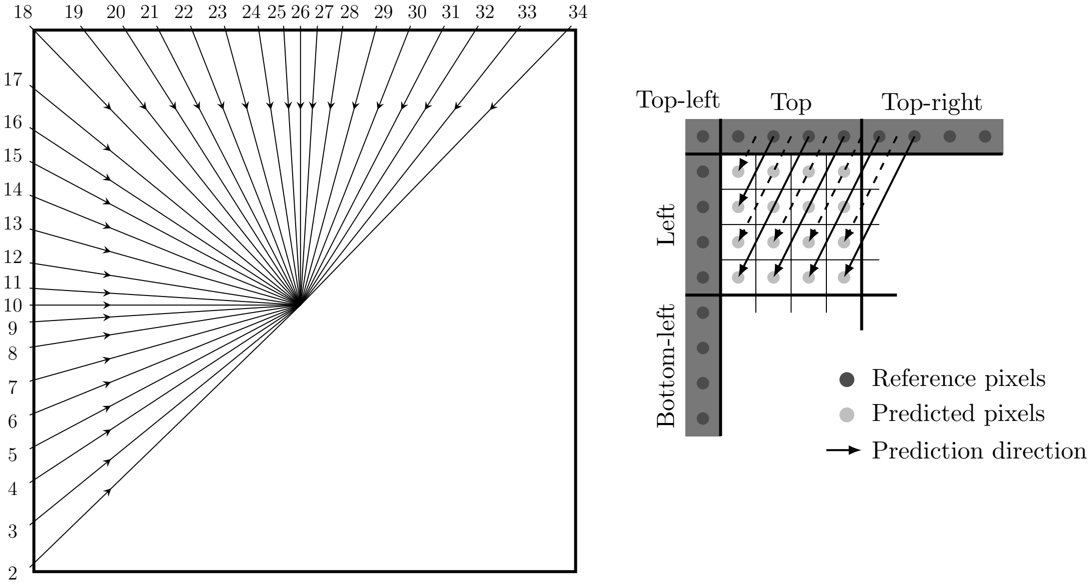


Figure 2.9: Directional intra prediction modes in HEVC (left) and example of intra prediction mode 32 (right).

as can be seen in the example shown on the right in Figure 2.9. Given a $N \times N$ PU, intra prediction requires the top neighbouring row of $2N$ pixels, the left column of $2N$ pixels and the top-left neighbouring pixel. Due to data unit boundaries (e.g., slice or tile boundaries) and constrained intra prediction (i.e., reference pixels belonging to inter-predicted PUs are not used in order to remove error propagation from potentially erroneous reconstructed reference frames), neighbouring pixels might not be all available to be used as reference for intra prediction, resulting in an incomplete set of neighbouring reference pixels. In order to overcome this issue and to allow all possible modes, HEVC uses reference sample substitution to replace the unavailable reference pixels with the closest available ones, so that all intra prediction modes can be used.

Angular modes

Directional intra prediction was first introduced in the H.264/AVC standard [2] to estimate regions with a structured texture or directional edges. Directional prediction projects the reconstructed samples in the block neighbourhood (see reference pixels in Figure 2.9) along a specific direction. The HEVC increases the number of possible directions from 8 (in H.264/AVC) to 33 [7]. The direction of each projection is illustrated in the left-side of Figure 2.9 by the directional modes 2 to 34.

As shown in Figure 2.9, the angular modes are designed to provide a dense coverage near the horizontal (i.e., mode 10) and vertical (i.e., mode 26) directions. Moreover,

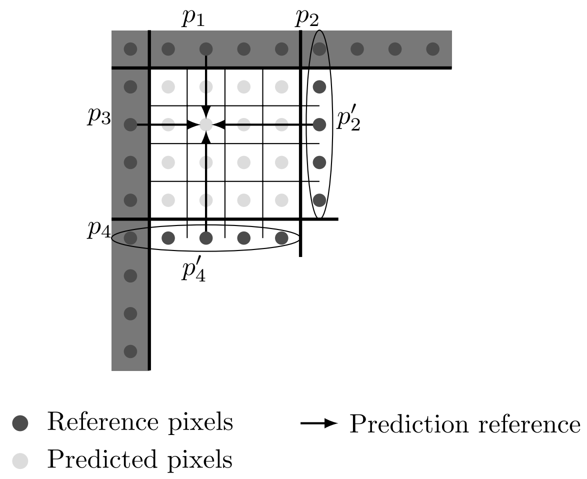


Figure 2.10: Example of Planar prediction for a 4 × 4 block.

the angular directions are coarser as it gets closer to the diagonals. This reflects the observed statistical prevalence of the angles and the prediction efficiency. To improve the efficiency of angular prediction, when the reference samples need to be projected, a bilinear interpolation is used from the two closest pixels using a 1/32 accuracy. The dashed arrows on the right of Figure 2.9 correspond to a projection of interpolated pixels.

Planar and DC

The Planar and DC modes are particularly efficient in the prediction of smooth regions. The DC mode uses the average or the neighbouring reference samples to generate a constant prediction for the current PU as a whole, i.e., all pixels are predicted from the same value. The Planar mode predicts the PU through a linear interpolation from the four closest neighbouring reference pixels. Figure 2.10 shows the reference pixels (i.e., p_1, p'_2, p_3 and p'_4) and arrows pointing to the pixel being predicted. Since the right and bottom neighbourhood is not available (not coded yet), the same pixel p_2 is repeated to form a column of reference pixels, marked with an ellipse in Figure 2.10. A similar approach is also used to create the bottom row of reference pixels from the pixel p_4 .

Sample smoothing

In HEVC sample smoothing is applied in two distinct cases to improve the overall performance of the prediction accuracy. Firstly, the reference samples are filtered and secondly the PU is filtered after prediction.

In HEVC, the reference sample smoothing is conditionally applied, based on the

PU and intra prediction mode. Similar to H.264/AVC, a three-tap smoothing filter (i.e., [1 2 1] /4) is applied, except for PU sizes of 4 × 4. The prediction resulting from Planar mode is also filtered when the PU size is greater than 8 × 8, and it is not used for the DC mode.

The second filter in HEVC is the boundary smoothing. This filter aims at removing the discontinuities along block boundaries due to intra prediction. This filter is only applied when the block size is smaller than 32 × 32 to the DC, and angular modes 10 and 26 (exactly the horizontal and vertical directions, respectively). For the DC mode a two-tap filter ([3 1] /4), fed by the original predicted value and the adjacent reference sample is applied to both the top row and the left-most column of predicted pixels. In the case of angular mode 10 (horizontal direction), only the predicted samples of the first column are changed by adding half of the difference between the adjacent reference sample and the top-left reference sample. A similar filtering is done for angular mode 26 (vertical direction), but only for the first row of predicted samples. Only the mentioned boundaries are filtered, as they are the ones most likely to introduce discontinuities.

2.6.2 Inter-frame prediction

The efficiency of video coding algorithms consistently relies on inter-frame prediction techniques to reduce temporal redundancy. The underlying idea of inter-frame prediction is to estimate the current frame from one or more previously encoded frames used as reference with block-based motion compensation. The most common technique used for motion estimation in HEVC is the block matching algorithm. In the HEVC standard an Advanced Motion Vector Prediction (AMVP) is used for efficient motion vector coding, which defines the Merge Mode and differential coding of the motion information.

Motion estimation and compensation

Motion estimation consists in searching for a block with the highest similarity with the current block to be predicted (PU in HEVC), using block-matching algorithms over previously encoded frames, i.e. the reference frames. Typically, reference frames correspond to past or future temporally adjacent frames. The frame encoding order determines which references are selected and which future frames are available. The reference block that results in the lowest error is selected as the best block for prediction. To identify the best matching block to be used for motion compensation, the difference between the target and the reference block positions is encoded as a two-dimensional

Motion Vector (MV). The HEVC allows for either one or two MVs to be used for each PU. Video frames of type P are predicted from one single reference (one MV), while frame of type B are predicted from two references (two MVs). This results in uni-direction or bi-directional coding, respectively. Furthermore, weighted prediction can also be used, which consists in scaling and offset operations performed on the prediction block to improve efficiency. The MVs only refer to frames included in a reference frame list, where each frame is identified by an index. Therefore, to fully describe inter-frame prediction a combination of two parameters is required: MV and reference frame index.

Since MPEG-2, motion estimation/compensation has been performed using blocks of variable size. In H.264/AVC block sizes ranging from 4×4 up to 16×16 pixels are allowed, while HEVC extends the possible block sizes up to 64×64 pixels, and also to asymmetric rectangular block sizes. By using larger prediction units less MVs are required, thus fewer bits are spent in coding motion information. In order to increase the efficiency of motion estimation fractional MVs are used, which has been proven to improve motion compensation accuracy. This is accomplished through interpolation to generate non-integer pixels in the reference frame, resulting in MVs up to quarter-pixel accuracy for luma samples. For chroma samples, the MV accuracy is determined according to the chroma sampling format, which for 4:2:0 results in eighth-pixel accuracy. In contrast with the two-stage approach used in H.264/AVC, HEVC uses a single separable approach for the interpolation process without intermediate rounding operations. This improves the precision and simplifies the architecture of the fractional interpolation. The partition size as well as the MV are selected using a R-D optimisation which considers both the number of bits required to code the motion information and the video distortion.

The block matching-based motion estimation is one of the most computationally intensive procedures in video encoders. In general, the optimal solution for motion estimation is provided by the full-search algorithm, which tests all the possible candidate blocks in the search area. However, although the full-search algorithm achieves the optimal performance, it requires the highest computational complexity. Therefore, sub-optimal algorithms have been proposed over the years to reduce the motion estimation complexity and decrease encoding time [38–40]. For example, the HEVC Test Model (HM) software implements a fast algorithm based on Test Zone Search [38]. Since motion estimation is a non-normative procedure, encoders may implement any alternative motion search algorithms.

Table 2.2: Example of a reference picture set.

Frame	RPSet ({frame, is used})		
I_1	–		
P_5	$\{I_1, 1\}$		
B_6	$\{I_1, 1\}$	$\{P_5, 1\}$	
B_7	$\{I_1, 1\}$	$\{B_6, 1\}$	$\{P_5, 0\}$
B_8	$\{B_6, 1\}$	$\{P_5, 1\}$	

Reference picture management

Another improvement in HEVC is related with the reference picture management scheme. HEVC introduces the Reference Picture Set (RPSet) concept [9], which defines how previously decoded pictures are managed in the Decoded Picture Buffer (DPB). Decoded frames in the DPB are grouped in one of the following categories: (i) *short-term reference*, (ii) *long-term reference* and (iii) *unused for reference*. Once a frame is marked as *unused for reference* it is no longer used for motion compensation and it will be discarded from the DPB after being displayed. In H.264/AVC a sliding window approach is used to implicitly manage the reference picture list. This process discards a given frame whenever the DPB has stored the maximum number of reference frames (which is defined and controlled by the encoder). Moreover, H.264/AVC uses an explicitly Memory Management Control Operation (MMCO) to send commands to the decoder and change the normal reference picture management process.

In the HEVC standard the status of the DPB is encoded for every slice, instead of the implicit managing used in the past. The most fundamental difference in RPSet concept compared to MMCO/sliding window of H.264/AVC is that each particular frame has a complete set of the reference frames that are used or will be used. Thus, a complete set of all frames that must be kept in the DPB is explicitly provided. This is different from H.264/AVC where only relative changes to the DPB are encoded. With the RPSet concept, no information from earlier frames in decoding order is needed to maintain the correct status of the DPB. Table 2.2 shows the RPSet for some frames illustrated in Figure 2.4. As expected, for the CRA frame (I_1) no information is sent to the decoder because there is no inter-prediction and the DPB should be empty. For the remaining frames, two types of information are encoded: frame number and a binary flag to indicate whether it is currently used. This is required because any frame not listed in a given RPSet will be discarded by the decoder and then subsequent frames cannot use them as reference.

The RPSet concept provides a basic level of error robustness to the reference picture

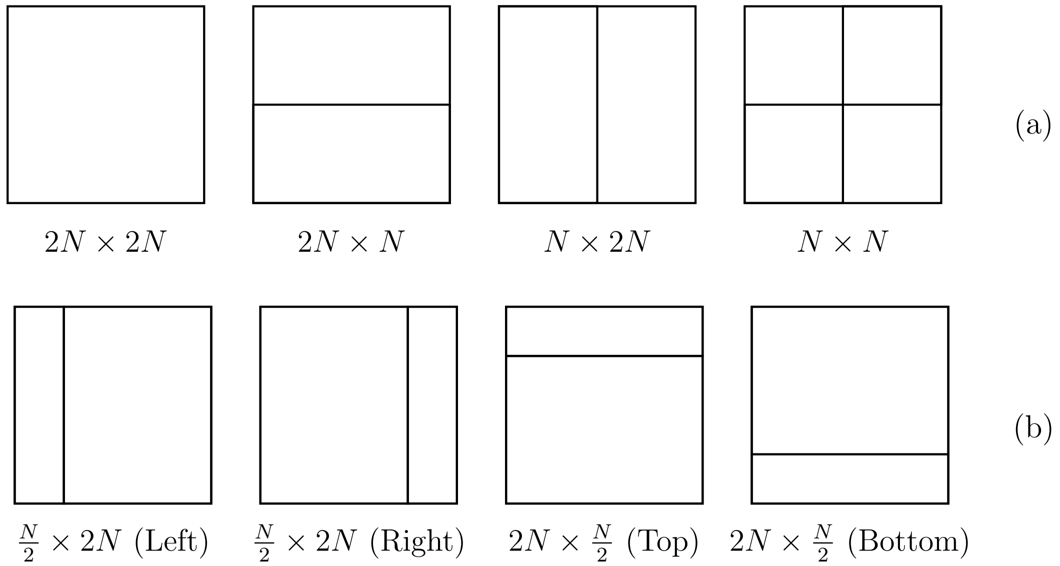

$2N \times 2N$
$2N \times N$
$N \times 2N$
$N \times N$
(a)
$\frac{N}{2} \times 2N$ (Left)
$\frac{N}{2} \times 2N$ (Right)
$2N \times \frac{N}{2}$ (Top)
$2N \times \frac{N}{2}$ (Bottom)
(b)

Figure 2.11: Partition types of an inter-coded CU (size $2N \times 2N$) into either (a) symmetric or (b) asymmetric PUs - the size and relative position of the smallest PU are indicated [6].

management. In H.264/AVC, whenever a reference frame is lost the decoder can look at the value of the *FrameNum* (incremental value associated with each reference frame) to detect the loss event. However, as gaps in *FrameNum* are allowed, the decoder might not be able to decide whether the missing frame was intentional or resulted from a transmission loss event. Using the RPSet concept, the decoder is always able to detect loss events and correctly identify the missing frames.

Asymmetric block partitions

When inter prediction mode is selected, a higher range of partitioning schemes is allowed, thus dividing the CU into one, two or four PUs. The splitting into four PUs is only allowed when the CU size is equal to the minimum. Figure 2.11 illustrates the possible CU partitioning into PUs. Figure 2.11 (a) shows the symmetric partitions. The $2N \times 2N$ corresponds to the case where the CU is not further divided. The HEVC standard defines the horizontal and vertical division into two PUs, corresponding to the $N \times 2N$ and $2N \times N$ cases, respectively. Finally, the $N \times N$ case results in four PUs. The partitions shown in Figure 2.11 (b) correspond to the asymmetric partitioning, which is one of the main improvements introduced in HEVC, as it allows higher flexibility in motion estimation, leading to higher compression efficiency [6]. These partitions types are only allowed for CUs with 16×16 pixels or larger.

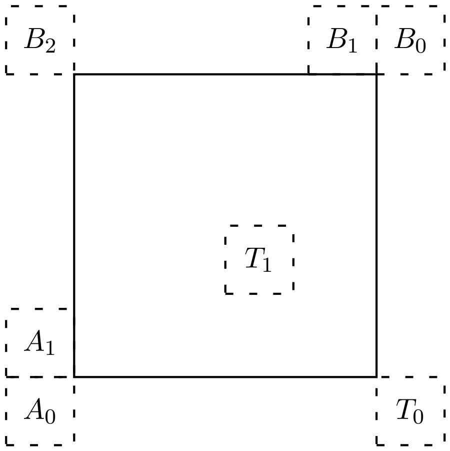


Figure 2.12: Illustration of the MV prediction candidates in the AMVP and Merge Mode mode.

Merge mode

The Merge mode was introduced in HEVC as a new technique to derive motion information from spatial and temporal neighbouring blocks [41]. This technique extends the concepts of the Direct and Skip modes used in H.264/AVC [2], using a more sophisticated approach with two key differences. Firstly, the Merge Mode allows for more MV candidates than the Skip in H.264/AVC. Thus, an index is used to identify one out of several MV candidates. Secondly, the reference picture index is explicitly identified. This results in improved flexibility, compared with the predefined values used in H.264/AVC. The MV candidates allowed in the Merge Mode [41] are illustrated in Figure 2.12. The figure shows the positions of the spatial MV candidates (A_0, A_1, B_0, B_1, B_2). Moreover, it also uses a temporal MV candidate, derived from the co-located position on the temporally adjacent frame, at positions T_0 or T_1. From the MV candidate list the unavailable candidates are not considered, and duplicated candidates (i.e. with the same motion information) are removed. The Merge Mode uses a $C - 1$ spatial MV candidates, selected from the aforementioned order, and one Temporal Motion Vector Predictor (TMVP) to fulfil the total number of C MVs. The value of C is defined at the slice header and a flag controlling the use of TMVP is encoded at the PPS-level. For B-slices, additional MV candidates are considered by choosing two existing candidates, according to a predefined order for reference picture list 0 and list 1. A total number of twelve combinations are specified in HEVC, which are listed in [5]. For P-slices, if the number of candidates is less than C, then zero MVs are included to fill the remaining entries.

The Merge Mode defines how to encode the MV candidate and the residual information. In HEVC, the Skip mode, i.e., transmitting only the MV without residual information, is treated as a special case of the Merge Mode. In this case, only a skip flag is enabled and the corresponding merge index is transmitted.

Motion vector prediction

The HEVC also improves the performance of MV coding for those cases where the Merge Mode is not used. A differential encoding approach is always used for efficiently representing the motion information. It considers the same MV candidates of the Merge mode, illustrated in Figure 2.12, to derive the MV predictors from which the differential vector is obtained. From these MV candidates only the first two available candidates are selected, based on the following order:

1. Left candidate: first available from A_0 and A_1.
2. Top candidate: first available from B_0, B_1 and B_2.
3. Temporal candidate: first available from T_0 and T_1.

Therefore, the use of the temporal MV candidate is limited to those cases where the spatial neighbours are not able to provide two valid candidates. For MV prediction only a much lower number of candidates is allowed, since the encoder can still send the MV difference. Moreover, the encoder performs motion estimation for each candidate, which increases the encoding complexity for every possibility tested.

2.7 Transform and quantisation

Similarly to the previous standards, HEVC uses a transform followed by quantisation to compress the residual signal. The residual block is partitioned into multiple square blocks, referred to as TUs.

Core transform

A 2D transform of the residual information is obtained by applying 1D transforms in both the horizontal and vertical directions. The core transform matrices are obtained from the Discrete Cosine Transform (DCT) basis functions to define a integer matrix of 32×32 points. Sub-sampled versions of this matrix are used to derive the remaining transform matrix sizes down to 4×4.

A special case is applied for coding the residual of 4×4 intra-coded block. In this case an alternative integer transform based on the Discrete Sine Transform (DST) is used. The motivation for DST is related to the statistics of the residual signal of these blocks, which tend to present higher residue as the distance to the boundary used for prediction (i.e, top or left boundary) increases. Simulation studies have shown that DST provides up to 1% of bitrate savings for the compression of 4×4 intra predicted blocks, without significantly increasing the computational complexity [42].

Quantisation

The HEVC follows the same principle as previous standards in terms of quantisation. After applying the transform operation to the prediction residue, a quantisation is applied to the resulting coefficients using a uniform-reconstruction quantiser scheme based on a QP. The standard supports QP values ranging from 0 to 51, where an increase of 6 in the QP corresponds to doubling the quantisation step. This procedure is the main cause for the coding distortion, as it performs a many-to-one mapping, which cannot be reversed.

2.8 Entropy coding

HEVC only uses the CABAC, which is an arithmetic coding method that only uses binary symbols, considering different probability models for each symbol [43]. Entropy coding efficiency is closely linked with the context model selected. Therefore, this was carefully designed in the HEVC standard, extending the functionality previously defined for H.264/AVC. For example, the HEVC standard exploits the depth of the partition tree or residual transform tree, in order to derive the context models for several syntax elements. Regarding the transform coefficient coding, CABAC uses a scanning method to firstly organise the coefficients and then encodes the position of the last non-zero transform coefficient. Moreover, a significance map is also encoded along the sign bits and the levels of the transform coefficients. Three coefficient scanning methods are available: the diagonal up-right, the horizontal and the vertical scan. The coefficient scanning is implicitly selected and always performed in 4×4 sub-blocks, for all transform block sizes.

2.9 In-loop filters

The HEVC standard uses two filtering procedures over the reconstructed pixels before writing them into the frame buffer, namely the Deblocking Filter (DBF) and the SAO filter. The main purpose of the deblocking filter is to reduce blocking artefacts caused by the block-based coding tools. DBF is applied to the samples that are spatially adjacent to the PU or TU boundaries, considering a 8 × 8 grid. Such grid-based restriction reduces the computational complexity and facilitates the parallel-processing, which is one of the main goals of HEVC design. Three strength levels can be used by the DBF, depending on the coded block characteristics.

The SAO is a non-linear filter, that is adaptively applied to all samples of the image, after the deblocking filter. In its operation, the SAO filter modifies the samples by adding an offset value, extracted from lookup tables transmitted by the encoder. For each CTU, the encoder decides whether to apply the SAO filter or not, and if used, one of the two filter types is applied, namely the band offset or the edge offset. In the band offset mode, the added offset value depends on the sample amplitude, while the edge offset uses the gradient to classify and derive the offset value.

2.10 Summary

This chapter was devoted to describe the main features of the HEVC standard with special emphasis on the high-level features, coding structures and newly introduced prediction modes. At the high-level, HEVC introduces new NAL types and also new pictures types. These allow the implementation of various random-access strategies, more flexibility and higher coding efficiency than its predecessor H.264/AVC. Another major improvement comes from the inter-frame prediction. On the one hand new high-level reference picture management was introduced, which increases the error robustness, as it explicitly defines the active reference frames and the DPB status. On the other hand it introduces the AMVP and the Merge mode, which efficiently reduce the required bits for coding the motion information. The newly developed coding tools and prediction modes described in this chapter have resulted in approximately 50% of bitrate savings for the same subjective quality in comparison with the previous standard. However, such high coding efficiency implies higher data dependencies, which are more sensible to data loss and increase the error propagation to subsequent coding structures. Therefore, evaluating the impact of network errors and error propagation in HEVC is of most importance to develop efficient techniques to recover from data loss and reduce the error propagation.

Chapter 3

Robust video coding: a review

3.1 Introduction

In general, digital communications through networks are prone to transmission errors and packet loss. These errors can be handled in different ways, such as retransmission, error correction, and EC [44, 45]. This chapter reviews the existing techniques for robust video coding that can be applied at the VCL to reduce the error propagation and increase the decoded video quality. Several techniques have been proposed in the past to deal with the problem of robust video transmission, especially for video coding standards with high temporal dependencies, such as H.264/AVC and more recently HEVC. These techniques are normally focused on the spatio-temporal dependencies in the coded streams, which are the basis of high coding gains but lead to low error robustness [15]. In this chapter the error resilience and error concealment schemes are classified in different categories, depending on their functionality and algorithms. Although each approach is based on different techniques they all aim at striking the best trade-off between coding efficiency and decoded video quality under lossy conditions.

The chapter is organised as follows. Section 3.2 provides an overall description of the different categories of error resilience. Section 3.3 describes the localisation techniques, which are able to reduce the temporal and spatial dependencies in coded bitstreams. Section 3.4 provides an overview on methods based on data partition, classification and bitstream redundancy. Section 3.5 covers the techniques used at the decoder-side, as post-processing, to reduce the effect of errors and improve visual quality and smoothness. Section 3.6 provides a description of methods that use EC techniques to cope with packet loss and improve error robustness. Finally, Section 3.7 concludes the chapter.

3.2 Robust video coding techniques

In general, most of the compression efficiency is achieved by exploiting both the spatial and temporal redundancy of video signals. Since there is high correlation between neighbouring information, predictive coding schemes are very effective to achieve higher compression gains by exploiting such correlations. However, the drawback of high predictive coding schemes is the strong dependencies between compressed symbols, which may have a great impact on error propagation. In order to increase the robustness of video transmission, several coding mechanisms can be employed at different elements of a transmission chain to mitigate the artefacts caused by errors and data loss. These mechanisms can be classified according to the functional characteristics and the type of problem they address [46]. In this chapter the robust video coding techniques are grouped into four categories:

- Localization techniques;
- Data partitioning and redundant coding;
- Error concealment at the decoder;
- Concealment-aware error resilience.

Table 3.1 provides a summary of the techniques covered in this section, some examples and the key advantages associated with each group. Localisation techniques mainly localise and break the predictive coding loop, reducing the spatial and temporal dependencies, so that if an error occurs, then it is not likely to affect other parts of

Table 3.1: Advantages of error robustness techniques by category.

Category	**Examples**	**Advantage**
Localization	• End-to-end distortion • Adaptive intra refresh • Reference picture selection • Leaky predictions	Reduces error propagation
Data partitioning and redundancy	• Slice classification • Flexible macroblock order • Redundant pictures • Multiple-description coding	Unequal error protection, transport priority and multi-channel transmission
Error concealment at the decoder	• Linear interpolation • Boundary matching • Motion vector extrapolation	Does not require extra information and adapts to every stream
Concealment-aware	• Error concealment R-D optimisation	Increases the error concealment performance

the coded video [47, 48]. Data partitioning relies on grouping the coded data into different categories depending on their relative importance. This can be used to ease the introduction of unequal error protection at the transmission layer. Redundant coding uses additional data blocks to enable correction of transmission errors and to achieve robust decoding in error-prone conditions. In case of the post-processing EC methods, the problem of robust transmission is addressed from the decoder-side. Finally, error concealment-aware techniques are used at the encoder, to either adapt the coding loop or introduce extra information, which is used to increase the performance of EC algorithms and increase the reconstruction quality of lost data at the decoder [49–52].

3.3 Localisation techniques

The basic principle of localization techniques is to break the predictive coding loop at certain predictive points, in order to reduce the likelihood of errors affecting large regions of the video signal, considering both the spatial and temporal domains. Since these techniques break the predictive coding, they have a negative impact on compression efficiency, as expected.

This group of error resilience techniques are able to reduce the error propagation at two levels, i.e., temporal and spatial levels. Figure 3.1 is used to demonstrate the principle of error propagation and how it can be reduced using localisation techniques. The top of the figure illustrates a lost block and the corresponding spatial and temporal error propagation on the left and right, respectively. Spatial localisation techniques are used to reduce the error propagation within a frame, where a row of refresh blocks can be introduced to stop the error propagation. To address the temporal error propagation, also refresh blocks (e.g., intra coded blocks) can be randomly introduced to break temporal dependencies and stop error propagation. As shown in Figure 3.1 the refresh blocks are not affected by error propagation. In both examples, the regions affected by errors are clearly reduced.

3.3.1 Error resilient rate-distortion optimisation

Common localization techniques deal with the error resilience problem within the scope of the R-D optimisation process used in video coding. Normally, such optimisation is equivalent to the minimisation of the Lagrangian cost (J) [53] given by:

$$J = D + \lambda R, \tag{3.1}$$

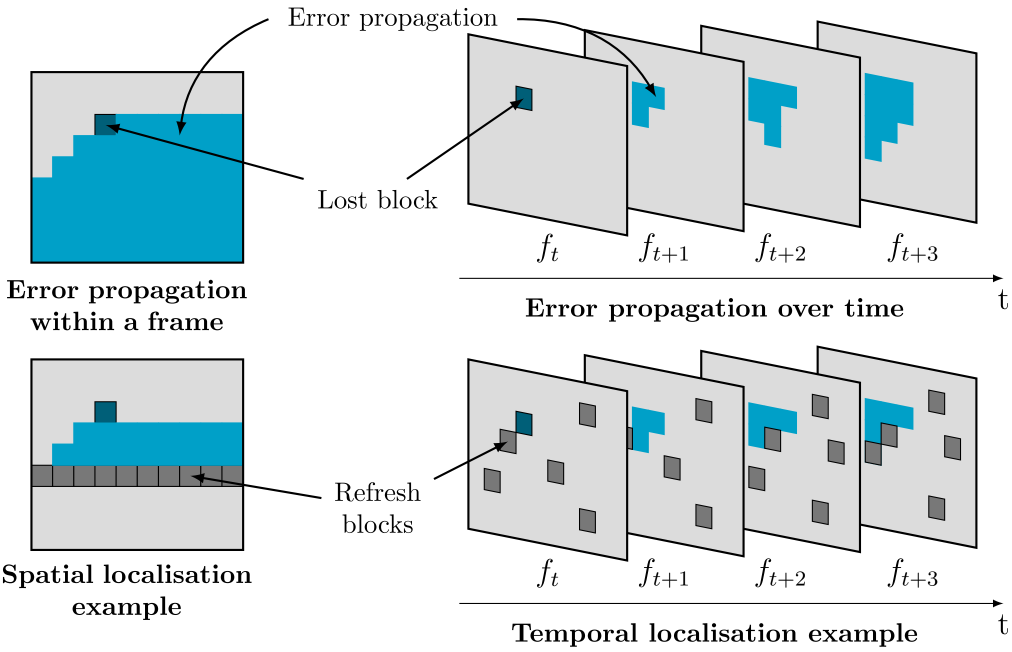

Error propagation
Lost block
Error propagation within a frame
f_t
f_{t+1}
f_{t+2}
f_{t+3}
t
Error propagation over time
Refresh blocks
Spatial localisation example
f_t
f_{t+1}
f_{t+2}
f_{t+3}
t
Temporal localisation example

Figure 3.1: Examples of techniques for spatial and temporal error resilience.

where D is the distortion of all pixels, R the bitrate and λ is the Lagrangian multiplier [54]. Such optimisation only takes into account the coding distortion (D), which is only suitable for environments without errors and data loss. However, some alternative approaches use error resilient R-D optimisation, by taking into consideration possible network losses [55]. Other advanced approaches can take advantage of a Recursive Optimal Per-pixel Estimation (ROPE) of the end-to-end distortion at the encoder-side [56]. This may be formulated as follows. Let f_t^i denote the original value of pixel i in frame at instant t, $\hat{f}_t^i$ denote its reconstruction in the encoder and $\tilde{f}_t^i$ the reconstruction in the decoder after EC. Since data loss may occur in the transmission channel, $\tilde{f}_t^i$ is modelled as a random variable at the encoder. Using the Mean Square Error (MSE) as a distortion metric, the overall expected distortion, $E\left\{d_t^i\right\}$, is given by:

$$E\left\{d_t^i\right\} = E\left\{\left(f_t^i - \tilde{f}_t^i\right)^2\right\} = (f_t^i)^2 - 2f_t^i E\left\{\tilde{f}_t^i\right\} + E\left\{\left(\tilde{f}_t^i\right)^2\right\}. \qquad (3.2)$$

The packet loss probability (p) is taken into consideration and used in the distortion measurement. In case of intra-coded pixels the following equations are used:

$$E\left\{\tilde{f}_t^i\right\} = (1-p)\,\hat{f}_t^i + pE\left\{\tilde{f}_t^i\right\}, \qquad (3.3)$$

$$E\left\{\left(\tilde{f}_t^i\right)^2\right\} = (1-p)\left(\hat{f}_t^i\right) + pE\left\{\left(\tilde{f}_t^i\right)^2\right\}, \tag{3.4}$$

while in the case of inter-coded pixels:

$$E\left\{\tilde{f}_t^i\right\} = (1-p)\left(\hat{e}_t^i + E\left\{\tilde{f}_{t-1}^j\right\}\right) + pE\left\{\tilde{f}_{t-1}^i\right\}, \tag{3.5}$$

$$E\left\{\left(\tilde{f}_t^i\right)^2\right\} = (1-p)\left(\left(\hat{e}_t^i\right)^2 + 2\hat{e}_t^i E\left\{\tilde{f}_{t-1}^j\right\} + E\left\{\left(\tilde{f}_{t-1}^j\right)^2\right\}\right) + pE\left\{\left(\tilde{f}_{t-1}^i\right)^2\right\}, \tag{3.6}$$

assuming that pixel i is predicted from pixel j in the previous frame. The prediction error e_t^i is quantised to the value $\hat{e}_t^i$, which is transmitted to the decoder along with the motion information. Equations (3.3) to (3.6) provide a measure of the distortion, taking into consideration not only the coding distortion but also the possibility of frame loss and consequent EC distortion.

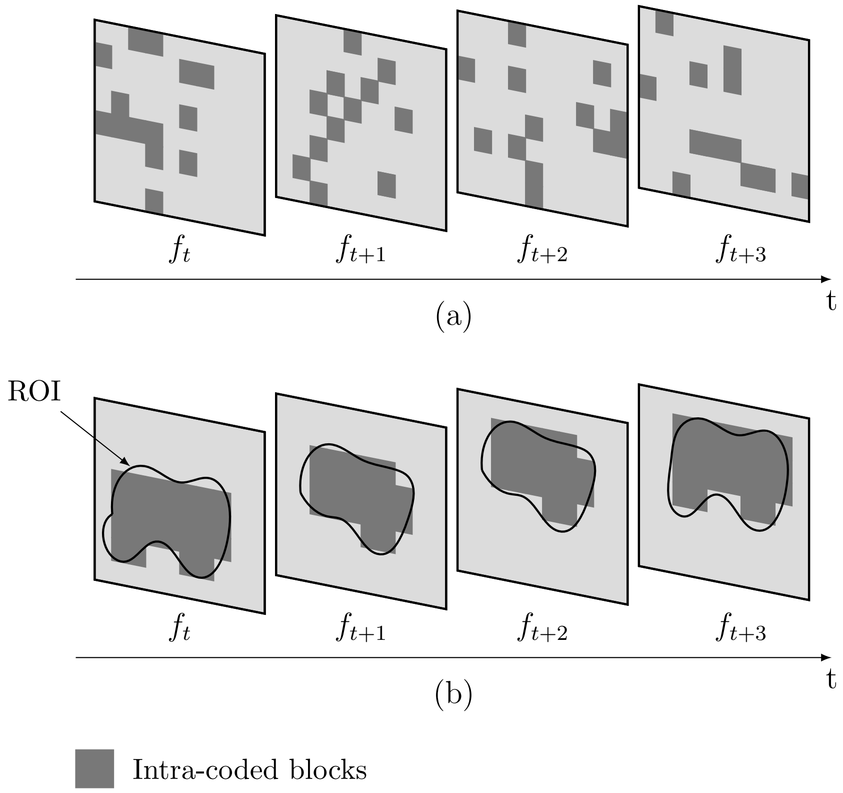
This method was used in [56] to switch between intra and inter coding modes with the aim of reducing the temporal dependencies, thus achieving enhanced error robustness in video transmission. In [57] this was optimised to quarter pixel estimation, and in [58] the same method was extended for burst losses, by introducing the burst loss probability into the original end-to-end estimation. The results obtained in [58] revealed that higher video quality was achieved by considering both probabilities. In [59] temporal EC is considered in the distortion estimation, which increases the performance of the proposed method, by avoiding frame-copy that provides a lower bound on reconstruction quality. More advanced temporal EC methods were also considered in the work presented in [60]. The MVs used for motion compensation reconstruction are obtained using boundary matching search, which is able to achieve lower distortion when compared with frame-copy. Using more advanced EC algorithms leads to a more accurate estimation of the end-to-end distortion, thus increasing the error robustness of the transmitted video. To take into account the delay introduced in the encoding loop due to hierarchical prediction structures, delay-based R-D optimisation was proposed in [61] by estimating both the distortion and the end-to-end delay that is minimised along with the required bitrate. Using this approach, the total delay can be controlled without severely compromising the coding efficiency. In the context of HEVC, end-to-end distortion estimation was also used in the R-D optimisation showing significant quality improvements when compared to with the reference implementation [47].

Different quality metrics can also be used for end-to-end distortion estimation. The method presented in [62] proposes to estimate the distortion by using the Structural Similarity Index Metric (SSIM) [63] in order to increase the correlation between the estimated distortion and the perceptual quality noticed by the human visual system [64]. Moreover, the SSIM quality metric was also used in learning-based R-D models [65], using a Convolutional Neural Network (CNN) to predict the end-to-end distortion and the required rate, showing good accuracy in the R-D cost estimation.

3.3.2 Intra-refreshing

In order to reduce the effects of error propagation within a single frame, spatial and temporal dependencies can be limited by introducing extra refresh points into the bitstream, which can be accomplished by forcing intra-coded blocks [66–73]. Figure 3.2 (a) shows an example where intra-coded blocks are forced at random locations (dark gray sqares) in order to stop error propagation. This technique is fairly efficient to improve the error robustness of video streams, by avoiding long error propagation not only within a single frame but also across several subsequent frames. The main characteristic of this group of techniques is to increase the number of intra-coded blocks as they have no dependencies to previous frames. However, this is done at the expense of using higher amount of coded bits [67]. Therefore, different alternatives have been proposed to optimise the number of intra-coded blocks, which can be introduced either randomly or periodically [70], based on the motion information or rate-distortion optimisation [74, 75]. Moreover, whenever possible, the encoder can use a feedback channel to decide where refresh regions should be dynamically introduced in the coded stream [66, 72].

A non-normative method for random intra refreshing is implemented in the H.264 reference software, which introduces cyclic refresh points using a random pattern. Figure 3.2 (a) shows an example of intra-coded blocks (dark gray squares) being randomly placed into the coded frames to prevent error propagation. This approach guarantees that all MBs are refreshed at some point, therefore errors do not propagate indefinitely. However, all MBs are refreshed exactly the same number of times without consideration for rate distortion optimisation neither the image region containing the MBs. In [74] a rate control scheme is used to decide whether a given block should be refreshed using intra-coded prediction. The decision is performed based on the ratio between the Lagrangian cost of the best intra and inter modes. Whenever this ratio is lower than a certain predefined threshold an intra mode is selected, which guarantees that intra refreshing is only forced when the Lagrangian cost of the intra mode is close to the



(a)

(b)

Figure 3.2: Examples of two intra-refreshing methods, where intra-coded blocks are placed (a) randomly or (b) inside a region of interest (ROI).

optimal coding method. This is further extended to adapt the intra refresh algorithm to the network conditions, i.e., the threshold value is obtained from the bitrate and the loss ratio [69]. Therefore, the number of intra-coded block automatically increases as the bitrate and the loss ratio increases. In the context of HEVC, intra refresh is also being used to reduce error propagation. Soft-refresh points using joint intra-inter prediction were proposed for this purpose [75]. Although the image regions cannot be correctly decoded in case of errors (as happens with intra-refresh), this method significantly reduces the bitrate increase due to extra intra-coded blocks. When compared with traditional intra-refresh, bitrate savings of up to 12% are achieved for a 5% packet loss.

A different approach is to introduce refresh points based on the video content. In [67] the error resilience of video streams is increased by combining rate control and attention-based models to define ROIs. A weighting factor is introduced in the conventional Lagrangian cost [53], to increase the number of intra-coded blocks in the image regions defined by the attention models. Figure 3.2 (b) shows an example with a ROI marked with a black line, which varies for every frame. In this case, the intra-coded blocks are only introduced inside the marked region, in contrast to what is shown in Figure 3.2 (a). Results show that by using the aforementioned method, less refresh points are required to improve the error robustness of the ROI. When comparing with [55], approximately 4 dB of Peak Signal-to-Noise Ratio (PSNR) gains are achieved in the ROI without compromising the overall video quality. A more recent approach was proposed using a perceptual noticeable difference model to estimate the noticeable distortion, and then using a visual attention model [76] to decide which blocks should be used as intra refresh in the bitstream [77]. This method is able to improve the results of a previous work [78], especially for high motion sequences. This is due to the fact that such previous approach only used an estimation of the error propagation without accounting for perceptual information.

A similar approach was also exploited for HEVC in [79, 80], by using a feedback channel to provide information about the network errors, which is then used by the encoder to introduce intra-coded blocks into the ROI. The simulation results reveal an increase on the video quality when higher number of intra blocks are used. Moreover, applying the proposed method only to the ROI, the bitrate overhead introduced is lower, leading to higher R-D gains.

3.3.3 Leaky-predictions

Directly using reconstructed past frames as reference frames for predictive coding, may result in substantial error propagation, whenever the reference frames are either lost or reconstructed from other distorted frames. To reduce error propagation one can employ leaky predictions, which scales down reconstructed frames before being used as reference [81, 82]. This leads to an exponential decay of error propagation, with direct impact on error robustness. In such approach, the reconstructed reference frames are scaled down by applying the following formula:

$$\check{f}_t^i = \alpha \hat{f}_t^i + (1 - \alpha)C, \tag{3.7}$$

where $\hat{f}_t^i$ is a reference frame, α is the leaky factor used to control the amount of information from the reconstructed frame present in the new frame $\check{f}_t^i$, and C is a constant value. Leaky prediction has been used in the past for error control, especially in scalable video coding [83], where C in (3.7) is replaced by a lower layer error-free reconstructed frame. In case of single layer coding, the use of a constant value has significant impact on the coding efficiency leading to higher bitrate, as shown in [84]. The α parameter allows to control the effect of the constant value, and thus the coding efficiency. By decreasing α, the coding efficiency increases, but the exponential decay

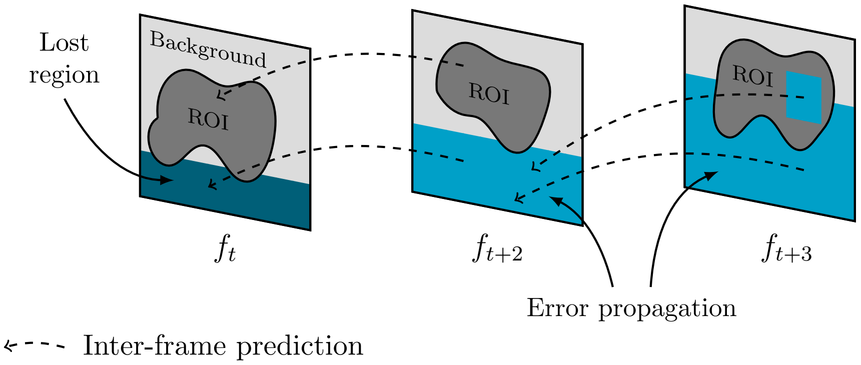


Figure 3.3: Example of error propagation across two different regions: errors from background affect the ROI.

in error propagation becomes less significant. To determine α, one can use network feedback information, such as packet loss probability p, resulting in $\alpha = (1 - p)$. Alternatively, empirical approaches using exhaustive search can also be used to find optimal values of α for each sequence [85].

In order to overcome the loss in coding efficiency, weighted prediction can achieve similar advantages, by using motion compensation from two previously encoded frames to generate a single prediction for the current frame [86]. Alternatively, the two previously encoded frames can be used to generate an interpolated frame ($\check{f}_t^i$) to be used as reference, using:

$$\check{f}_t^i = \alpha \hat{f}_t^i + (1 - \alpha)\hat{f}_{t-1}^i. \tag{3.8}$$

Therefore, it is possible to decrease the error propagation using $\check{f}_t^i$ as reference frame for prediction, if only one of the original encoded frames is lost (i.e., $\hat{f}_t^i$ or $\hat{f}_{t-1}^i$). This is due to the fact that $\check{f}_t^i$ contains information from both $\hat{f}_t^i$ and $\hat{f}_{t-1}^i$. In (3.8) $\hat{f}_t^i$ and $\hat{f}_{t-1}^i$ always correspond to the encoded frames and the generated frames ($\check{f}_t^i$) are not used to interpolate subsequent frames. This corresponds to a finite impulse response filter. In the case of burst losses where both $\hat{f}_t^i$ and $\hat{f}_{t-1}^i$ are missing at the decoder, the interpolated frame ($\check{f}_t^i$) suffers from severe drift, which significantly decreases its quality, because $\hat{f}_t^i$ and $\hat{f}_{t-1}^i$ are recovered using error concealment algorithms, i.e., they are not correctly decoded.

The method presented in [82] uses leaky prediction confined to a pre-defined ROI, in order to increase the error robustness with lower impact on the coding efficiency. Figure 3.3 shows an example of error propagation within two different regions: ROI and background. As show in this figure, since the ROI uses background regions for predictions, error propagation occurs across regions. To overcome this problem, the method in [82] uses a weighted prediction for the ROI region obtained from the following predictions: (i) a globally optimal prediction obtained by unrestricted motion estimation, (ii) a prediction obtained by restricting motion estimation within the ROI region. This limits the impact of errors in the background region to affect the ROI. Moreover, since the leaky predictions are only confined to some specific regions, higher coding efficiency is achieved when compared with other methods.

The method proposed in [85] extends the traditional leaky prediction expressed in (3.8) by replacing the previous encoded frame ($\hat{f}^i_{t-1}$.) with the previous interpolated frame ($\check{f}^i_t$). This results in the following equation:

$$\check{f}^i_t = \alpha \hat{f}^i_t + (1-\alpha)\check{f}^i_{t-1}, \tag{3.9}$$

which corresponds to an infinite impulse response filter. Comparing (3.8) and (3.9) for the same α, the latter implies stronger filtering than the former. Therefore, reference frames generated with heavier filtering have higher resilience against error propagation, but less correlation with the original frame, which negatively impacts inter predictions and hence coding efficiency. The idea of combining multiple predictions was also applied in the HEVC standard by combining it with end-to-end distortion to select the optimal predictions [87].

3.3.4 Reference picture selection

Reference Picture Selection (RPS) has been widely used in the past to reduce the temporal dependencies and reduce the error propagation [88–91]. In [88], an optimisation algorithm was devised to select between short and long term references based on the ROPE end-to-end distortion estimation method. This was used to increase the prediction distance and reduce error propagation. The same method can also be used in combination with feedback information from the network to improve the efficiency of RPS. In the decoder-side, an EC method was later used to take advantage of such prediction structure [92].

A network feedback mechanism was proposed in [89] to reduce error propagation at end-user decoders, by dynamically selecting the reference frame only among those correctly received. Figure 3.4 shows an example of an error resilience transcoding scheme, where the original temporal predictions (represented by the solid arrows) are transcoded to only use correctly received frames as reference, i.e., prediction represented by the dashed arrow. This is done based on network feedback. The length of the Group of Pictures (GOP) has several implications on the performance of this transcoding scheme. For instance the complexity increases with the GOP length, because the

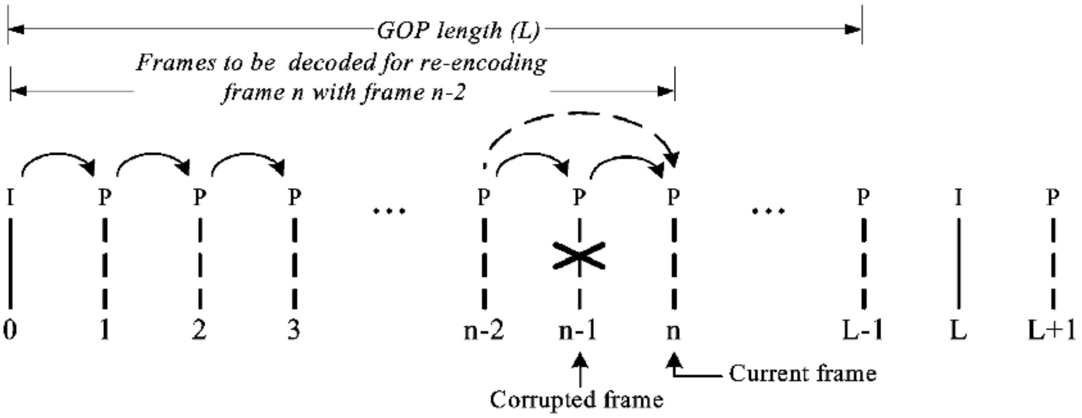


Figure 3.4: Examples of dynamic reference picture selection used in error resilient video transcoding [89].

number of frames that must be decoded also increases [89]. An absolute optimal GOP length was not found in [89], since it mostly depends on the application requirements (i.e., random-access, maximum refresh period, etc). A more advanced scheme may be considered, where only the regions affected by mismatch predictions are processed in the compressed domain, and subsequently transcoded to limit the dependencies from lost frames. To achieve low complexity, MV concatenation can be used, in order to obtain new inter-predictions for error-free frames [91].

In [93] network feedback is combined with path diversity to select the reference frames that are more likely to be received at the decoder. In such approach, the feedback information is not directly used to select the reference frames. Instead, the reference frames are selected through an end-to-end distortion estimation process, taking into account the coding distortion and distortion resulting from channel errors (EC distortion is also taken into account). In [48] a ROI is defined based on the moving objects to limit the regions where the RPS is applied. Intra refresh is also used, based on the temporal distance to the previous intra-refresh frame. This method not only achieves noticeable quality gains, when compared with the reference HEVC implementation, but also when compared with traditional reference selection schemes. In [68] an algorithm to introduce intra refresh points at the MB level was proposed. This method takes into consideration the decoding distortion, assuming that frame-copy is applied for EC when packet loss occurs. The refresh blocks are encoded using either intra or inter prediction through a secondary algorithm to select the most reliable reference frame (i.e., frame with less total expected distortion).

In the methods mentioned above, the dependencies between frames are analysed and reduced in order to decrease the error propagation, in case of errors or data loss.

Since the existing video coding standards are strongly based on predictive coding, these methods present an efficient approach in highly erroneous conditions. However, although these closed-loop techniques may offer a reliable control of the decoder distortion, the feedback required from the receiver side and multipath networking are neither always available nor feasible in real-time and unidirectional channels (e.g. broadcast).

3.4 Data partitioning and redundant coding

In compressed video streams different symbols have different relevance in terms of their impact on error propagation. Therefore, there is an unequal contribution of different syntax elements to the decoding distortion and error propagation, depending on the symbols that are lost in each case. Data partition techniques have been developed to provide a mechanism to cope with the unequal importance of coded information by grouping different parts of a compressed bitstream, according to their importance to the decoded quality. Thus, some of these groups can be protected more efficiently than others using Unequal Error Protection (UEP) schemes. These are normally associated with Forward Error Correction (FEC) mechanisms, which are typically applied at the channel level without taking into consideration that not all video data is equally important. This leads to a sub-optimal error protection [94]. In general, when using UEP, source-channel optimisation is applied, whereby multi-levels of FEC are used to achieve different levels of error protection according to the importance of coded video data. This is presented and discussed with more detail in Sub-section 3.4.2

In single channel transmission, the most important data or syntax elements (e.g., headers) can be protected with stronger channel codes than the remaining ones. In case of multi-channel communications, the important data can be routed over the most reliable channels. Moreover, to achieve higher quality in multi-channel transmission, one can use extra information to increase the error resilience of video streams, in order to increase the probability of being able to decode them in case of errors. For instance, redundancy can be added either using explicit (e.g., redundant picture) or implicit (e.g., multiple description coding) mechanism. Section 3.4.3 and 3.4.4 present a detailed review of these two approaches, respectively.

3.4.1 Methods for data partitioning and macroblock ordering

The concept of data partitioning was firstly introduced in the H.263 coding standard to group the coded data into different categories [31]. Then, it was further improved in H.264/AVC, allowing the symbols in each slice to be divided into three partitions:

- Type A: headers, MB types, quantisation parameters and motion information.
- Type B: intra-coded information, such as, intra blocks signalling and intra coefficients. This type cannot be used without the partitions of type A being present.
- Type C: inter-coded information and inter coefficients. This type also requires the information of type A, but can be decoded without intra-coded information (type B).

These partitions have different levels of importance, according to the type of information carried in each one. Data partitions of type A are the most important as they carry the header information that is required to decode the remaining partitions in the video stream. Then, since partitions of type B carry intra-coded information, they are more important than type C, because they can stop temporal error propagation.

An error resilience tool named FMO technique was introduced in H.264/AVC, allowing each MB to be freely assigned to a specific slice group (i.e., a set of slices), organised in a mixed-up fashion [95]. It supports up to eight slice groups in each picture, and the MBs within a group can be assigned to several slices in the default raster scan order. The case where only one slice group is used corresponds to the case where the FMO tool is disabled. FMO increases the error resilience by partitioning the bitstream into several groups, which is able to reduce the error propagation. Simulation results show that the use of FMO contributes to enhance the subjective and objective quality up to 3 dB of PSNR for 5% of packet loss, at cost of an affordable overhead [95].

Six types of slice groups are supported by FMO. Figure 3.5 shows FMO types 0 to 5 only, since the type 6 has no defined pattern. In the FMO type 0 each slice group is defined by a maximum number of MBs, and a new slice group begins whenever his maximum value is reached. The same slice group can be cyclic repeated within a frame. FMO type 1 corresponds to scattered slices, where the MBs are assigned to each slice group with a fixed formula. Figure 3.5 represents the FMO type 1 with two slice groups resulting in a checkerboard pattern. In the FMO type 2, slice groups are defined by rectangular shapes bounded by the top-left and bottom-right corners. Although the slice groups are allowed to overlap each other, each MB may only belong to one slice group. Similar to FMO type 0, in the types 3, 4 and 5 in Figure 3.5, slice groups are defined by a maximum number of MBs. However, contrary to FMO type 0, the slice groups are allowed to change every frame for types 3, 4 and 5. In order to keep the signalling overhead affordable, only two slice groups are allowed in these types. The

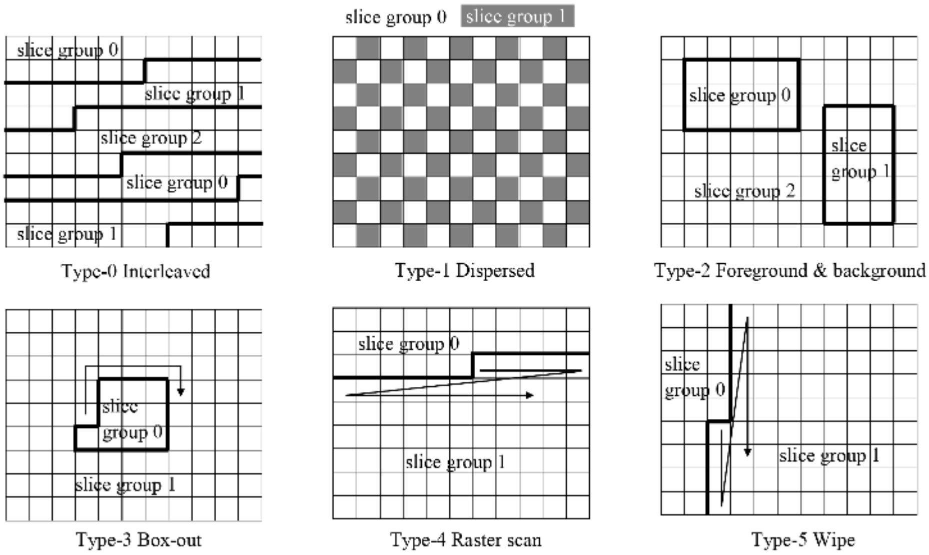


Figure 3.5: Different types of flexible macroblock ordering [95].

FMO type 6 is the most generic case, where the slice groups are explicitly defined using high level packets (i.e., signalling NAL units), allowing more flexibility at the cost of higher signalling overhead.

In the past, FMO has been widely used to increase the error robustness of video transmissions in different communication scenarios [96–100]. In [96] FMO is used to increase the EC performance and thus reduce the error propagation in prioritised video transmission. The method divides the MBs into two slice groups based on an estimation of the EC distortion using a two-pass approach. Then, the slice group which has the most impact on the EC distortion is transmitted using a reliable channel, and it is assumed to be received without errors. The experimental results shows that the proposed approach is able to outperform the dispersed mode (FMO type 1) [96]. Slice classification can also be based on the weighted sum of the motion vector information and the rate-distortion cost [97]. As in [96], the MBs are ranked based on their importance and distributed in the two groups with the same number of blocks. In order to reduce the mapping complexity of the MBs into groups, the same map can be repeated for several subsequent P-frames [98]. This reduces the computational complexity of the MB grouping while achieving similar error resilience performance. The method in [99] splits the video frames into two regions, which are then grouped in different slice groups. To classify the different MBs, spatial information is extracted from the MB bit allocation and temporal information is extracted based on the estimation of the EC distortion. This technique is able to reduce the number of erroneous MBs up to 82%, achieving a quality gain of up to 3.97%, in terms of PSNR, in comparison with the case where FMO is not used [99]. Motion energy was also used to classify MBs into different groups [100], by comparing the motion energy of a given MB with the motion energy of its neighbours and the average motion energy of the entire frame. This allowed the authors to obtain a dynamic threshold which is then used to determine the MB importance. In comparison with other UEP error protection schemes, this method achieves good quality improvements under different Packet Loss Ratios (PLRs).

In [101], a new FMO type is proposed combining the slice groups of type 1 and 3 in order to obtain a checkerboard pattern in a box-out scanning order, with the objective of increasing the scattering of lost MBs. Results show that the proposed FMO type is able to outperform the existing types under single and burst loss events. The method described in [102] introduces a new FMO type in order to efficiently divide the video content of each frame in two regions with different levels of importance, based on the FMO type 6. By proposing a new FMO type they are able to signal several ROI areas achieving lower overhead than using the original FMO type 6.

3.4.2 Unequal error protection

UEP techniques are used to protect different parts of a video stream against errors according to their importance for the decoding quality. UEP approaches take advantage of the video stream characteristics that result from different types of content, both at the temporal and spatial levels. This allows categorisation of coded data into different groups, by exploiting the fact that different regions are affected differently by transmission errors. This not only ensures that more important regions will be delivered with higher video quality, but also leads to efficient use of the total overhead required to achieve increased error robustness in a video transmission.

A wide variety of UEP schemes have been presented in the literature [94,103–105] to optimise the performance of video communication over networks where the probability of packet loss is not negligible. It has been observed that, within a GOP, different frame types (e.g. I, P, or B frames) have different levels of contribution to the reconstructed video quality, when transmission losses occur. This has been exploited for UEP in video transmission, as shown in [94,103]. In [104] an evaluation of layered video was proposed to increase the error robustness, by using redundant packets for helping the recovery of lost ones. In order to achieve reduced overhead for error protection, each video layer has a different importance and is protected using different levels of redundancy, thus guaranteeing that packets carrying data of lower layers have higher probability of being decoded without errors. This is a simple approach, not depending on the network capabilities, neither requiring a deep knowledge of the video contents. A lightweight model is used to assign packet priorities to be used by UEP algorithms [105]. Such model is based on a weighted sum of four components: (i) a slice type parameter giving higher importance to refresh and reference frames; (ii) a binary flag indicating whether a network packet transports video stream headers, i.e., NAL or slice header; (iii) a parameter indicating the distance in packets to the next slice; (iv) a parameter indicating the temporal distance to the next refresh frame. These parameters can be obtained by parsing the NAL header, which makes it suitable for low-end decoders or even prioritising each packet at a streaming level.

However, UEP techniques usually require a deep video analysis, which in general contributes to increase the encoding complexity. Nevertheless, one can use information already needed for encoding, such as motion, to classify different frames or even different regions within a frame [94]. The motion information is used to determine the motion energy, which is given by the L2-norm of the MVs within a predictive regions. For instance, in [106] the motion energy of each frame is computed and compared with neighbouring frames. The energy of co-located block in adjacent frame is compared, and higher importance is given to those frames that have higher motion energy differences. This method was compared with equal error protection, resulting in higher objective and subjective quality and revealing the relation between motion information and impact on errors. This work was extended in [107, 108] to use a two-level UEP approach, where besides taking into consideration the MB and frame importance, the combined importance of different frames is also used to determine the level of importance of each GOP.

Another work dealing with UEP is described in [109], where the transmission distortion is estimated at the packet level, to assess the impact of each packet in the error propagation. In this case, different priorities are applied to each packet, allowing different levels of error protection. This was also used to schedule the packet transmission through multi-path networks [110]. Furthermore, in order to improve the classification of different regions of video streams, a perceptual algorithm can be used. This strategy increases the correlation between the classification and subjective impact of errors. In [111] a method based on a combination of two parameters is described: a spatial parameter, based on a perceptual attention model, and a temporal parameter, based on the frame position within the GOP, which are used to improve the efficiency of FEC assignment. Experimental results show a significant quality improvement, both using objective metrics and subjective evaluation.

The scalable video coding framework provides implicit support for UEP through the layered structures, defined with different levels of quality [112–115]. In [113], an algorithm to recursively estimate the total expected distortion in the decoder at the picture-level for each layer is proposed. This method takes into consideration Enhancement Layer (EL) truncation, i.e., a specific layer being discarded, and EC to measure an accurate estimation of the end-to-end distortion. The estimation is carried out for every layer in the video stream, in order to apply an optimisation of different levels of error protection across several layers. Also based on scalable coding, the work in [114] proposes a technique for finding the optimal FEC coding rates that leads to correctly decode the EL when errors affect the base layer. This prevents discarding correctly received ELs frames due to missing dependencies, increasing the overall video quality.

In the context of HEVC, there are also methods to achieve UEP. Since techniques like FMO are not available, different approaches were investigated [116–119]. The work in [116] proposes a resource allocation scheme based on prioritisation at the slice level (i.e., NAL unit level) for Long Term Evolution - Advanced (LTE-Advanced) networks [117]. The prioritisation is performed by analysing the inter-frame predictions dependencies, giving higher rankings to slices having higher number of dependent pixels. In [118] a quality degradation model is proposed for HEVC video, by simulating error events and measuring the quality using the PSNR metric. From empirical results, a model was devised to predict the quality under different events. Then, the model is used to control the amount of redundant codes to be applied at the network level. The scalable extension of the HEVC standard was also used to support UEP for video storage in cloud environments [119], by combining the layer information (i.e., base or enhancement layer) and the picture type to maximise the performance of the FEC codes, both at the temporal and quality/resolution levels.

3.4.3 Redundant pictures

A well-known technique adopted in the H.264/AVC standard is the use of redundant picture, which allows for different representations of the same coded picture to be transmitted using different encoding parameters [120]. For example, the primary picture may be coded with fine quantisation, while the redundant one may use coarse quantisation resulting in a lower bitrate. On the one hand, if the primary picture is received, the redundant information is discarded. On the other hand, if errors occur and the primary picture is lost, the redundant one is used to provide a lower level of reconstructed quality and limiting the error propagation.

Zhu *et al.* investigated the paradigm of redundant picture coding [120], considering

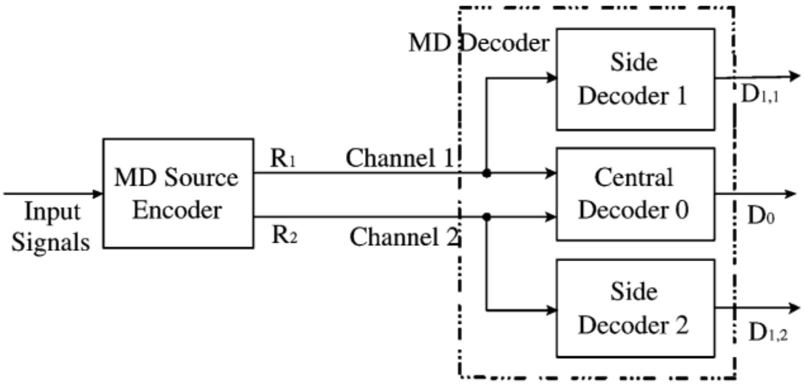


Figure 3.6: Architecture of multiple description system [124].

different approaches to developing a mixed method based on both RPS to obtain a clean decoding, without error propagation, and hierarchical allocation of the redundant pictures to increase error robustness. In [121] the key frames are encoded and transmitted at a lower bitrate to be used in case of error, to replace the missing frame. Redundant pictures are a good mechanism to improve overall video quality under error prone networking conditions, since they introduce secondary pictures that can be decoded when the primary picture is not available.

Moreover, redundant pictures might also work as localisation techniques when different reference frames are selected. As the number of redundant pictures increases, the amount of mismatch temporal predictions in case of data loss is reduced, since the redundant picture can replace the erroneous ones. This achieves a similar goal as reducing the video stream dependencies, i.e., localisation techniques. The main drawback of such approaches is the large amount of redundant data and overheads that are required for their efficient implementation. In [122] a newer approach was proposed to reduce the amount of information needed to encode redundant pictures, by introducing only redundant MVs. This was also used in [123], where the motion information was used to recover the missing frame. These approaches increase the overall system performance by presenting an efficient trade-off between the redundant bitrate and the error robustness.

3.4.4 Multiple description coding

In Multiple Description Coding (MDC) schemes two or more bitstreams, named descriptions, are encoded and transmitted through different channels, such that an acceptable level of quality is achieved when only some of them reach the decoder [124,125]. Figure 3.6 illustrates a basic architecture of a MDC encoder and decoder using two descriptions. In this example, the encoder-side creates two descriptions, i.e., R_1 and R_2, which shall be transmitted using two different channels, reducing the probability of the same video data being lost at the same time in both descriptions. As shown in this diagram, three different outputs may be produced at the decoder-side: both descriptions are correctly received and the central decoder is used (D_0) or, either one of the descriptions is not received and then only one side decoder is used ($D_{1,1}$ or $D_{1,2}$). When both bitstreams are received, the central decoder combines the information carried by each description achieving higher quality than any single side decoder.

In MDC encoding, the redundancy among descriptions can be controlled by changing the correlation between them. Higher correlation leads to higher redundancy and also higher robustness, also leading to higher quality in the output of the side decoders. In contrast, with the case of redundant pictures, the redundant MDC streams do not carry the same information, which means that if both are received, then the decoded quality may be increased. However, a combination of MDC and redundant pictures can also be used to achieve higher quality gains [126]. Different types of MDC approaches [127,128] can be used to improve the quality under transmission errors, either by taking advantage of hierarchical B-pictures [127] or by introducing extra auxiliary information to limit the error propagation in case of single description decoding [128]. Although these techniques can be very effective in multi-path networks, a significant amount of overhead is required to create the different descriptions [129].

3.5 Error concealment by post-processing

In general, EC functions are implemented as a post-processing step in the decoder, thus they are the last module in the communication chain capable of handling errors [130]. In this case, reconstruction of missing data should take place within the video decoder to minimise the perceived effects of lost data. This section reviews several EC methods based on spatial and temporal redundancies, as these play an important role in the overall performance of robust video communication systems.

3.5.1 Spatial error concealment techniques

A typical approach to recover missing regions in an image is to estimate the lost regions based on linear interpolation of neighbouring pixels [131, 132]. Figure 3.7 shows the estimation process of a lost pixel (lp), within the lost region (R_L), using the four closest pixels (p_i) as reference, and the distance between the lost pixel and the reference ones

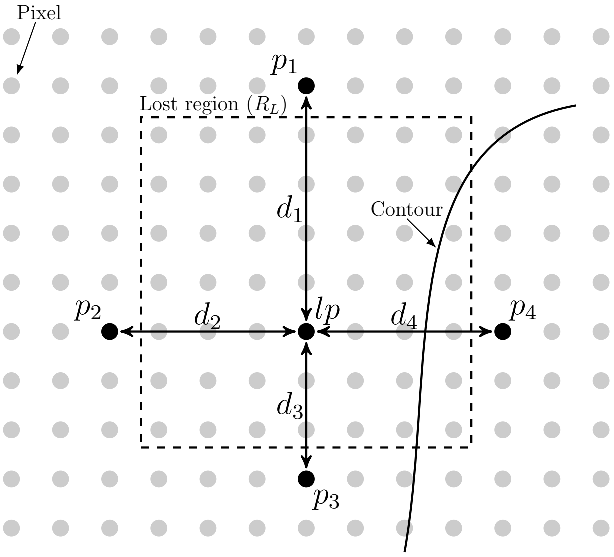


Figure 3.7: Lost pixel interpolation from neighbouring pixels correctly decoded.

(d_i). The pixel value (lp) is recovered using a weighted average of n neighbouring pixels, defined as follows:

$$lp = \frac{\sum_{i=1}^{n} \frac{p_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}}, \tag{3.10}$$

where $n = 4$ in the example of Figure 3.7. Liner interpolation is a simple yet effective method to recover lost regions in digital images. However, as each lost pixel is recovered by giving the same importance to each reference direction, the image discontinuities are not taken into account. Thus, some pixels used as reference may belong to distinct regions of the image, without significant correlation with the lost region itself. This leads to a poor estimation of lost pixels, resulting in noticeable artefacts in the reconstructed image. For example, in Figure 3.7 the lost pixel (lp) and the pixel p_4 belong to different regions separated by an edge. Therefore, p_4 may not be suitable to recover the lost pixel value. So, to improve the estimation accuracy of the lost region, the edges and image transitions must be considered, by selecting correct pixels to be used in the interpolation function. To overcome this problem, directional interpolation can be used, by recovering the lost pixels based on those within the same region, according to the edge directions.

Directional interpolation improves the accuracy of the pixel recovery, by firstly estimating the lost edges, using either correctly received or previously recovered pixels. To obtain the best interpolation direction, edges are estimated using the image gradient, for example applying the Sobel operator [133]. This operator is an efficient algorithm to detect edges, as it can reach extreme values at the image edges. The usage of this operator for EC was evaluated in [134–138], and it is characterised as a fast and efficient method due to its small kernel [139].

3.5.2 Temporal error concealment techniques

The conventional methods used to recover missing frames in the temporal domain are the so-called Frame-Copy (FC) and Motion-Copy (MC). These EC methods were implemented for the H.264/AVC reference software in [140]. The FC method consists in repeating (copying) the frame that has the shortest temporal distance, in order to substitute the missing one, into the time instant where the loss occurred. The MC method uses the motion information of the closest frame to recover the missing one, thus the missing pixels are recovered through motion compensation. Previous works and simulation results showed that the MC method achieves higher quality in the reconstructed frames in most cases [24, 141]. However, the complexity increases with the implementation of the MC method, when compared with FC.

In order to improve the motion-copy performance, two reference frames can be used as sources of MVs to reconstruct the missing one [141]. The average value of two vectors is used to reconstruct the missing motion information. Results in [141] show quality gains up to 4.99 dB (PSNR) when compared with MC method. In [24] the MC method performance was increased by using a refinement algorithm to improve the accuracy of the MVs obtained from the closest decoded frame. The MV differences between the vectors of successive frame pairs are recursively calculated along with the next refinement, until the changes become negligibly small (e.g., close to zero). This approach has shown to improve the performance of the MC method, with gains up to 0.47 dB (PSNR) for a PLR of 20%.

Boundary matching algorithms

Boundary Matching Algorithms (BMA) are used to recover a missing region based on the available neighbouring regions. By using the missing block neighbourhood, a new MV can be estimated for the missing block. Then, the recovery of the missing pixels is accomplished based on motion compensation. A low complexity approach may use the boundary pixels to search for the MV, based on a spatial smoothness constraint imposed on the boundaries of the lost block [142]. Figure 3.8 represents the boundary information used to estimate the MV for the lost block from a set of different candidates. The BMA uses the available MVs provided by the top (T), bottom (B), left (L) and right (R) blocks to simulate different reconstructions of the lost block. Then,

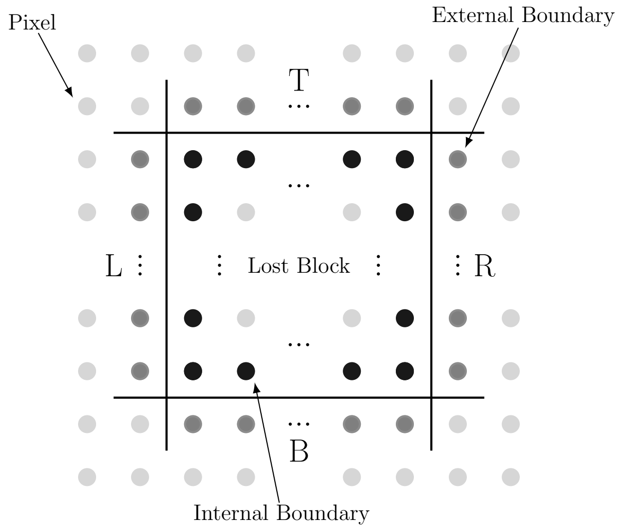


Figure 3.8: Illustration of the lost block and the neighbouring information used in boundary matching algorithm.

the optimal reconstruction is selected according to the sum of differences, between the pixels on the internal boundary in the recovered block and the corresponding ones on the external boundary (see Figure 3.8).

An implementation of the BMA was proposed in [143], by analysing the change of magnitude and angle between the MVs of the neighbouring blocks and the co-located ones in the previous frames. The average change is applied to the MV of the block co-located with the lost one to compute the MV candidate for recovery. An improvement to BMA was developed using the outer boundary matching algorithm [144], also known as decoder MV estimation algorithm in [145]. In this method the distortion is calculated by the difference of two outer boundaries of reconstructed block, instead of an internal and external boundary in BMA. Although BMA and outer BMA have the same design principle of minimising the mismatch distortion, in the latter the distortion is measured in the outer boundaries, *i.e.*, external boundaries of the reconstructed block. It was observed in the experimental results of [144] that the outer BMA has better performance in video recovery than conventional BMA. The performance improvements are explained by the distortion minimisation of the outer boundary, which incorporates the edge information located in the neighbourhood of the lost block in the decision process to select the most accurate MV. In [142] a novel distortion function was proposed, to choose among different candidate MVs. The distortion function exploits both spatial and temporal smooth constraints. The temporal function is based on the outer boundary distortion, as in [144]. The spatial distortion aims to reduce the gradient variances in boundaries of the reconstructed block, so the image edges may be guaranteed. This method is able to achieve higher quality in the reconstructed frame when comparing with the BMA method. However, although subjective analysis revealed that this method is able to recover the edges in the lost image, complex mathematical operations are required, which may not be practical in real time applications.

Motion vector extrapolation

As mentioned before, the MC method is a simple way to reconstruct an estimation of a missing frame using motion information from previously decoded frames. Although this method can provide an acceptable performance comparing with the FC, as shown in [140], more complex algorithms that use the assumption of motion continuity across consecutive frames are used to recover the missing information.

Motion Vector Extrapolation (MVE) [146, 147] methods are efficient techniques to reconstruct the MVs of a lost frame, which in turn are used for motion compensation. These MVs are extrapolated from the last decoded frame (reference frame) onto the missing one [146]. Each block in the reference frame is projected to the missing one with its own direction given by the corresponding extrapolated MV (mv'). An example of MVE is shown in Figure 3.9, where the vectors of frame $t-1$ are projected onto the missing frame at instant t. Note that, in the reference frame (f_{t-1}) the MVs may point to different frames (f_{t-2} and f_{t-3}), previously decoded, so MVE must take into account the temporal distance between different frames. To improve the accuracy of MVE, the damaged region is divided smaller blocks and, according to the matching area between the extrapolated block and the damaged one, the best MV is chosen. In other words, the MV for each block is obtained from the extrapolated block with the highest number of pixels matching the missing locations. Note that the final motion information need to be aligned to a fixed size grid, e.g., 4×4 grid, which leads to a misalignment of the extrapolated MVs. Alternatively, in [148] several blocks are extrapolated to the missing frame, resulting in overlapped regions, which are recovered using a weighted average of the multiple pixels candidates. The weight coefficients assume a maximum value in the center of the extrapolated region, and a minimum value at the boundaries. An extension to the previous approach was considered in [149], by using the MVs of subsequent available frames. Thus, to estimate the lost motion information, forward and backward extrapolations are performed, in order to obtain two candidates to each lost pixel. Then, the lost pixels are obtained through motion compensation using the MVs determined by the average of two MV candidates. The experimental results revealed that the bi-directional Pixel-based Motion Vector

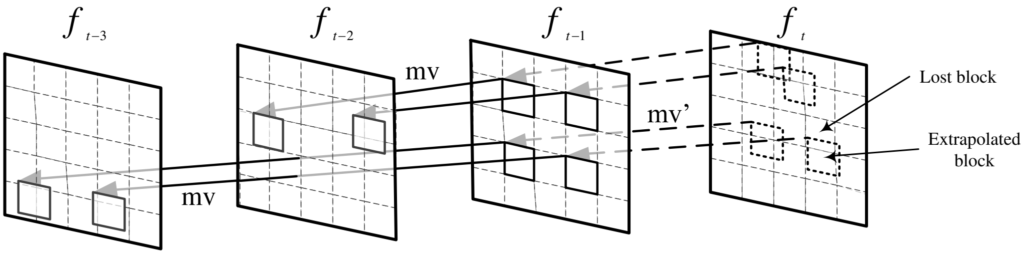


Figure 3.9: Motion vector extrapolation with the corresponding extrapolated block, based on the method presented in [146].

Extrapolation (PMVE) outperforms the isolated use of forward or backward PMVE. Moreover, the bi-directional PMVE achieves gains up to 1.85 dB and 1.32 dB (PSNR) comparing with the work in [150] and MVE [146].

In [151] a Hybrid Motion Vector Extrapolation (HMVE) algorithm was proposed, to improve the accuracy of the MVs of the PMVE method, by using the extrapolated vectors at the pixel and block level. Moreover, the MVs that were wrongly extrapolated are discarded in order to obtain a more accurate MV. In this EC method, firstly a pixel-based motion extrapolation is performed and the pixels are organised into three groups. Then, through block-based MVE two new MVs are determined: the first one corresponds to the MV of the most overlapped block with the current one, and second one is obtained by the weighted average value of the MVs corresponding to different overlapped blocks. To discard the wrongly extrapolated MVs and finding the most accurate ones, further refinement is applied based on the distance between each pair of MVs. This algorithm is able to outperform the PMVE for different packet loss probabilities, as shown by the results in [151]. The method was used to recover both intra and inter frames in video sequences compressed with the H.264/AVC, achieving PSNR gains above 1.09 dB in the erroneous frames.

Previous works presented for the HEVC standard are mainly based on temporal EC approaches. For instance, in [25] the block partitions of the neighbouring frames were used to improve the performance of MVE. The reconstructed frame keeps smooth boundaries, leading to higher objective and visual quality. Moreover, in order to select only the reliable motion information, the residue information can also be used to decide the optimal MV, as shown in [152].

Overall, the aforementioned methods rely on post-processing without any assistance from the encoder-side, which usually leads to inaccurate recovery of the lost information. Thus, in order to achieve higher reconstruction quality, further information needs to be shared with the decoder to minimise the residual distortion of EC.

3.6 Error concealment-aware error resilience

The last category of techniques for robust video transmission addresses those algorithms that help the EC process at the decoder-side. Although such techniques might overlap with some of the concepts described above, such as, data partition or redundant information, they have the main goal of improving the EC performance. This class of algorithms combine the error resilience capabilities with EC methods to achieve an improved overall video quality, when comparing with individual EC methods.

One of the approaches that can be used to improve the EC performance is to estimate, at the encoding side, the EC distortion introduced by the decoder in case of data loss. In [49] the EC distortion estimation is used for slice classification and UEP. Such distortion is also used to optimise motion estimation [50], by including the distortion of the subsequent frame (assuming it is lost and recovered) in the R-D cost optimisation of the current frame. This was also used in the context of HEVC to improve its error robustness, as shown in [47]. This kind of approaches take advantage of location techniques and data partitioning to improve the reconstruction quality achieved by EC algorithms. Normally, they use a single EC method that is implemented both at the encoder and decoder, based on temporal algorithms, e.g., MC and MVE.

Another approach is the use of EC MVs, which are vectors that might be carried by intra-coded units for the purpose of error recovery. When an error occurs in a given coding unit, the EC MV from the neighbouring units can be used to form a prediction from previously received frames. This concept was exploited in [51] by using data-hiding to carry additional motion information in the DCT coefficients. This improves the reconstruction quality, but transfers the computational complexity from the decoder to the encoder.

A different approach to improving the EC performance was presented in [45]. Pixel interlacing is used to divide the frame into different packets in order to enhance the performance of intra-frame EC by using linear interpolation. This technique reduces the probability of burst losses, therefore, the neighbouring pixels surrounding the lost regions are available to be used as reference for the concealment operation. These methods improve the EC performance, either through slice reordering or based on MVs, but they always rely on the assumption that a fixed EC method is used at the decoder. In [153] a low-resolution video is transmitted alongside the main bitstream in order to be used as reference for EC. Such redundant stream is used by the decoder to improve the estimation accuracy of motion information. This method was extended in [154] by using asymmetric coding of the redundant information based on Itti's attention model [76].

The work presented in [52,155] proposes testing several methods with transmission of extra symbols to indicate the best method to be applied by the decoder, using a ROI to identify the most sensitive image area. Although such an approach achieves reasonably good performance, it relies on the spatial neighbouring information, which may not be available in case of full frame loss. Moreover, it requires transmission of the ROI map to the decoder, which increases the overall bitrate. More recently an EC signalling methods was proposed for scalable video [156], in order to select the best frame to replace the missing one.

Summarising, EC-aware error techniques allow the combination of efficient algorithms, e.g., redundant pictures, EC-aware motion estimation or embedded information, with algorithms at the decoder that take advantages of this extra information. Such techniques are able to achieve quality improvements in comparison with error recovery approaches at the decoder on its own. Although effective techniques have been proposed in the past, the exploration of dynamic techniques, which explore the potential of different EC techniques with adaptable decoder, have not been fully exploited.

3.7 Summary

This chapter presented an overview of the different error resilience techniques proposed over the last years, based on their main characteristics. Different approaches were described covering techniques for robust video transmission, like dependency reduction at the encoder-side and EC, by post-processing at the decoder. The techniques described in this chapter provide relevant insights for the methods investigated in this work. Different algorithms were considered and combined to increase the error robustness of highly efficient coded video and improve the reconstruction quality in case of transmission errors.

Chapter 4

Error robustness of HEVC: evaluation and improvements

4.1 Introduction

The use of HEVC in different type of application and video services over networks that are prone to errors and packet losses, requires mechanisms to minimise the effect of such probabilistic events. This chapter starts by presenting an evaluation study about the error resilience of HEVC in Section 4.2, with the aim of defining the weak elements of coded streams and the level of error robustness under different transmission loss conditions. The evaluation study covers different coding configuration to evaluate the individual impact of some new coding tools introduced in this standard, e.g., increased partition size, intra-refreshing, slice partitioning and advanced motion vector prediction.

Then, in Section 4.3, based on the results of the simulation study, a novel method to increase the error robustness of the MV coding is presented and compared to existing techniques. The proposed method introduces refresh points in the motion prediction used in the HEVC, which significantly reduces the impact of mismatch MV predictions. Finally, Section 4.4 concludes the chapter.

4.2 Evaluation study of the error robustness

4.2.1 Motivation and objectives

As mentioned before, the HEVC standard has improved over its predecessor in terms of coding efficiency, at cost of higher complexity and data dependencies, mostly resulting from advanced prediction structures. Although the increased complexity may be compensated with more efficient algorithms and faster hardware technologies, the impact of higher coding efficiency on reducing error robustness strongly affects the video quality under transmission losses.

A critical aspect related to the performance of robust video streaming methods is the error detection accuracy. Since HEVC streaming is expected to use known transport technologies, such as RTP [16], MPEG-2 TS [17], or more recent standards as MPEG-DASH [18] and MPEG Media Transport [20], the error detection capabilities of the underlying transport protocols have significant impact on the overall system performance. In general, the error detection operation at the transport layers is decoupled from the video coding layer. To cope with data loss, the encoder implements robust coding methods, while the efficient recovery of lost video data is implemented in the decoder, through EC algorithms. Interlayer communication must be used to cope with different protocol layers, ensuring that the video decoder is given the necessary information to identify the lost slices/frames in the video stream. Therefore, the error robustness of video streams can be investigated on its own, assuming any possible form of streaming/transport technologies and error detection. This is the main focus of the study presented in this section.

There are some studies presented in the past regarding the error robustness of the HEVC standard, which are worth to be mentioned as they provide relevant insights for the research carried out in this work. In [12,157] the coding efficiency and error robustness of the HEVC are evaluated and compared against the H.264/AVC. Experimental results confirm that HEVC has reduced error robustness despite its increased coding efficiency. A subjective evaluation study is presented in [13, 14] revealing that packet loss ratios higher than 3% have significant effect on the perceived video quality. The vulnerability of MV prediction is investigated and its impact on the error robustness is evaluated in [158]. Moreover, in the same study, a method to introduce MV refresh points is proposed, achieving average PSNR gains of approximately 3 dB. Summarising, the decrease in error robustness is mostly due to the strong data dependencies imposed by the highly complex prediction modes of HEVC. Since these are selected according to a rate distortion optimisation criterion that assumes packet loss-free transmission, the decoder faces different problems whenever errors or data loss occur in the transmission network [15].

The following sub-sections present an error robustness evaluation of HEVC focused on the impact of the novel coding techniques introduced in the standard, extending the generic studies presented in previous works. This study provides relevant insights for research directions and motivation to investigate solutions for the error robustness problem in HEVC. The following case-studies are covered in this experimental evaluation:

1. different distances between intra frames, to evaluate the impact of introducing different number of refresh points;
2. different maximum CU sizes to evaluate the impact of larger frame partitioning;
3. different slice partitioning schemes;
4. different MV coding schemes using different types of MV candidates, with analysis at the spatial and temporal levels.

4.2.2 Characterisation of the experiments

The experimental setup used in this study to evaluate the error robustness of the HEVC is based on seventeen test sequences, which were chosen to cover different types of motion and texture complexity. Table 4.1 presents a summary of the main characteristics of the test sequences. Appendix A shows a single frame of each test sequence. Considering a test sequence with N frames $f(t)$ the pixel values (Luma component) of the frame at instant t, the spatial and temporal complexity of the video sequences was measured by using the following metrics [159]:

- Spatial Information (SI) based on Sobel filter. Each video frame is firstly filtered, then the standard deviation (*std*) of the pixel values is computed. Finally, the maximum value over time ($\max_{t\in[1..N]}\{\ \}$), is chosen to represent the SI metric. This process can be represented by the following equation:

$$SI = \max_{t\in[1..N]} \left\{ std\Big(sobel\left(f\left(t\right)\right) \Big) \right\}. \tag{4.1}$$

- Temporal Information (TI) based upon the motion difference, which is given by the difference between the pixel values at the same spatial location but in

Table 4.1: Test sequences used in the experiments.

Sequence	**SI / TI**	**Inter (%)**	**Bitrate (Mbps)**	**Description**
Basketball Drill $832 \times 480@50Hz$	33.4 / 14.4	88.1	4.50	High motion with several basketball players captured from a high point
Basketball Drive $1920 \times 1080@50Hz$	33.0 / 15.2	85.0	15.00	Several basketball players passing a ball captured from a low point
Book Arrival $1024 \times 768@30Hz$	28.4 / 21.7	90.7	1.50	Moderate translational motion with two moving persons
Bosphorus $3840 \times 2160@120Hz$	13.4 / 3.80	91.6	12.00	Boat shipping at low motion with moderate complex background
BQSquare $416 \times 240@60Hz$	63.2 / 11.5	93.7	4.00	Moderate outside motion with moving camera capturing from high point
Cactus $1920 \times 1080@50Hz$	30.5 / 11.4	89.7	20.00	Several objects with high details with moderate motion
Four People $1280 \times 720@60Hz$	31.3 / 6.90	93.3	2.00	Four people talking of the desk and passing objects to each other
Jockey $3840 \times 2160@120Hz$	11.5 / 16.2	82.8	12.00	High motion with one horse rider
Kendo $1024 \times 768@30Hz$	19.6 / 16.1	88.0	0.60	Moderate motion with two moving persons, and moving camera
Kimono $1920 \times 1080@24Hz$	23.4 / 32.5	83.0	8.00	Capture of a women moving in a forest with moderate motion
Kristen and Sara $1280 \times 720@60Hz$	25.6 / 6.30	93.1	15.00	Two persons talking to each other with moderate motion
Park Scene $1920 \times 1080@24Hz$	31.3 / 11.6	92.0	10.00	Moderate motion with cyclists passing across the scene
Party Scene $832 \times 480@50Hz$	52.8 / 11.4	89.8	4.50	Moderate motion with three children playing
People on Street $2560 \times 1600@24Hz$	40.0 / 25.4	87.4	12.00	Elevated capture point of people moving; high motion and texture complexity
Race Horses $832 \times 480@30Hz$	43.7 / 24.4	79.7	8.00	High motion with several horse riders
Tennis $1920 \times 1080@30Hz$	20.3 / 45.3	76.2	5.00	High motion with one moving person in the scene
Traffic $2560 \times 1600@30Hz$	28.4 / 22.3	93.0	6.00	Elevated camera capturing highway with several cars with moderate speed

consecutive time instances. TI is computed as the maximum displacement over time of the standard deviation over space. Larger displacements in adjacent frames will result in higher values of TI. This process can be represented by the following:

$$TI = \max_{t \in [2..N]} \Big\{ std\Big(f\left(t\right) - f\left(t-1\right) \Big) \Big\}. \tag{4.2}$$

The values of SI and TI calculated for the test sequences are shown in the Table 4.1 (second column). The percentage of inter-coded blocks and the bitrates are also shown in the table and explained below.

For the simulation carried out in this study the HM reference software, version 16.2, was used [160]. The sequences were encoded with all the coding modes enabled,

Table 4.2: Characteristics of the loss patterns used in the experimental simulations.

Packet Loss Ratio (PLR)	Average burst length	Maximum burst size
1%	1.24	2
3%	1.47	3
5%	1.83	5
10%	2.05	7

following two recommended configurations: Low-Delay (LD), i.e., P-frames with 4 references frames, and Random-Access (RA), i.e., B-frames with 2 reference frames in each list [161]. These prediction structures cover the use of P-frames and hierarchical B-frames [162], which are commonly used in HEVC streaming [163]. An IDR period of 16 frames is used, derived from the recommended configurations. Further IDR periods are evaluated in Section 4.2.4. In order to achieve high quality images for all coded sequences (i.e., PSNR around 40 dB) different bitrates were used, as listed in Table 4.1. The filtering operations were disabled at slices boundaries, in order to keep the slices self-contained. Moreover, the ratio of inter-coded CUs of each coded sequence using the default coding parameters is also shown in the third column of the table since this is an important parameter to analyse error propagation.

To simulate the lossy transmission environment, each frame was divided into several slices and each slice encoded and packetized into one NAL unit, which were then transmitted as the payload of independent packets. In these simulations a whole packet is lost whenever an error occurs, originating a lost slice. Random packet loss was simulated using a two-state Markov model, derived from the classical Gilbert-Elliot [164, 165], which can be used to describe error patterns in real time video applications [166]. In this study a simplified version of the Gilbert-Elliot model was used by only defining the loss probability without control over burst loss events. The application of this model resulted in loss patterns with approximately the same average burst loss as the ones provided in [167]. For each test condition, 50 runs were tested and the average quality was measured. Four different PLRs with the characteristics presented in Table 4.2 were used. The remaining configurations have been kept at their default values, as given in the common test conditions [161].

4.2.3 Evaluation of the impact of the partition units size

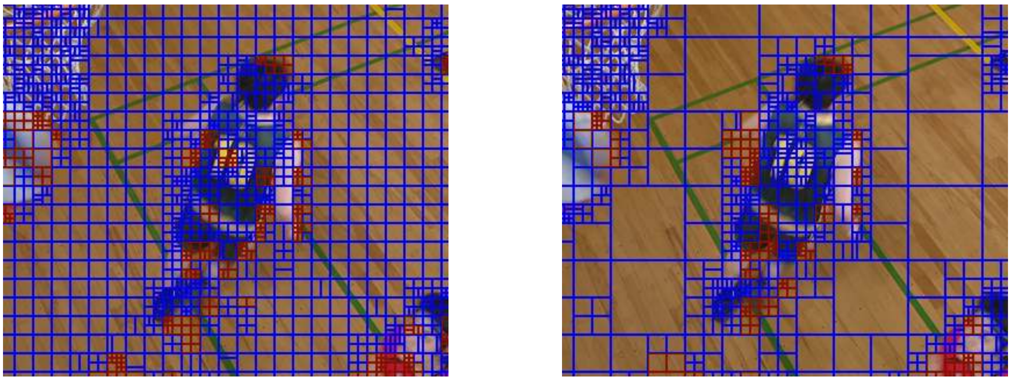
In the first set of experiments the impact of the maximum CU size in the decoded video quality is investigated. In this experiment the maximum CU size is set to three different sizes, i.e., 16×16, 32×32 and 64×64 pixels. The lower size is targeting the

Table 4.3: Average PSNR (dB) for different maximum CU size.

Sequence	**Maximum CU size**	**No-Loss**	**Packet Loss Ratio**		
			1%	**5%**	**10%**
Basketball Drill	16×16	39.25	37.82	34.66	32.67
	32×32	+0.47	+0.38	+0.66	+0.63
	64×64	+0.65	+0.64	+0.87	+1.01
Book Arrival	16×16	39.85	38.11	33.63	30.60
	32×32	+0.25	+0.19	+0.21	-0.38
	64×64	+0.43	+0.71	+1.06	+0.75
BQSquare	16×16	38.99	38.11	35.04	32.75
	32×32	+0.22	+0.17	+0.33	+0.34
	64×64	+0.31	+0.37	+0.74	+0.88
Cactus	16×16	37.50	37.19	36.72	36.45
	32×32	+0.39	+0.38	+0.41	+0.39
	64×64	+0.50	+0.52	+0.55	+0.48
Four People	16×16	39.54	39.26	38.42	37.54
	32×32	+0.44	+0.43	+0.38	+0.34
	64×64	+0.59	+0.60	+0.59	+0.57
Kendo	16×16	39.76	38.47	33.90	29.57
	32×32	+0.99	+1.04	+0.74	+1.01
	64×64	+1.39	+1.34	+1.00	+1.14
Kimono	16×16	42.42	41.69	40.64	40.13
	32×32	+0.36	+0.39	+0.32	+0.33
	64×64	+0.46	+0.56	+0.47	+0.51
Park Scene	16×16	39.03	38.49	37.64	37.18
	32×32	+0.33	+0.26	+0.38	+0.30
	64×64	+0.44	+0.41	+0.43	+0.38
People On Street	16×16	35.20	34.05	32.66	31.83
	32×32	+0.37	+0.31	+0.34	+0.26
	64×64	+0.70	+0.69	+0.57	+0.57
Race Horses	16×16	39.19	37.92	36.33	35.56
	32×32	+0.51	+0.55	+0.49	+0.47
	64×64	+0.68	+0.81	+0.79	+0.66
Tennis	16×16	39.27	37.91	35.39	34.04
	32×32	+0.95	+0.83	+0.73	+0.58
	64×64	+1.41	+1.33	+1.15	+1.02
Average Difference	**32×32**	**+0.48**	**+0.45**	**+0.45**	**+0.39**
	64×64	**+0.69**	**+0.73**	**+0.75**	**+0.72**

MB size used in the H.264/AVC.

Table 4.3 shows the average PSNR (dB) obtained for the three test conditions and for three different packet loss ratios. Results for the No-Losscase shows a consistent quality increase as the maximum CU size increases, when comparing the maximum CU size of 64 × 64 with 16 × 16. This confirms the principle used by the HEVC standard, whereby larger coding units lead to higher coding efficiency. For the Tennis sequence, which has the highest TI parameter (see Table 4.1), a maximum of 1.41 dB is achieved. Moreover, it is also noticeable that sequences with higher amount of inter-coded blocks take less advantage of the larger CUs. This is noticeable in the results of the BQSquare



(a) 16 × 16 (b) 64 × 64

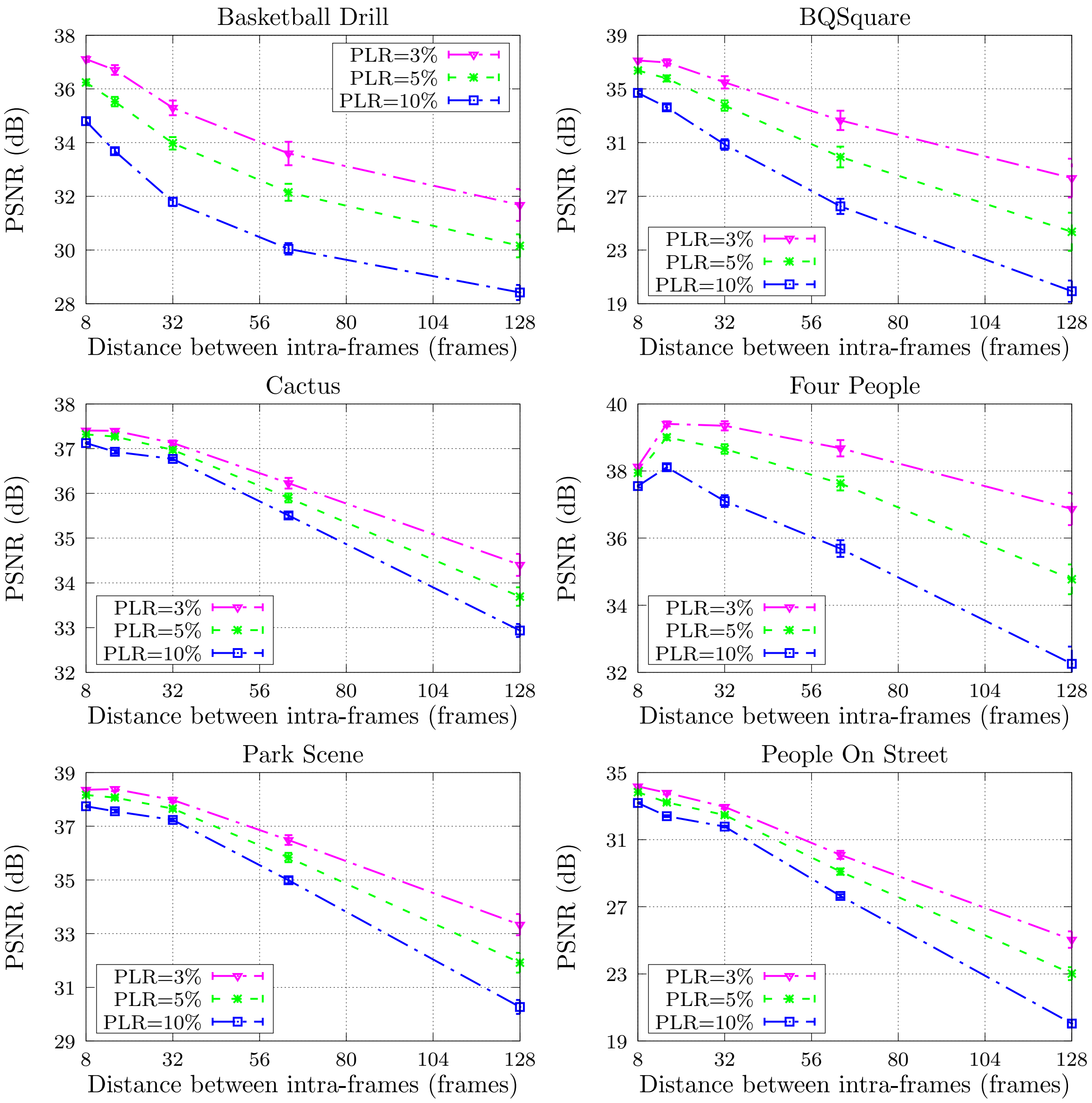
Figure 4.1: Block partition and prediction mode for a region of Basketball Drill.

sequence (i.e., highest ratio of inter-coded blocks and lowest quality gain in the No-Losscase). Overall, comparing the results with and without packet loss, it is noticeable that similar quality gains are obtained, mostly due to the superior coding efficiency of larger CUs. In most cases the quality gains slightly decrease for increasing PLRs. Therefore, one may conclude that the error robustness of HEVC streams is not affected by limiting the maximum CU size.

Figure 4.1 illustrates an example of the frame partitions for a region of the Basketball Drill sequence for two cases: (i) maximum CU size of 64 × 64 and (ii) maximum CU size of 16 × 16. As shown, in the moving regions, e.g., moving player, the partitions obtained for both cases have a similar pattern, mainly composed of CUs with small size. Larger CUs are only used in the floor, which has a lower spatial complexity and suffers small changes between frames, therefore, easier to compress with 64 × 64 CUs. Moreover, those regions can be more accurately recovered, by using neighbouring information correctly decoded (spatial and temporal), thus not incurring in severe error propagation. Therefore, the overall error robustness is not affected by reducing the maximum CU size. One may conclude that the use of larger coding units is able to improve the overall HEVC efficiency communications networks with and without packet loss.

4.2.4 Evaluation of the impact of intra refreshing

In this sub-section the relation between the temporal distance of consecutive intra-coded frames, i.e., period of the intra frames, and the video quality under packet loss is evaluated. As the intra-coded frames have no dependencies to prior coded pictures, they introduce a refreshing point into the bitstream, which stops temporal error propagation along several frames or slices. This allows the decoder to recover



Basketball Drill

BQSquare

Cactus

Four People

Park Scene

People On Street

Figure 4.2: Decoded video quality with a 95% confidence interval for different intra periods.

from frame/slice loss and continue error-free decoding in a seamless manner.

Figure 4.2 shows the average decoded video quality (PSNR) obtained with different error patterns with a 95% confidence interval for different distances of intra frames, covering PLRs between 3% and 5%. Results in the figure indicate that as the distance between intra frames increases, the decoded video quality significantly decreases. The quality decreasing is especially noticeable for intra periods higher than 32 frames, indicating that temporal distances greater than approximately one second are not suitable for video streaming under error-prone conditions. When a high temporal distance between intra frames is used, the quality decreasing is more noticeable in the People On

Street and BQSquare sequences, which have higher spatial complexity (see values for SI), leading to a reduction of 13 dB and 14 dB, respectively. As spatial complexity increases, the prediction residue energy also increases, which leads to more erroneous information in case of data loss, reducing the video quality under errors. Finally, based on the confidence interval one may conclude that for different error patterns the average PSNR of all frames does not significantly change. A maximum variation of 1 dB, approximately, is observed in Figure 4.2.

One should also note from the results of Figure 4.2 that, using a small distance between intra frames is not always more efficient. Since intra frames require higher amount of bits, the overall coding efficiency is reduced, resulting in lower overall quality in case of data loss, when comparing with other cases. This is clearly noticeable for the Four People sequence, where replacing the intra period from 8 to 16 frames leads to a quality increase of 1 dB for 5% of PLR. For the remaining sequences it can be concluded that using sixteen frames (not higher than half a second) between refresh points (i.e., intra frames) yields an acceptable efficiency under errors. This provides a good trade-off between compression efficiency and error robustness in case of data loss. For this reason for the remaining experimental evaluation an IDR of 16 frame will be used.

4.2.5 Evaluation for different slice partitioning schemes

The partitioning of a frame into different slices increases the flexibility of video stream packetisation, along packets of different sizes, thus providing adequate adaptation of coded data units to different transport layers and networks. By partitioning the video frame into different slices, spatial refresh points are introduced, limiting the error propagation across different CTUs. In this sub-section the effect of the slice size is evaluated, to study the influence of different slice partitioning schemes on the error robustness of HEVC streams. A comparison of the quality obtained for different partition modes is also carried out, in order to find the most efficient mode for video transmission under data loss conditions.

Figure 4.3 shows the average PSNR with a 95% confidence interval for four different slice partitioning modes and different PLRs. The first three modes are characterised by a different amount of bytes per slice (B/S), namely 1200, 2400 and 4800 bytes. In these cases each frame has a variable number of slices. The last mode uses a single slice per frame (referred to as *SS*), corresponding to the case where slice partitioning is disabled. The results shown in Figure 4.3 reveal a clear impact of the slice partitioning

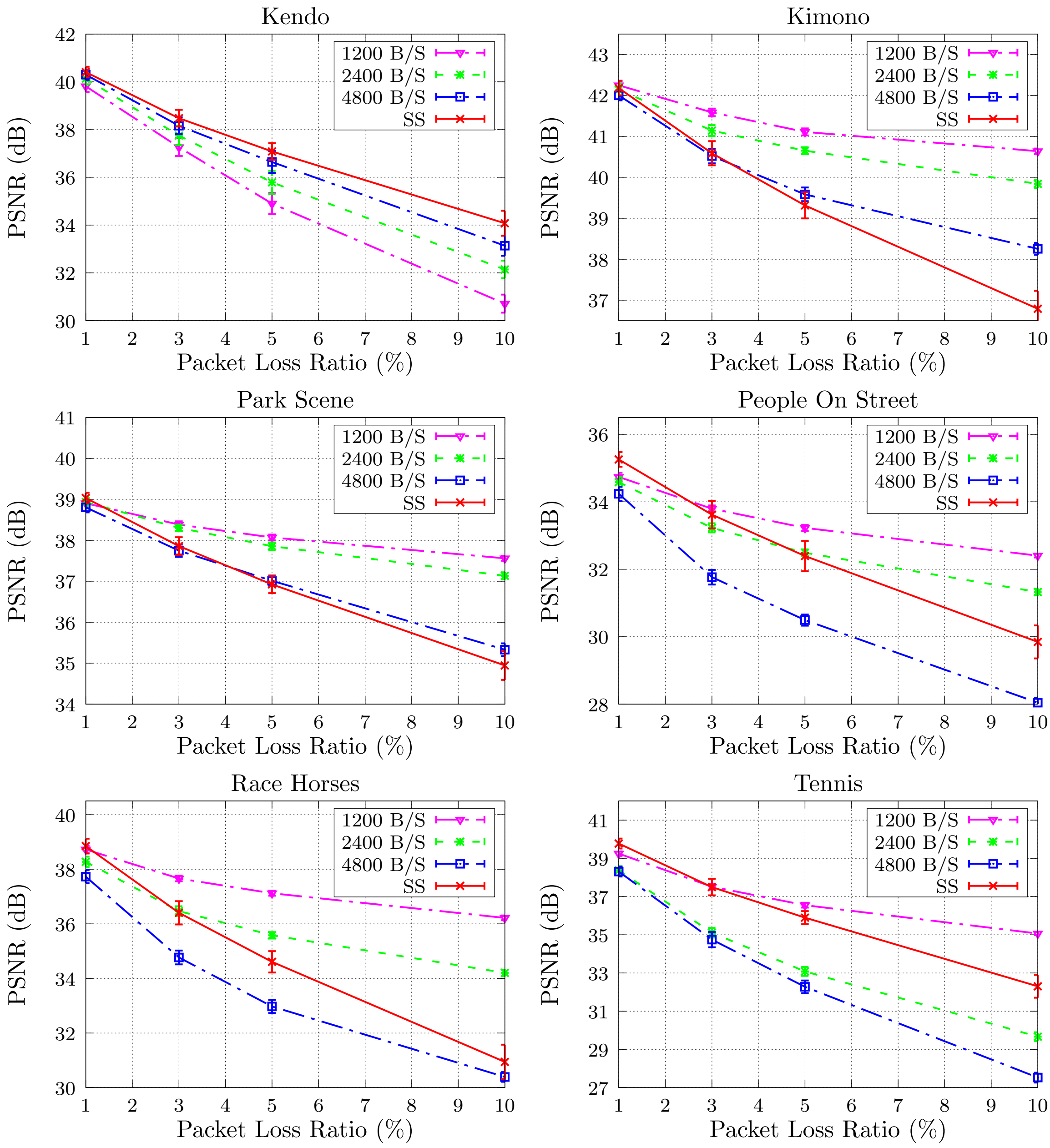


Figure 4.3: Decoded video quality with a 95% confidence interval for different slice partitioning modes.

mode on the error robustness of HEVC, as the quality degradation for increasing PLR changes significantly for each mode. Comparing the cases with fixed amount of bytes per slice, it is clearly noticeable that a higher amount of bytes also corresponds to a higher quality degradation. This is true for the results of most test sequences shown in Figure 4.3. Although the use of smaller slices increases the overhead required for slice headers, the superior error robustness achieved with 1200 B/S is able to overcome the inherent degradation on coding efficiency. This is due to the lack of refresh points within a frame, resulting in errors that affect larger regions of the frame, which are propagated to the subsequent inter predicted frames.

The results for of *SS* (slice partitioning disabled) reveals that this approach also leads to lower error robustness when comparing with those cases where the frames are divided into multiple slices. In this case the 95% confidence interval is greater than the others shown in Figure 4.3. This is because the packet size is variable, which implies higher variability in the amount of lost information. Although higher coding efficiency is achieved in this mode, the lower error robustness leads to an overall poor performance under frame or slice loss conditions. It is worthwhile to notice that the results for Kendo sequence do not follow the main trend. This can be explained by the low bitrate required to compress the video signal, which is the lowest when compared with the other test sequences (see Table 4.1). Therefore, most of the inter-coded frames can be transmitted using a single slice, even with 1200 bytes per slice, resulting in a similar partitioning for all tested modes, and thus similar error robustness. In this case, the higher coding efficiency achieved by the *SS* case also leads to higher quality in case of errors.

4.2.6 Motion vector coding impact on error propagation

In HEVC, the new MV prediction modes known as AMVP and Merge Mode [168], may be significant contributors to the poor performance under network errors. These modes are characterised by coding the motion information using previously encoded MVs, which obviously leads to prediction mismatches at the decoder-side when such neighbouring information is missing due to network losses. Therefore, in this subsection the impact of using spatial and temporal MV candidates on the decoded video quality under errors is evaluated.

Spatial MV candidates

The first set of experiments aimed to investigate the impact of the amount of spatial MV candidates on the video quality. As mentioned before, HEVC allows up to five candidates to be used, which can be provided solely by spatial neighbours. In order to check the impact of the MV candidates, different coding configurations were used, each one characterised by different amounts of spatial MV candidates, i.e., 1, 3 and 5 MVs. Figure 4.4 shows the average quality results for the Basketball Drill and People On Street sequences, under different packet loss conditions. The results show that, for different number of MV candidates, similar quality decreasing is achieved as the

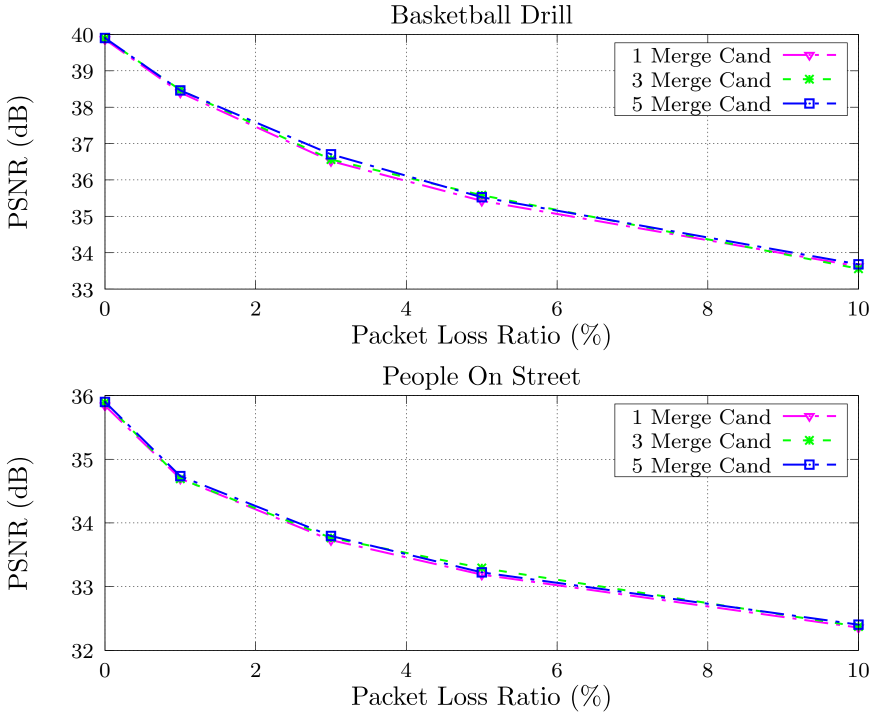


Figure 4.4: Average quality using different number of spatial MV candidates.

Table 4.4: Relation between the number of available MV candidates and its usage

Max. number of candidates	**Candidate selected (ratio)**				
	1	2	3	4	5
3 MVs	66.17%	22.79%	11.04%	-	-
5 MVs	64.54%	22.24%	7.38%	3.37%	2.47%

PLR increases. Therefore, one may conclude that varying the number of available MV candidate does not affect the error resilience of the HEVC.

The occurrence rate for each spatial MV candidate was also analysed in order to check whether they are uniformly used. Table 4.4 shows the usage percentage of each MV candidate when three and five candidates are available for the slice partitioning mode of 1200 B/S. The results indicate that most predictions use the first candidate and minor differences occur by adding 2 or more MVs. Since the mostly used MV candidates are the initial ones, changing the number of available candidates does not significantly affects the error resilience. Moreover, the slice partitioning mode used only 1200 bytes per slice, which introduces refresh points in each frame, reducing the spatial error propagation due to mismatched MV predictions.

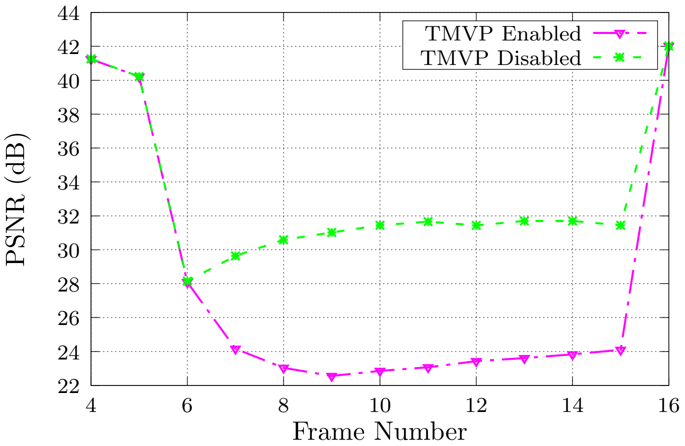


Figure 4.5: Error propagation for Kendo sequence when Frame #6 is lost for the LD configuration using different TMVP configurations.

Temporal MV candidates

The second set of experiments aimed to find the influence of losing temporal motion predictors on the error resilience of HEVC. In order to perform this test, a single frame loss in enforced, and the error propagation is evaluated for two configurations: (i) with the temporal MV candidate enabled and (ii) disabled. Figure 4.5 illustrates the error propagation, when a single frame is lost. For this particular experiment an intra-refreshing period of 32 frames was used to evaluate the impact on long error propagation periods. This is commonly used in quality evaluation of erroneous video decoding and follows the reference conditions [161]. The results show that when the TMVP mode is enabled, the loss of Frame #6 leads to significant reduction in the reconstruction quality of the subsequent frames. On the contrary, when the temporal MV candidate is disabled, the motion information on subsequent frames is not affected by the loss of a single frame and significantly higher quality is obtained. This is the effect of breaking the temporal dependencies between MVs, which prevents erroneous MVs to propagate throughout the set of dependent frames.

In order to further evaluate the impact of the temporal MV dependencies on the error propagation, the HEVC streams were subject to random packet losses. Table 4.5 presents the average quality results, measured using the PSNR. As expected, the results indicate that in the those cases without packet loss (No-Loss) higher quality is achieved by enabling the TMVP candidate, since higher compression efficiency is obtained because more MV predictors are used. This is true for both test conditions, *i.e.*, LD and RA. However, in the presence of errors, the video quality significantly decreases due to low error robustness when temporal MV candidates are used. Alternatively, av-

Table 4.5: Average PSNR (dB) under random losses for different TMVP configurations.

Sequence	**TMVP**	**No-Loss**	**Packet Loss Ratio**		
			1%	**5%**	**10%**
Low-delay configuration (LD)					
Basket Drill	Enabled	38.69	34.73	27.92	25.41
	Disabled	-0.04	+1.97	+3.93	+3.43
Book Arrival	Enabled	40.62	36.78	30.78	27.29
	Disabled	-0.01	+1.98	+3.32	+3.33
BQSquare	Enabled	39.64	31.99	23.45	20.95
	Disabled	-0.05	+6.14	+10.6	+10.7
Kendo	Enabled	43.31	38.80	29.80	25.45
	Disabled	-0.04	+2.04	+4.51	+4.98
Park Scene	Enabled	36.89	33.78	28.09	25.77
	Disabled	-0.06	+2.33	+5.72	+6.05
Race Horses	Enabled	37.06	29.64	22.92	20.35
	Disabled	-0.07	+3.72	+4.78	+4.19
Tennis	Enabled	41.61	33.74	25.73	22.68
	Disabled	-0.04	+3.10	+4.01	+3.52
Average difference		**-0.04**	**+3.04**	**+5.32**	**+5.17**
Random-access configuration (RA)					
Basket Drill	Enabled	39.48	36.84	30.50	27.15
	Disabled	-0.03	+0.25	+0.71	+0.61
Book Arrival	Enabled	40.94	38.38	32.43	28.55
	Disabled	-0.01	+0.38	+0.61	+1.00
BQSquare	Enabled	41.47	35.86	26.77	23.13
	Disabled	-0.08	+0.74	+1.58	+1.7
Kendo	Enabled	43.90	40.08	29.77	25.71
	Disabled	-0.04	-0.05	+0.55	+0.45
Park Scene	Enabled	37.90	35.26	30.51	27.74
	Disabled	-0.06	+0.67	+1.29	+1.34
Race Horses	Enabled	37.43	32.36	25.35	22.47
	Disabled	-0.05	+0.61	+1.08	+1.00
Tennis	Enabled	41.78	36.33	28.80	25.49
	Disabled	-0.04	+0.68	+0.74	+0.67
Average difference		**-0.04**	**+0.47**	**+0.94**	**+0.97**

erage quality up to 5.17 dB of PSNR is achieved when the temporal MV predictors are disabled in the LD configuration. The results also show that the negative impact of using temporal MV dependencies is smaller when B-frames are used. For the Random-Access, the average gain of disabling the TMVP is 0.78 dB. A higher negative impact is obtained when only P-frames are used (LD configuration), since subsequent frames are temporally closer, leading to more accurate temporal MV candidates. Moreover, these results also indicate that for higher PLR the negative impact of using the TMVP does not increase, as both configurations are subject to high quality degradation.

Remarks on error robustness of HEVC

Overall, the results presented in the previous sub-sections show that HEVC has poor robustness against network losses, resulting in significant quality degradation, though depending on the video content and coding configurations. As shown, some coding configurations, such as slice partition and intra refreshing period, have high correlation with the decoded video quality when errors affect the compressed streams. Therefore, these are important aspects to take into consideration in the design and implementation of video delivery services and applications over error-prone networks.

One of the main improvements of the HEVC standard is the support for larger coding units up to 64×64, which does not have a negative impact on the error robustness of coded video streams, indicating that they should be allowed to achieve higher coding performance. Regarding the use of neighbouring information for MV coding, on the one hand the spatial MV candidates do not significantly decrease the error robustness. On the other hand, the use of temporal MV candidates significantly contribute to increase the temporal dependencies, leading to more severe quality degradation in the presence of packet loss.

Based on these evidences, the next section presents an efficient method to improve the error robustness of HEVC. The proposed method is based on the underlying principle of reducing the mismatch of MV predictions, by selectively disabling some MV predictions at the encoder-side in order to reduce the overall temporal dependencies.

4.3 Dynamic motion vector refreshing

The aforementioned results revealed low error resilience in motion prediction because any error is easily propagated through several frames. Based on this evidence, this section proposes a method to make intelligent use of temporal MV candidates during the motion estimation process, in order to decrease the temporal dependency, and improve the error resiliency without penalising the R-D performance.

4.3.1 Proposed method

The new approach to dealing with temporal MV dependencies is accomplished by modifying the R-D optimisation that is used to select the best MV for inter-prediction. The objective is to increase the number of MVs that can be independently decoded, in order to improve the error recovery capability after an arbitrary frame loss event.

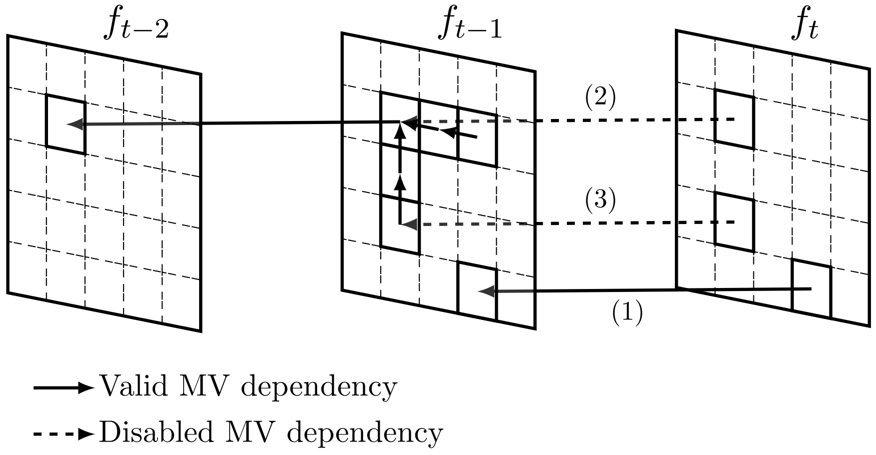


Figure 4.6: Proposed approach to constrain the temporal MV dependencies (the arrows point to the MV predictors).

The proposed method also overcomes the limitation of the method presented in [158], which breaks the motion information dependencies between frames. Instead of disabling the temporal motion candidates at the frame level, the proposed method aims to remove the dependencies at the CU level. To achieve this goal, the temporal dependency of all MV candidates is firstly analysed for each block, in order to select the suitable candidates to be used as predictors for the current MV. Based on the analysis of the temporal dependency, some temporal candidates are marked as unsuitable predictors of the current MV, and for this reason they are not selected for MV prediction. In the proposed approach, a temporal MV candidate is marked as unsuitable, based on the following criteria:

1. if it is encoded based on another temporal MV candidate from a previously encoded frame;

2. if it is predicted using a spatial neighbour that was previously encoded using a temporal MV candidate.

Figure 4.6 can be used to explain the proposed method, illustrating different blocks and different MV dependencies represented with arrows. Note that the squares illustrate a generic block size; the representation of the temporal dependencies is simplified, and in some cases may not correspond to the exact co-located block in the previous frame. In frame f_{t-1} there are several MVs encoded using spatial predictions and one MV encoded using a temporal prediction (arrow pointing to f_{t-2}). In frame f_t, the MV corresponding to the block (1) may use the temporal MV candidate, since the co-located block in f_{t-1} is not temporally dependent. However, in the proposed scheme the MV prediction identified with the arrow (2) is not allowed, which prevents the

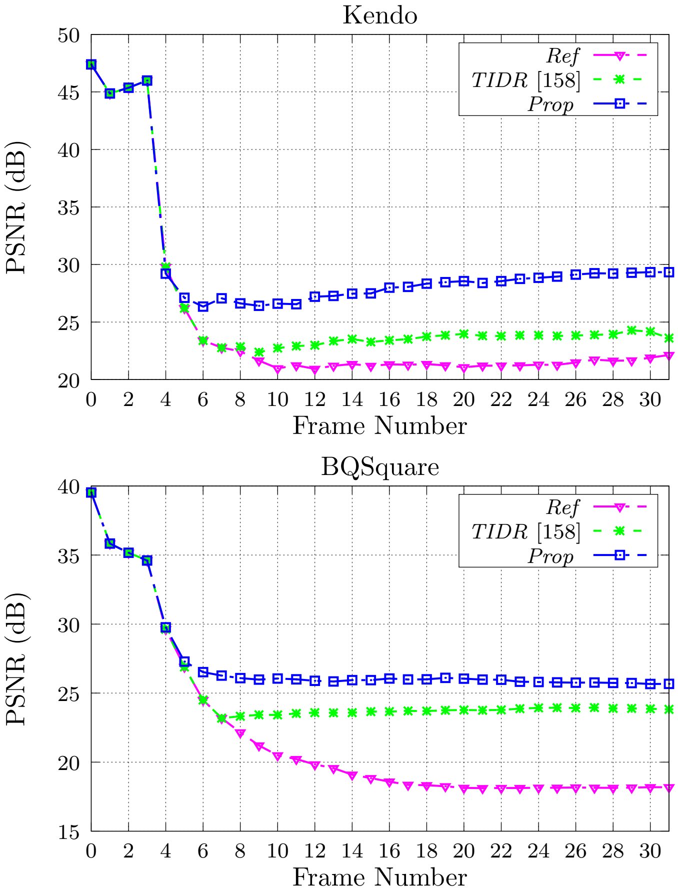


Figure 4.7: Quality obtained with the proposed scheme, *TIDR* and reference HEVC when the frame #5 is missing for Kendo and BQSquare sequences.

propagation of temporal dependencies. Moreover, the block corresponding to the arrow (3) cannot use the temporal MV candidate since it already depends on a MV that was previously encoded using the temporal candidate. The proposed method reduces the temporal dependencies of the MVs at the block level, providing improved error robustness to the encoded streams. Using this approach, the dependencies are selectively removed, improving the error recovery accuracy, without fully disabling the temporal MV candidates for a given frame. Therefore, coding efficiency is not significantly penalised.

4.3.2 Performance evaluation

The performance of the proposed scheme (*Prop*) was evaluated against two other approaches: reference HEVC with TMVP enabled and the *TIDR* method proposed in [158].

Table 4.6: Average standard deviation of the PSNR (dB) for the proposed scheme, *TIDR* and reference HEVC.

Sequence	**TMVP configuration**	**Packet Loss Ratio**		
		1%	**3%**	**5%**
Basketball Drill	*Ref*	5.09	7.51	7.70
	TIDR [158]	5.00	7.30	7.52
	Prop	4.78	7.03	7.24
Book Arrival	*Ref*	5.45	8.74	9.08
	TIDR [158]	5.24	8.48	8.88
	Prop	4.84	7.83	8.26
BQSquare	*Ref*	5.98	8.86	9.14
	TIDR [158]	5.20	7.51	8.01
	Prop	4.94	7.28	7.68
Kendo	*Ref*	7.56	11.58	12.05
	TIDR [158]	7.25	11.00	11.55
	Prop	6.62	10.06	10.66
Race Horses	*Ref*	5.74	8.31	8.55
	TIDR [158]	5.47	7.78	8.07
	Prop	5.29	7.63	7.88
Tennis	*Ref*	6.70	9.55	10.07
	TIDR [158]	6.55	9.25	9.83
	Prop	6.42	9.11	9.67

Error propagation under single loss events

Firstly, the performance of the proposed method was measured under a single loss event (i.e., only one frame is affected by errors). Figure 4.7 shows the error propagation for the different methods. The results show a significant quality difference between the proposed method, the *TIDR* method and the reference HEVC. When the temporal MV candidate is disabled (proposed and *TIDR* method) higher error robustness is attained. The results also show the benefits of using the proposed method over the *TIDR* method. While the *TIDR* method is only able to recover from the frame loss after the refreshing frame, *i.e.*, frame #8, especially for the BQSquare sequence, the proposed method is able to recover right after the frame loss.

Quality evaluation under random errors

Secondly, further simulations were performed to evaluate the effectiveness of the proposed error resiliency method under realistic packet loss patterns. Table 4.6 illustrates the average standard deviation of the PSNR, for three of the test sequences. The results show that increasing the PLR leads to higher quality variations, thus the impact of errors in the subjective quality experience is higher. However, reducing the temporal dependencies of the MVs using the proposed method decreases such quality variations, achieving the lowest variation value for all tested scenarios. The proposed method is

Table 4.7: BD-PSNR (dB) differences for the proposed method and *TIDR* at different frame loss conditions.

Sequence	TMVP configuration	No-Loss	Packet Loss Ratio 1%	5%
Basketball Drill	*TIDR* [158]	-0.011	0.09	0.37
	Prop	-0.054	0.22	0.82
Book Arrival	*TIDR* [158]	-0.003	0.10	0.43
	Prop	-0.012	0.36	1.44
BQSquare	*TIDR* [158]	-0.005	0.42	1.57
	Prop	-0.032	0.54	1.97
Kendo	*TIDR* [158]	-0.008	0.15	0.62
	Prop	-0.031	0.38	1.54
Race Horses	*TIDR* [158]	-0.010	0.18	0.69
	Prop	-0.040	0.24	0.91
Tennis	*TIDR* [158]	-0.003	0.10	0.49
	Prop	-0.041	0.21	0.93
Average	*TIDR* [158]	-0.006	0.17	0.70
	Prop	-0.034	0.34	1.34

able to reduce the quality variations up to 0.94 dB for Kendo sequence (3% of packet loss) compared to the *TIDR* method. Therefore, since the proposed method achieves the lowest PSNR standard deviation, one may conclude that it is able to reduce the video quality variations leading to a better visual quality.

Table 4.7 shows the Bjontegaard's average PSNR increase [169]. The results are obtained taking the standard HEVC encoder as reference. The results presented in Table 4.7 show that the proposed method presents practically the same rate-distortion performance when compared to the *TIDR* method. In general, the average quality, in the No-Losscase, only decreases 0.028 dB in comparison to the *TIDR* method and 0.034 dB in comparison to the reference encoder. However, the proposed method widely outperforms the reference *TIDR* technique in terms of error resiliency performance, achieving in average higher overall quality under error-prone conditions. The effectiveness of the proposed method increases at higher packet loss rates (*e.g.*, 5%), achieving an average gain of 0.64 dB (1.34 – 0.70), and a maximum gain of 1 dB (1.44 – 0.43) for the Book Arrival sequence , when compared to the *TIDR* technique. Higher gains are achieved when compared with the reference HEVC encoder, as shown in Table 4.7.

4.4 Summary

In this chapter, the error resilience performance of HEVC was studied in detail under error prone conditions. This study was partially discussed and published in the paper (C3). The simulation results revealed that new coding tools introduced in HEVC to improve coding efficiency are partially responsible for higher error propagation. This motivated further research on new techniques for highly efficient encoders and decoders, in order to increase error resiliency and to improve the overall performance of video delivery services and applications. Then, a new approach to deal with the weak robustness of MV coding was proposed, addressing MV dependencies at CU level. The proposed method selectively disables the temporal MV candidates, in order to break the dependencies across more than one frame. This method was proposed in the publication (C1). The experimental evaluation was able to demonstrate that the proposed method increases the error robustness and outperforms existing state-of-the-art methods and the reference codec. The results also show that the proposed method is able to provide a good trade-off between coding efficiency and error resilience. Overall, these new findings demonstrate that selective usage of temporal MV candidates is an efficient approach to enhance error resilience in high efficiency video encoders.

Chapter 5

A two-stage approach for robust coding and streaming

5.1 Introduction

This chapter proposes a new two-stage approach to improving the error robustness of HEVC streaming, by reducing temporal error propagation in case of frame loss. This is accomplished by reducing the prediction mismatch that occurs at the decoder after frame loss. The proposed approach comprises the following two stages: (i) at the encoding stage, the reference pictures are dynamically selected based on constraining conditions and Lagrangian optimisation, which distributes the use of reference pictures throughout the frames, reducing the number of PUs that depend on a single reference; (ii) at the streaming stage, a MV prioritisation algorithm, based on spatial dependencies, selects an optimal sub-set of MVs to be transmitted as side redundant information to reduce mismatched MV predictions at the decoder.

This chapter is organized as follows. Section 5.2 presents the problem definition and motivation for the method proposed in the chapter. Section 5.3 provides an overview of the proposed method, describing the functional blocks that compose the two-stage approach for robust coding and streaming. Section 5.4 describes the encoding stage of the proposed method, that is used to select the optimal reference picture for inter-prediction. Section 5.5 describes the streaming stage where a sub-set of MVs are selected to reduce the mismatch motion predictions. Section 5.6 provides a performance evaluation of the proposed method, illustrating the advantages of each stage and how they are combined to achieve superior error robustness in HEVC standard. Finally, Section 5.7 concludes the chapter.

5.2 Problem definition

As previously mentioned, when HEVC bitstreams are subject to transmission losses, the decoded video signal incurs in significant degradation of both objective and subjective quality. This is mostly due to the strong decoding dependencies imposed by the highly complex prediction modes, which are selected according to R-D optimisation criteria, which always assume transmission without errors [15]. Therefore, high compression efficiency is achieved, but resilience to errors and data loss is penalised.

Highly efficient video encoders support the use of multiple reference frames for inter-frame prediction, which are selected based on minimisation of R-D cost that may lead to an unbalanced use of various reference frames. In order to analyse the distribution of the reference frames selected by the R-D decision method used in the reference HEVC encoder, different test sequences were encoded using two reference frames, referred to as r_1 and r_2, and their selection ratios were analysed. In this study, reference r_1 corresponds to the closest neighbouring frame and reference r_2 the second closest. Figure 5.1 illustrates the average usage percentage of reference r_1 and r_2 for different sequences. These results show that, in general, reference r_1 is selected more often than r_2. This is mainly true for sequences with higher motion. For those sequences with lower motion, but high spatial complexity, e.g., BQSquare (see SI in Table 4.1), both reference frames are equally (approximately) selected for temporal predictions, with r_1 being used by 58% of cases and r_2 by 42%. This reveals that, even though two references are used, the encoder mainly uses the temporally closer one as the reference for inter-frame predictions. Therefore, if such reference frame is lost, a higher number of mismatched predictions will occur, inevitably leading to higher error propagation.

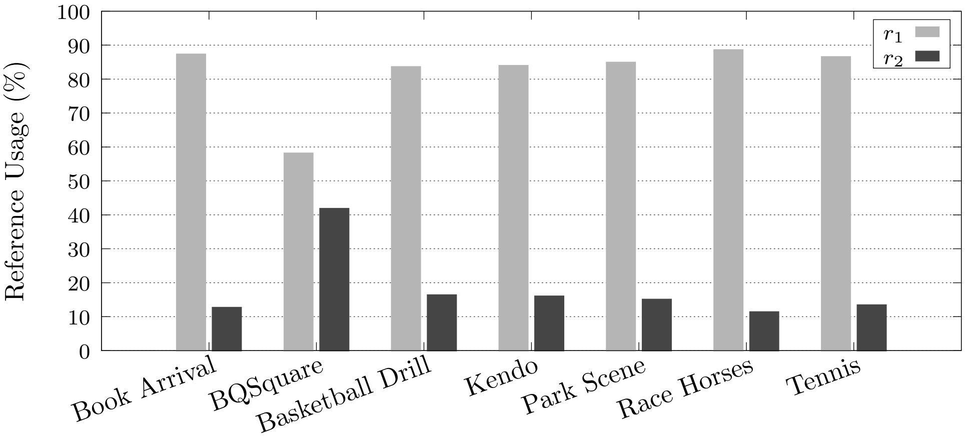


Figure 5.1: Average usage of the reference frame 1 (r_1) and 2 (r_2) for different sequences.

Another coding technique that contributes to increase error propagation is the use of predictions in MV coding [168], which also contributes to increase the total amount of spatial and temporal dependencies. As shown in Section 4.2.6, temporally predicted MVs are responsible for a greater impact on error propagation, as they propagate errors in the MV decoding across different time instances.

To overcome these issues a two-fold approach was devised to increase the error robustness in video streaming by jointly considering two different stages: (i) while encoding live video, by selecting robust coding modes and (ii) after encoding, by adding robust features to the compressed streams. While in the former this is accomplished by optimising coding parameters and decisions to increase robustness, in the latter some type of compressed domain processing is used to increase error robustness of pre-encoded streams, e.g., error resilient transcoding. Inevitably, in both cases this is achieved at the expense of some loss in coding efficiency, therefore the challenge is to find an optimum trade-off between robustness and bitrate overhead. The proposed method is described in the remaining of this chapter.

5.3 Proposed two-stage architecture

The goal of the proposed method is to reduce the error propagation in decoded video when the received stream incurs frame loss. The underlying idea is to constrain inter-frame prediction dependencies, in order to minimise the frame distortion D_t given by

$$D_t = \tilde{f}_t - \hat{f}_t, \text{when } \tilde{f}_{t-1} \neq \hat{f}_{t-1}, \tag{5.1}$$

where $\hat{f}_t$ and $\tilde{f}_t$ are the encoded and decoded frames, respectively, at instant t.

Figure 5.2 illustrates the block diagram of the two-stage architecture proposed in this work. At the encoding stage, a reference frame selection mechanism is used for each PU within a CTU, by forcing the use of different reference frames for adjacent CTUs. The decision criterion contributes to reduce the amount of predictions from a single reference frame and forces to interleave different reference frames along all encoded PUs, based on R-D optimisation. At the streaming stage, the coded stream is parsed to extract the MV information, which allows ranking MVs according to a relevance criterion. Then, a sub-set of these MVs is selected to be transmitted as side information. The amount of selected MVs is not fixed and can be optimised for different content and network characteristics. Note that the streaming stage is an independent and complementary subsystem of the encoder. Thus, it is able to operate at network

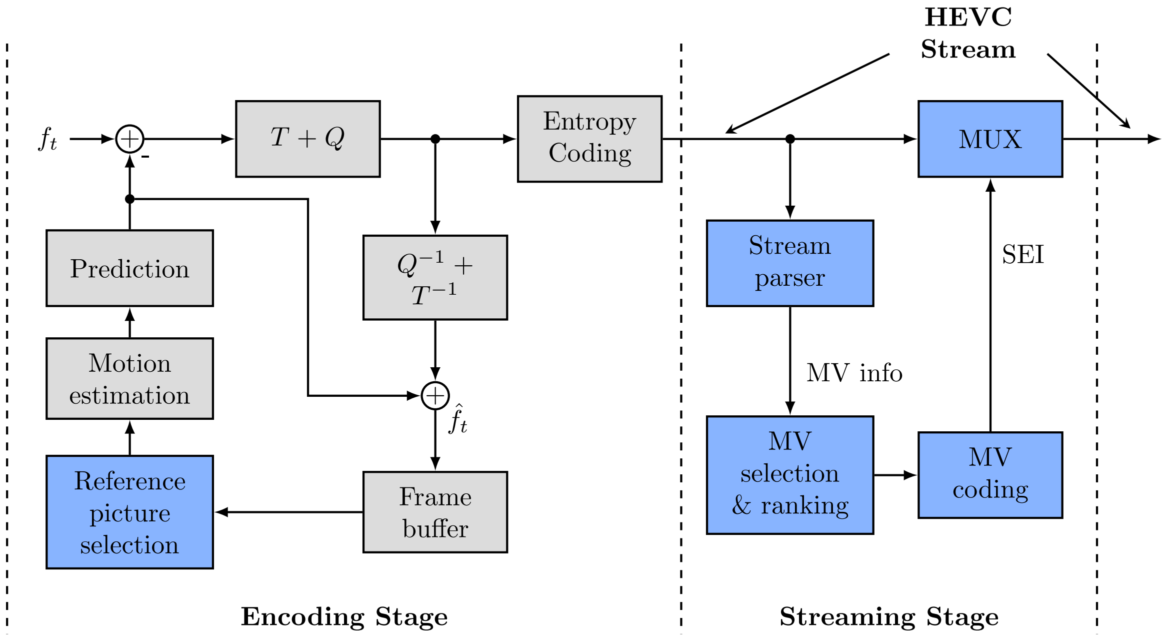


Figure 5.2: Functional blocks of the proposed method (blue blocks) in the context of the HEVC coding and streaming.

nodes, including streaming servers delivering pre-encoded streams.

In case of frame loss, these two stages have the common objective of helping the decoder to recover from mismatched predictions at the expense of a small increase in bitrate. This is equivalent to reduce the error propagation by means of two different mechanisms: (i) reducing the mismatch of inter-frame predictions, by encoding each PU from a different reference frame and (ii) reducing the mismatch of MV predictions by sending the most important MVs as side information. These are the two stages of the proposed method described in the following sections.

5.4 Encoding stage: reference frame selection

The reference picture selection mechanism attempts to uniformly distribute the use of different reference pictures across the CTUs of a coded video frame. In case of frame loss, the error propagation in the decoder is reduced because the number of CTUs predicted from the lost frame decreases and thus EC benefits as more data is correctly decoded from other reference frames. The underlying principle of this method is to minimise the probability of using the same reference picture for neighbouring CTUs.

In a coding structure with two reference frames, as shown in Figure 5.3, the reference frames r_1 and r_2 are used to predict the same number of CTUs in current frame (C),

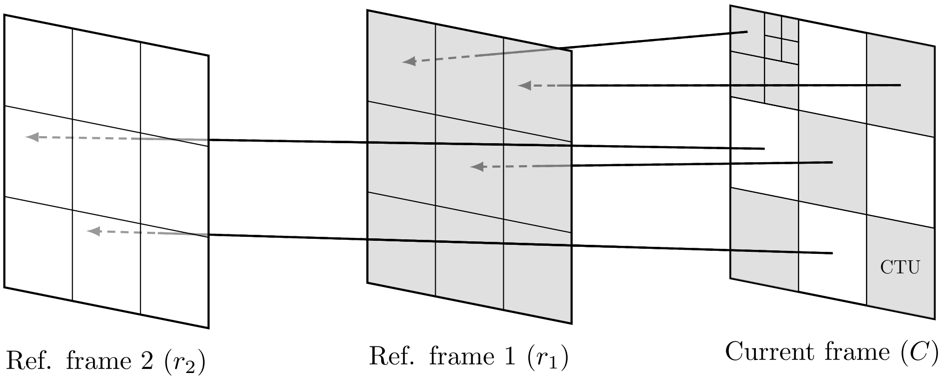


Figure 5.3: Example of the reference picture selection applied to the current frame based on the checkerboard at CTU level (PUs are represented by small partitions).

resulting in a checkerboard structure, where a different reference frame is used for predicting every other CTU within the current frame. If each reference frame is used to predict half of the pixels in the current frame, when any of these references is lost, the subsequent predicted frame is only partially distorted by mismatched predictions. This is because only half of its predictions depend on the missing reference frame, which limits the overall error propagation. Moreover, since the motion information of the predicted frame is encoded using, not only the missing frame as reference but also the past ones, those MVs that cross over a missing reference frame, pointing to the previous ones, are still useful for reconstructing accurate MVs and also for EC. In the limit, a straightforward implementation of the previous example leads to a simple interleaved uniform distribution of the reference frames over all CTUs in the current frame, forcing the checkerboard structure shown in Figure 5.3. In the case of two reference frames, 50% of the CTUs would use one reference frame, while the other 50% would use the other reference. However, such solution would not take into consideration coding efficiency, since the selection process of the reference frame would not be R-D optimised for each PU within a CTU.

Therefore, a more efficient approach to improving error robustness was devised to dynamically choose the coding mode for each PU, that ensures a balanced use of all possible reference frames $F = \{r_1, r_2, ...r_S\}$, where S is the total number of references, across the CTUs of the current frame C, also optimising the R-D cost. This leads to the overall goal of finding the optimum R-D coding mode for each PU, as the one which tends to minimise the standard deviation of the reference frames use count $\mathbf{N}_C(r_i)$, considering all encoded PUs of the current frame C, as follows:

$$\sigma_C = std\Bigg(\mathbf{N}_C(r_i)\Bigg), i = 1, 2, \cdots S, \tag{5.2}$$

where σ_C measures the deviation from the checkerboard solution, which corresponds to $\sigma_C = 0$, since all possible reference frames $r_i \in F$ are uniformly used throughout the PUs of C.

To attain the goal defined above, the best coding mode for each PU is chosen through a Lagrangian R-D optimisation that tends to minimise σ_C. This is done by penalising (or benefiting) the R-D cost of those coding modes which intend to use reference frames that have already been used more (or less) frequently than their average use in previous PUs of the current frame. For each PU, this mechanism has the effect of reducing (or increasing) the probability of using each possible reference frame $r_i \in F$, according to the number of previous PUs that have used each one. Such approach allows to select the reference frames, by independently processing each PU in order to minimise the global standard deviation given by (5.2). Thus, the optimum coding mode ϕ^*, selected from a set of possible coding modes M, is derived based on the following Lagrangian optimisation:

$$\phi^* = arg \min_{\phi \in M} J(\phi), \tag{5.3}$$

$$J(\phi) = \Big(D(\phi) + \lambda \times R(\phi)\Big) \times e^{W(r_\phi)}, \tag{5.4}$$

where $D(\phi)$ and $R(\phi)$ are the distortion and rate, respectively, associated with the coding mode ϕ. $J(\phi)$ is Lagrangian cost that measures the the trade-off between distortion and rate. The exponential weighting factor $e^{W(r_\phi)}$ is used to penalise (or benefit) the Lagrangian cost of a given coding mode ϕ associated to reference frame r_ϕ. $W(r_\phi)$ is obtained according to the global and local deviation between the number of times that r_ϕ was used and the average use of all $r_i \in F$ across the PUs of current frame C, as follows,

$$W(r_\phi) = \Big(\Delta_G(r_\phi) + \Delta_L(r_\phi)\Big) \times 2^{-T_{ID}} \times \gamma, \tag{5.5}$$

where T_{ID} is given by the temporal hierarchy in HEVC, and γ is a parameter to control the slope of $W(r_\phi)$. The global and local deviations, $\Delta_G(r_\phi)$ and $\Delta_L(r_\phi)$, respectively, are computed as follows,

$$\Delta_\zeta(r_\phi) = \mathbf{N}_\zeta(r_\phi) - \frac{1}{S}\sum_{i=1}^{S} \mathbf{N}_\zeta(r_i), \;\; \zeta = G, L. \tag{5.6}$$

For $\zeta = G$, function $\Delta_G(r_\phi)$ measures the difference between the number of times that

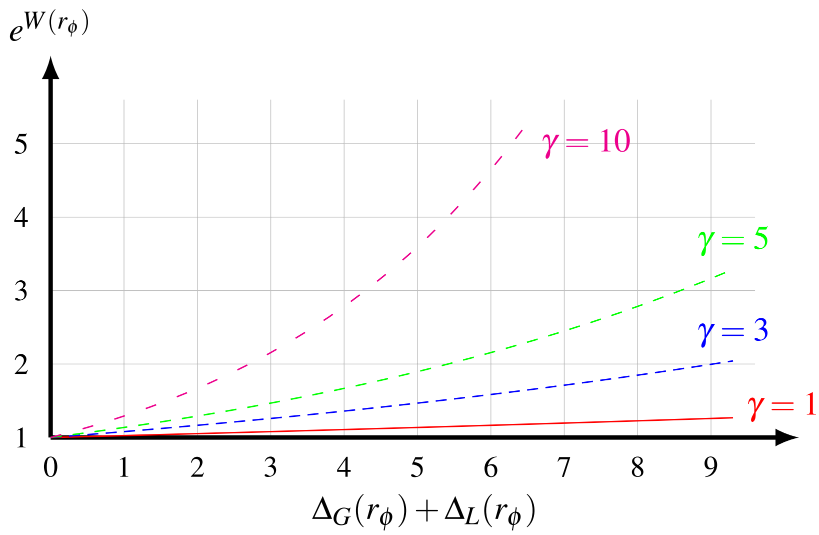


Figure 5.4: Exponential weight as a function of the global and local deviations.

reference frame r_ϕ was used and the global average use of all $r_i \in F$ in the PUs of C that were encoded before the current PU. $\mathbf{N}_G(r_i)$ is the number of times that each reference frame r_i was used across all previously encoded PUs. For $\zeta = L$, function $\Delta_L(r_\phi)$ only considers the three top-left neighbouring CTUs for counting the number of times each reference frame was used, which is expressed by $\mathbf{N}_L(r_i)$. Overall, (5.6) is used to adjust the penalty weight defined by (5.5), in order to achieve an approximately uniform use of all possible reference frames. This corresponds to the general objective of minimising the values of $\Delta_\zeta(r_\phi)$. The optimum is reached when all reference frames are used to encode the same amount of PUs (i.e., $\Delta_G(r_\phi) = \Delta_L(r_\phi) = 0$), and the original Lagrangian cost is not affected, i.e., $e^{W(r_\phi)} = 1$ in (5.4). Note that an exponential cost increase is used in (5.4) in order to increase the penalty associated to any reference frame that tends to be used much more often than the average.

In (5.5), the constant γ controls the slope of $W(r_\phi)$, leading to higher weights as the value of γ increases. On the one hand a very high value for γ leads to the checkerboard pattern in the reference picture usage, as previously described. On the other hand, a low value of γ reduces the importance of the choice of the reference frames in the Lagrangian cost. Therefore, the impact of the proposed method on the coding efficiency can be controlled through γ. Another way of controlling the impact of the proposed method is the use of the T_{ID}. This allows using lower weights in higher temporal layers, since such frames cannot be used as references for others with lower temporal ID, thus having less impact on error propagation. Therefore, there is a reduction of the impact of the proposed method when hierarchical coding configuration is adopted without compromising the error robustness.

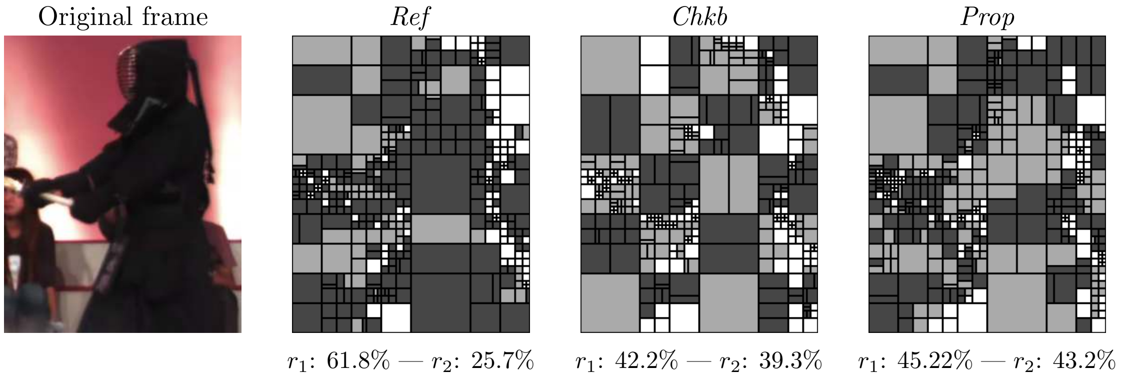

Original frame
Ref
Chkb
Prop
r_1: 61.8% — r_2: 25.7%
r_1: 42.2% — r_2: 39.3%
r_1: 45.22% — r_2: 43.2%

Figure 5.5: Example of the partitions and references pictures selected using the different encoding methods (the dark gray squares indicate that the MV is using the reference r_1, light gray squares correspond to the reference r_2 and unfilled regions are intra-coded blocks).

As an example, Figure 5.4 illustrates the exponential weight used in the proposed method for different values of γ and $T_{ID} = 0$. The horizontal axis represents the sum of $\Delta_G(r_\phi)$ and $\Delta_L(r_\phi)$. For instance, for $\gamma = 5$, when a reference frame r_ϕ is used to predict 3 more CTUs ($\Delta_G(r_\phi) + \Delta_L(r_\phi) = 3$) than the other reference frames (all PUs within those CTUs use the same reference), then the Lagrangian cost is increased by 50% ($e^{W(r_\phi)} \approx 1.5$).

Figure 5.5 illustrates the distribution of reference frames achieved by the proposed method across different PUs. This result is compared against a fixed checkerboard approach (*Chkb*) and the reference HEVC (*Ref*). The dark-gray and light-gray squares indicate the reference frame in use, either r_1 or r_2, respectively. The unfilled squares indicate intra-coded PUs. Below each method, the percentages of the image areas predicted by each reference frame are shown. As shown in this figure, the proposed method is able to distribute the use of the reference frames closely to the *Chkb* method, i.e., r_2 is used as reference for approximately the same amount of PUs as r_1 (see use percentages of the predicted image). However, the predicted regions in the proposed method do not exhibit the same regular pattern as in *Chkb* method. It is noticeable in Figure 5.5 that for several CTUs the *Prop* method is able to keep the same partitions and reference frames when compared to the optimal encoding (*Ref*). This can be observed, for example in the small partitions on the left, where the *Prop* method uses the same reference frame as *Ref*. Thus, the proposed method is expected to achieve better coding efficiency than *Chkb*, since more optimal coding modes may be chosen.

Summarising, by using (5.4) and (5.5), the cost of using a given reference frame increases as the number of PUs using it for prediction increases. However, the reference frames are not hard-selected without taking into account the R-D cost. In general, it is more difficult to reach a uniform use of reference frames in complex sequences than in simpler ones. Note that standard compatibility is maintained, since the proposed optimisation method does not require any syntax change. Moreover, since the proposed approach is employed at the Lagrangian optimisation already implemented in the HEVC reference software, the complexity of the encoder is not significantly affected.

5.5 Streaming stage: MV selection

Since HEVC uses differential and predictive MV coding, whenever a MV is lost the subsequent ones will also be affected, until a refresh point (i.e., MV coded on its own, without using MV prediction) is reached to break the dependency chain. Thus, the amount of error propagation strongly depends on whether the TMVP is used. In the streaming stage, to increase the number of possible refresh points and improve error robustness, a small number of MVs are selected to be transmitted as side information. The selection of these redundant MVs is based on a trade-off between the image area covered by such MVs and the number of bits to encode them. The best trade-off corresponds to the MVs that cover the largest image area using the lowest amount of encoding bits.

Figure 5.6 illustrates how MVs have different importance in error propagation. In this figure, each block corresponds to a PU, which has an associated MV. The arrows point to the PU containing the MV used as predictor. In the current frame (f_0), two PUs have temporally dependent MVs. While the MV of PU (a) is used to predict a total of four other MVs, the MV of PU (b) is only used to predict one single MV. Thus, in case of data loss (e.g., frame f_{-1}), the MV of PU (a) has more impact on error propagation to the frame f_0 and consequently quality degradation than the MV of PU (b), because of its higher number of dependent MVs. For this reason, in the MV selection process, the proposed method assigns higher importance to the MV of PU (a) than PU (b). Moreover, in Figure 5.6 it can be observed that, although the number of MVs predicted from the MVs of PU (c) is higher than PU (a) (6 and 4 MVs, respectively), the image area that depends on the former is smaller than the latter. This is due to the variable size of PUs. Therefore, the image area covered by all dependent PUs is separately considered in the MV selection process.

Based on the evidences mentioned above, the best MVs to be encoded as side information for each frame are selected through the following procedure:

1. the MVs with temporal dependencies are firstly selected as the most important set: $V = \{$*MV: MV is temporally dependent from others*$\}$.

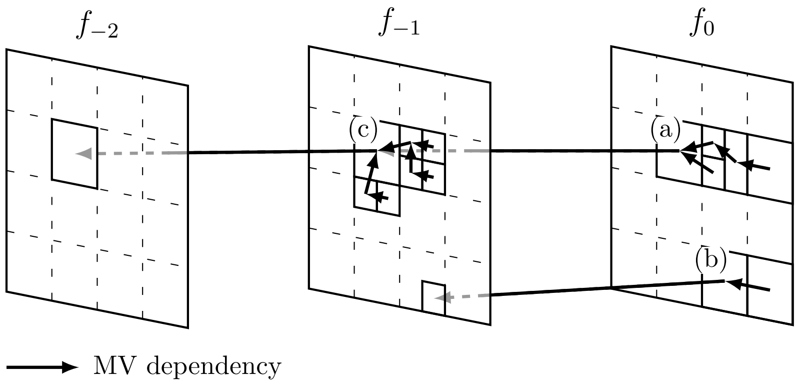


Figure 5.6: Example of temporal and spatial MV dependencies in HEVC.

2. the elements of $MVs \in V$ are ranked according to the number of spatial dependencies associated to each one.

3. the optimal sub-set $U \subset V$ is found as the best trade-off between the image area covered by the dependent PUs and bitrate overhead, i.e., the one that maximises the cost function in (5.7).

This can be formulated as follows. Considering V the ordered set of MVs encoded for the current frame, i.e., $V = \{MV_1, MV_2, \cdots MV_s\}$ and $U \subset V$ always starting in the first vector, i.e., $U_n = \{MV_1, MV_2, \cdots MV_n\}$ with $n \leq s$, the best set of MVs (U_n^*) is selected based on the optimisation procedure given by (5.7). Note that any value of n corresponds to a unique subset U.

$$U_n^* = \arg\max_{n \leq s} \left\{ \sum_{i=1}^{n} A_D(i) - \alpha \frac{t_C - t_I}{T_I} R_U \right\}, \tag{5.7}$$

where R_U is the amount of bits required to encode the n MVs of U and $A_D(i)$ is the total area of the PUs whose MVs are dependently encoded from MV$_i$:

$$A_D(i) = \sum_{j=1}^{DV_i} (w_j \times h_j), \;\; i = 1, 2 \cdots s, \tag{5.8}$$

where DV_i is the number of MVs dependent from MV_i, and w_j and h_j are the width and height of the corresponding PU. A weighting factor is used to control the amount of bitrate overhead (R_U), by increasing the cost of the number of encoded bits based on the duration of the error propagation. This duration ($t_C - t_I/T_I$) is given by the ratio between the distance from the current frame (t_C) to the previous refresh frame (t_I) and the IDR period (T_I) and varies in the range $[0; 1]$, achieving the lowest value for intra-coded frames and the maximum for the frame located immediately before an intra-coded one. The parameter α is used as a global control for redundancy of the entire stream, where higher values of α correspond to less redundant MVs. As pointed in Section 5.3, this streaming stage operates on coded data and may be implemented at any point of the network, to increase robustness without fully decoding the stream. Note that the parsing operation represented in Figure 5.2 does not include decoding of the video frames. Furthermore, only one frame delay is introduced because the MV selection process is performed on a frame-by-frame basis.

To maintain compatibility with the HEVC standard, the redundant motion information is transmitted using the SEI NAL units [9]. Thus, this information can be multiplexed into the coded bitstream without affecting the compressed video. The subset of selected MVs is independently encoded, ensuring that no reference information is needed in the decoder to properly decode the redundant MVs. In this work arithmetic coding [170] was used for the side-information, but any other entropy coding method can be used.

5.6 Performance evaluation

In this section the performance of the proposed two-stage architecture is evaluated and discussed for two cases: (i) only using the encoding stage and (ii) using the streaming stage in addition to the encoding stage. In the former case, identified as "*Prop*", only the dynamic reference picture selection method described in Section 5.4 is evaluated. In the latter case, identified as "*Dyn*", besides the previous method of reference picture selection, redundant MVs are encoded as described in Section 5.5, resulting in a combination of both stages.

As references for comparison, three other methods are used: (i) a fixed checkerboard approach (*Chkb*), (ii) the reference picture generation method based on leaky predictions proposed in [85], identified as "*Ref [85]*", and (iii) long-term reference frames (*Long*). The *Long* method only uses the key frames as reference, i.e., in the RA configuration only one out of eight frames is used as reference for prediction, allowing a more robust transmission, as the reference frames are less likely to be hit by errors.

The experimental setup used in this section is the same as the one previously described in Section 4.2. Two test cases were considered for simulation: (i) in IDR_{Loss} all frames may be equally affected by packet loss, including IDR frames; (ii) in IDR_{NoLoss} IDR frames are assumed to be prioritised in the network and delivered without errors packet loss, or using some combined approach as recently proposed in [21]. At the

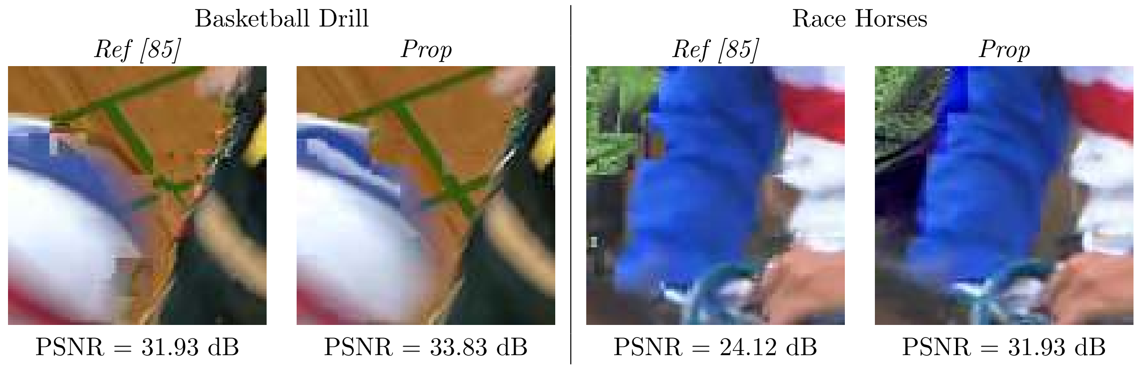


Figure 5.7: Zoom of recovered Frame #10 after a loss event at Frame #9 for Basketball Drill and Race Horses sequences.

decoder-side, EC is based on motion compensation from previously decoded frames. The motion field of a missing frame is obtained by applying MV extrapolation from the two closest neighbours. Afterwards, the final MVs are selected based on the residual energy of their original prediction [152]. The evaluation of the encoding-stage is carried out separately from the streaming-stage.

5.6.1 Encoding stage

The performance of the encoding stage is evaluated on its own, in order to only determine the improvements achieved by the dynamic reference frame selection method. In this case, the reference frames were reconstructed with distortion, as a result of non-perfect recovery. Since the encoding stage does not deal with errors in temporal MV predictions (this is handled by the streaming-stage), in this evaluation the temporal MV predictor is disabled. This is necessary to avoid masking the results with the errors caused by wrong MV predictions.

The performance of the encoding stage (*Prop*) was evaluated using $\gamma = 5$. This value was chosen to allow increasing the Lagrangian cost by about 50%, when a reference frame is used for prediction of more than three CTUs in comparison to the usage of the other reference frames (see Figure 5.4). Nevertheless, further results for a wide range of γ values are also presented.

Quality evaluation under error-prone conditions

The error robustness achieved by the encoding-stage was firstly evaluated in regards to the error propagation, resulting from a single loss event in the LD configuration. Figure 5.7 shows the visual impact of error propagation on a region of the decoded Frame #10 when Frame #9 is affected by errors, for Basketball Drill and Race Horses

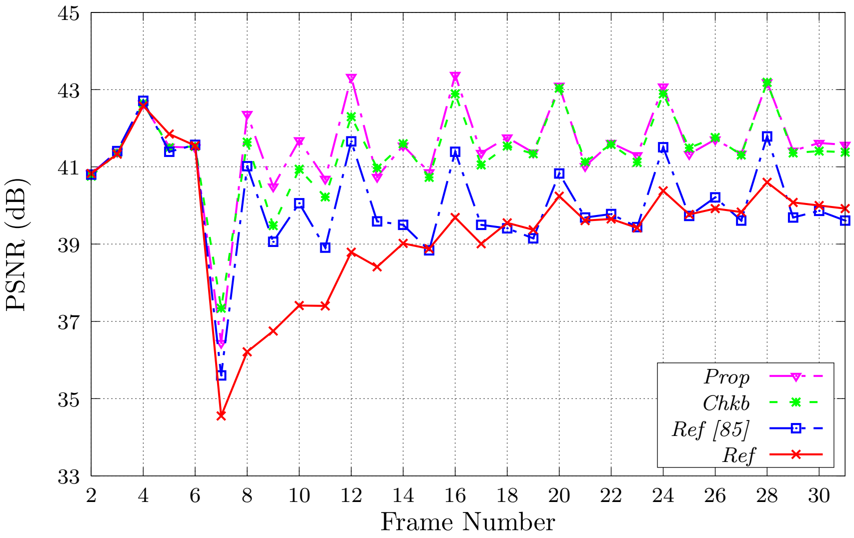


Figure 5.8: Error propagation when errors affect Frame #7 in Kendo sequence (IDR period of 32 frames) for the proposed reference frame selection algorithm in comparison with existing methods.

sequences (the PSNR corresponds to the entire frame). The proposed method is compared with the method in [85], as both aim at reducing the error propagation. The method in [85] uses filtered frames as reference for predictions, therefore, in case of errors, the effects of mismatched predictions are significantly reduced by the filtering operation (see Sub-Section 3.3.3). As shown in Figure 5.7, the proposed method is able to reduce the impact of mismatch decoding in comparison with the method in [85]. Although both methods still reveal some predictions mismatch leading to inaccurate reconstructed of object contours, the proposed method is able to achieve a smoother reconstruction, which leads to higher quality in terms of PSNR. This is because, in frame #10, these predictions that depend on the missing frame #9 are not used throughout the entire frame, thus, approximately half of the predictions are not affected by errors. In contrast, the method in [85] requires several filtering operations to be applied (i.e, different reference frames need to be filtered) in order to reduce the effect of mismatched predictions. Summarising, the reconstructed frames illustrated in Figure 5.7 reveal the higher efficiency achieved by distributing the use of the references frames.

Figure 5.8 shows the PSNR over a GOP with an IDR period of 32 frames, using the same constant bitrate for all streams. This figure shows that all methods are able to outperform the reference HEVC (*Ref*), gaining approximately 3 dB in reconstruction quality. The similar quality levels achieved by *Chkb* and *Prop* reveal that the dynamic selection of the reference frames used in *Prop* is more robust, because a better trade-off between coding efficiency and error resilience is attained. Moreover, results in

Table 5.1: Average PSNR (dB) under random loss events obtained with the proposed reference frame selection and three reference methods for the Low-delay configuration (LD).

Sequence	**Method**	**Packet Loss Ratio**					
		IDR_{Loss}		IDR_{NoLoss}			
		1%	**5%**	**1%**	**3%**	**5%**	**10%**
Basketball Drill	*Ref*	32.25	24.43	36.52	33.70	31.70	28.51
	Ref [85]	-1.05	+0.50	-1.95	+0.12	+1.32	+2.90
	Chkb	-0.39	+0.60	-0.09	+0.69	+0.97	+1.34
	Prop	+0.08	+0.86	+0.45	+1.26	+1.54	+1.78
Book Arrival	*Ref*	34.47	26.95	38.53	35.82	33.63	30.67
	Ref [85]	-1.84	-0.72	-1.76	-0.64	-0.05	-0.04
	Chkb	-0.08	+0.81	+0.49	+1.33	+1.66	+2.15
	Prop	-0.37	+0.56	+0.35	+1.17	+1.41	+1.55
Cactus	*Ref*	34.37	31.94	37.04	35.67	34.51	32.65
	Ref [85]	-0.13	+0.20	+0.31	+0.46	+0.35	+0.83
	Chkb	-0.34	-0.12	+0.02	+0.37	+0.71	+1.00
	Prop	-0.23	+0.04	+0.06	+0.21	+0.43	+0.54
Bosphorus	*Ref*	31.79	24.00	42.19	41.31	40.56	38.74
	Ref [85]	-0.09	-1.04	-0.34	-0.15	+0.09	+0.67
	Chkb	+0.16	-0.49	-0.38	-0.13	+0.13	+0.72
	Prop	+0.23	-0.31	-0.47	-0.18	+0.07	+0.69
BQSquare	*Ref*	32.47	25.11	38.20	36.08	34.57	31.93
	Ref [85]	-4.42	-3.79	-4.42	-3.79	-3.73	-3.74
	Chkb	-0.38	+0.63	-0.12	+0.29	+0.44	+0.67
	Prop	+0.74	+1.37	-0.14	+0.66	+0.88	+1.35
Kendo	*Ref*	38.49	30.89	40.45	36.79	33.80	30.35
	Ref [85]	-1.82	+0.07	-1.82	+0.10	+1.34	+1.30
	Chkb	+0.09	+1.16	+0.24	+1.38	+2.27	+2.15
	Prop	+0.51	+1.33	+0.19	+1.34	+2.26	+2.26
Park Scene	*Ref*	29.81	25.29	36.06	34.83	33.64	31.63
	Ref [85]	-2.29	-1.63	-3.49	-3.40	-3.34	-3.45
	Chkb	-0.05	+0.10	-0.69	-0.20	+0.22	+0.71
	Prop	-0.08	-0.04	-0.50	-0.05	+0.29	+0.86
People on Street	*Ref*	26.07	21.36	32.42	29.61	27.85	25.00
	Ref [85]	+0.26	+0.82	-0.66	+0.42	+0.83	+1.26
	Chkb	+1.05	+0.43	+0.43	+0.95	+1.13	+1.38
	Prop	+0.94	+0.34	+0.58	+1.02	+1.17	+1.42
Race Horses	*Ref*	29.34	23.88	33.00	29.41	27.49	24.55
	Ref [85]	-0.10	+1.25	+0.60	+2.71	+3.31	+3.88
	Chkb	+0.69	+0.64	+0.46	+1.35	+1.47	+1.64
	Prop	+1.11	+0.84	+0.79	+1.80	+1.94	+2.11
Tennis	*Ref*	32.50	25.69	36.50	31.90	29.67	26.15
	Ref [85]	-1.04	+0.12	-1.74	-0.72	-0.08	+1.71
	Chkb	+0.48	+0.45	+0.69	+1.19	+0.81	+1.17
	Prop	+1.15	+1.17	+1.20	+2.03	+1.95	+2.12

Figure 5.8 are in line with the reconstruction quality shown in Figure 5.7, confirming that the proposed method is able to outperform the reference frame generation technique in [85]. Overall, the method proposed for the encoding stage is able to increase the error robustness of HEVC without severely compromising the coding efficiency.

Further tests were run to evaluate the effectiveness of the proposed method under

Table 5.2: Average PSNR (dB) under random loss events obtained with the proposed reference frame selection and three reference methods for the Random-access configuration (RA).

Sequence	**Method**	**Packet Loss Ratio**					
		IDR_{Loss}		IDR_{NoLoss}			
		1%	**5%**	**1%**	**3%**	**5%**	**10%**
Basketball Drill	*Ref*	31.36	25.07	36.90	33.54	31.05	27.88
	Long	+0.09	-0.12	-0.02	+0.31	+0.57	+0.54
	Chkb	+0.12	+0.20	+0.15	+0.46	+0.60	+0.69
	Prop	+0.39	+0.37	+0.26	+0.57	+0.79	+0.87
Book Arrival	*Ref*	33.39	25.48	38.71	35.90	33.01	29.53
	Long	-0.14	+1.01	+0.18	+0.38	+0.96	+0.94
	Chkb	+0.30	+0.10	+0.07	+0.06	+0.31	+0.70
	Prop	+0.26	+1.16	+0.34	+0.28	+0.67	+0.97
Cactus	*Ref*	34.58	32.36	36.92	35.08	33.92	32.12
	Long	-0.32	-0.10	-0.34	-0.32	-0.30	-0.30
	Chkb	+0.15	+0.22	+0.21	+0.36	+0.42	+0.53
	Prop	+0.19	+0.13	+0.19	+0.34	+0.46	+0.50
Bosphorus	*Ref*	32.43	24.86	41.64	39.70	38.28	36.05
	Long	+0.09	-0.87	-0.60	-0.62	-0.61	-0.66
	Chkb	-0.22	+0.37	+0.19	+0.43	+0.55	+0.63
	Prop	+0.21	+0.49	+0.18	+0.47	+0.59	+0.69
BQSquare	*Ref*	33.85	23.63	37.85	33.45	30.44	27.10
	Long	-0.24	+0.46	-0.75	-0.06	-0.04	-0.41
	Chkb	+0.32	+0.62	+0.66	+0.65	+1.06	+1.05
	Prop	+0.42	+0.31	+0.50	+1.11	+0.93	+0.90
Kendo	*Ref*	36.81	28.11	39.78	34.69	30.73	26.62
	Long	+1.04	+0.97	+0.59	+1.43	+1.44	+0.80
	Chkb	+0.55	+1.09	+0.24	+0.62	+1.07	+0.98
	Prop	+0.84	+1.65	+0.37	+0.95	+1.53	+1.42
Park Scene	*Ref*	30.03	25.37	36.74	32.20	29.60	26.35
	Long	-0.15	-0.27	-0.56	-0.44	-0.39	-0.12
	Chkb	+0.24	+0.16	+0.17	+0.51	+0.56	+0.58
	Prop	+0.36	+0.31	+0.65	+0.95	+0.93	+1.04
People on Street	*Ref*	26.04	20.91	30.86	27.21	25.17	22.39
	Long	+0.26	+0.14	-0.22	-0.27	-0.30	-0.27
	Chkb	+0.56	+0.81	+0.56	+0.64	+0.68	+0.68
	Prop	+0.74	+0.83	+0.76	+0.87	+0.96	+0.93
Race Horses	*Ref*	30.43	24.11	33.22	29.14	26.84	23.93
	Long	-0.17	-0.20	+0.09	+0.50	+0.56	+0.34
	Chkb	+0.50	+0.43	+0.29	+0.69	+0.83	+0.89
	Prop	+0.50	+0.38	+0.54	+1.08	+1.10	+1.08
Tennis	*Ref*	31.89	25.89	36.74	32.20	29.60	26.35
	Long	+1.12	+0.58	+0.09	+1.55	+1.61	+1.25
	Chkb	+0.16	+0.12	+0.17	+0.51	+0.56	+0.58
	Prop	+1.46	+1.09	+0.65	+0.95	+0.93	+1.04

various packet loss rates for both test cases, i.e., with and without IDR frame loss, IDR_{Loss} and IDR_{NoLoss}, respectively. The use of long-term reference pictures (*Long*) is only tested for the RA configuration, where key frames are available, while method [85] is tested for the LD configuration. For each test condition, 50 trials were performed and the average quality (PSNR) obtained across all trials for both coding configurations are shown in Table 5.1 and 5.2, for the LD and RA, respectively. In both tables the absolute PSNR is shown for the reference HEVC case (*Ref*), while the PSNR difference is presented for the other cases. Results show that the proposed method is able to outperform other methods for the same loss conditions. An average quality gain up to 2.26 dB is obtained for the LD configuration (see Table 5.1), in comparison to the *Ref*, while the fixed checkerboard pattern approach (*Chkb*) only achieves up to 2.15 dB. In [85], the use of interpolated references in the encoding loop helps to reduce error propagation, at a cost of poor predictions, thus degrading the coding efficiency. This is the main reason that justifies a lower overall performance of the method proposed in [85] (see results for "*Ref [85]*"). The proposed method is also able to outperform the HEVC reference software when an hierarchical coding structure is used (see results for RA configuration in Table 5.2), which implicitly limits the error propagation. In comparison to the reference HEVC, quality improvements up to 1.42 dB at PLR=10% are obtained for Kendo sequence.

When comparing the results for both test cases, it is noticeable that the proposed method is able to outperform the reference methods for both cases, regardless whether IDR frame loss is allowed or not (IDR_{Loss} and IDR_{NoLoss}, respectively). As the PLR increases, the gains of the proposed method decrease in comparison with the case where the IDR frames are delivered without errors (i.e., case IDR_{NoLoss}), due to the higher error propagation of the missing IDR frame. Moreover, the quality gains are lower in the IDR_{Loss} case when comparing with IDR_{NoLoss} for the same PLR, for both the LD and RA. Taking into account the above results and the spatio-temporal information shown in Table 4.1, one can observe that the performance of the proposed method is lower for sequences with high spatial details and low temporal activity (e.g., BQSquare and Park Scene), which indicates higher effectiveness for video content with higher motion activity, as can be seen by the results of Race Horses and Tennis sequences.

Influence of parameter γ

The influence of γ parameter used in (5.5) on the video quality was also evaluated, in order to check how the cost constrains affect the performance of the proposed method. Figure 5.9 shows the average quality gains of the proposed method in comparison to the reference HEVC (*Ref*) across all test sequences, for different values of γ. When no packets are lost, the results in this figure show that increasing γ leads to a decrease in the coding efficiency, especially in the LD configuration. The reason is that higher values of γ lead to higher values of $W(r_\phi)$, therefore the reference picture selection cost is severely constrained by the exponential weights. Since T_{ID} is constant for all

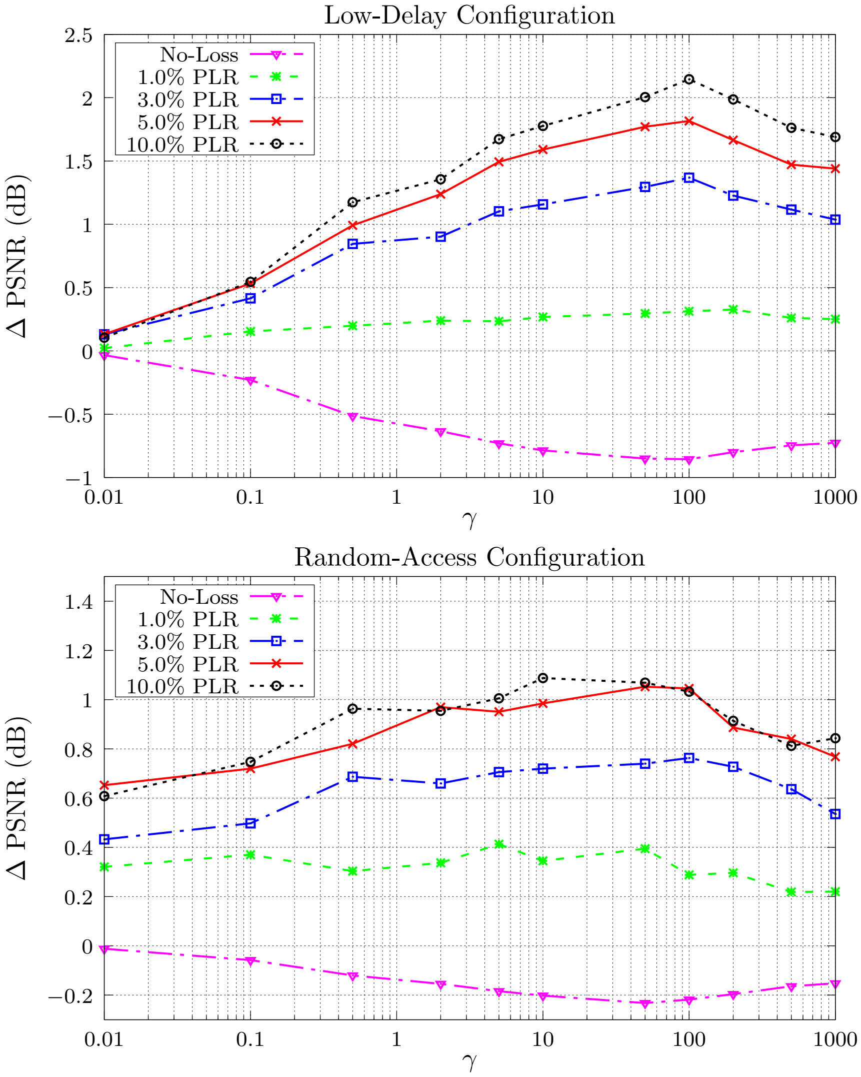


Figure 5.9: Influence of the γ parameter in the average quality gains: comparison between the proposed method and *Ref*, for different PLRs.

frames in the LD, higher weights are used for all frames (T_{ID} is used to reduce the weights in case of hierarchical coding) and the encoder is not able to select the best reference frame in the R-D sense. In the case where packet loss occurs, higher values of γ lead to improved error resilience. Since higher weights are used in the R-D cost, the encoder is constrained to distribute the use of the reference frames of different PUs, reducing the error propagation. For very high values (e.g., $\gamma = 1000$) the encoder is forced to uniformly select the reference frames, resulting in a fixed checkerboard pattern. Overall, the γ parameter is able to control the reference frame selection constrains, which allows to achieve a better trade-off between improved error resilience and loss of coding efficiency. The value of γ is a design parameter that can be selected according to the application requirements and networking conditions, e.g., available

Table 5.3: BD-PSNR (dB) of the proposed reference frame selection algorithm (*Prop*) in comparison with existing methods for the case where no packets are lost.

Sequence	Low-Delay			Random-Access		
	Ref [85]	*Chkb*	*Prop*	*Long*	*Chkb*	*Prop*
Basketball Drill	-1.14	-1.15	-0.51	-0.85	-0.29	-0.22
Kendo	-2.10	-1.39	-0.99	-0.91	-0.39	-0.28
Park Scene	-1.79	-1.38	-0.81	-0.68	-0.38	-0.28
Traffic	-1.91	-1.21	-0.71	-0.55	-0.36	-0.21
Bosphorus	-0.97	-0.95	-1.10	-0.64	-0.20	-0.20
Average	**-1.58**	**-1.22**	**-0.82**	**-0.73**	**-0.32**	**-0.24**

bandwidth or packet loss probability. A possible objective criterion is to choose γ based on the predefined increase percentage of the Lagrangian cost and balancing constrains of reference frame usage. In this work 50% increase of the Lagrangian cost was defined as adequate whenever the unbalance of the reference frame usage is greater than 3 CTUs (see Figure 5.4).

BD-PSNR under packet loss free conditions

This sub-section evaluates the R-D penalty incurred by the proposed method in comparison with *Chkb*, as well as the other reference methods. Table 5.3 shows the Bjontegaard's Delta PSNR (BD-PSNR) values compared to the reference HEVC encoder for different coding configurations in error-free video decoding. Four QPs were used from the common test conditions: 22, 27, 32, 37.

The results in Table 5.3 indicate a small loss of coding efficiency of the proposed method, as well as, the reference methods tested. The fixed checkerboard approach leads to an average quality reduction of 1.22 dB for the LD (0.32 dB for RA), while the proposed method only loses 0.82 dB and 0.24 dB in the LD and RA configurations, respectively. These results clearly show that dynamically selecting the reference frames is more efficient than using a fixed pattern, confirming the previous results for error-prone conditions. Moreover, in comparison to *Long* and the method in [85], the proposed method is still more efficient, because it consistently achieves better results. Comparing, the results for both packet loss free and error-prone conditions, one may conclude that the proposed method not only achieves higher quality gains, but also reduces the impact on coding efficiency. Overall, the loss in the quality compared to the reference HEVC is considered acceptable, given the increase in robustness obtained in lossy transmission.

Table 5.4: Relative CPU times of the tested methods.

Sequence	**Method**		
	Ref [85]	*Chkb*	*Prop*
Basketball Drill	+4.19%	+2.13%	+2.93%
Kendo	+4.96%	+1.83%	+2.05%
Park Scene	+4.98%	+3.90%	+4.04%
Traffic	+8.65%	+2.28%	+3.05%

Complexity overhead

The implementation of this method slightly increases the complexity of a standard non-robust encoder. Therefore, the computational complexity has also been considered as a performance metric in this research, which is measured as the average encoding time of five runs, in a controlled hardware platform. The reference used for comparison with other methods is the HEVC reference software encoding time, running on the same platform with the same coding configuration. The relative complexity increase of the robust encoder using *Prop*, *Chkb* and the method presented in [85] is shown in Table 5.4.

As shown in this table, the method presented in [85] is the most time consuming because it requires post-processing of the encoded frames in order to interpolate new references. The proposed scheme, using a reference picture selection, presents a lower complexity increase than [85], while the fixed checkerboard scheme (*Chkb*) presents a slightly lower complexity than *Prop* method. This is due to the fact that in the *Chkb* method there is no R-D optimisation process for the reference frame selection (fixed pattern is used), thus reducing the overall computational complexity.

5.6.2 Streaming stage

In this section the coding performance that results from transmitting redundant MV (streaming stage) in combination with the reference picture selection scheme (encoding stage) is presented. In these experimental tests the temporal MV candidates are enabled and MVs are dynamically added as side-information into the stream, as described in Section 5.5. Different values of α are used to achieve different overhead ratios. All streams were encoded with the same total bitrate, including the redundant bits, in order to make a fair quality comparison.

Dynamic MV selection performance

Firstly the quality obtained by the proposed method, which dynamically selects the amount of redundant MVs, is compared to cases where a fixed amount of MVs per frame are transmitted (i.e., the most important 10%, 30% and 50% MV $\in V$). These cases are referred to as "*Dyn* " and "*Fixed* ", respectively.

Figure 5.10 illustrates the relation between the quality obtained (PSNR) and the amount of redundancy used. The results indicate that for approximately the same amount of redundant bits are used in both cases, however by using the proposed method to dynamically select the MVs to be used as side-information (*Dyn*) is able to outperform the fixed case for all test sequences. Results also show that for both approaches the decoded video quality increases as the redundancy of the side-information increases, since a higher number of mismatch MV predictions are able to be recovered.

It is noticeable in Figure 5.10 that for higher PLRs, the proposed method is able to achieve higher gains when compared with the fixed approach. This reveals that when data loss increases it is more important to optimise the amount of MVs selected per frame, in order to use more MVs in those frames that have a higher impact on the error propagation. For instance, the proposed method is able to gain up to 1.5 dB for the BQSquare sequence subject to a PLR=10%. For lower PLRs, since fewer frames are affected by data losses, the use of a fixed approach achieves similar performance as the proposed method (*Dyn*).

Quality evaluation

Table 5.5 presents the average PSNR of decoded video for different PLR (columns 4 to 8) and for various percentages of MV redundancy (third column). The absolute quality values are shown for the reference case (*Without MVs*) and PSNR gains are shown for three different levels of redundancy introduced by the proposed method (*Dyn*). This study was also conducted for the two different cases, with and without IDR loss.

It can be confirmed from the results of Table 5.5 that, for each sequence, as α decreases more redundant MVs are used, leading to higher overhead rations. This results in higher objective video quality for all tested sequences. These results also confirm that redundant MVs are able to minimise the error propagation, which is accomplished at the cost of an acceptable increase in the bitrate, since only a sub-set of MVs is transmitted. The use of redundant MVs yields better video quality for PLR as low as 1%, achieving higher quality gains when IDR frames are not lost (IDR_{NoLoss}). This is because of the higher quality degradation when IDR frames are lost, which is

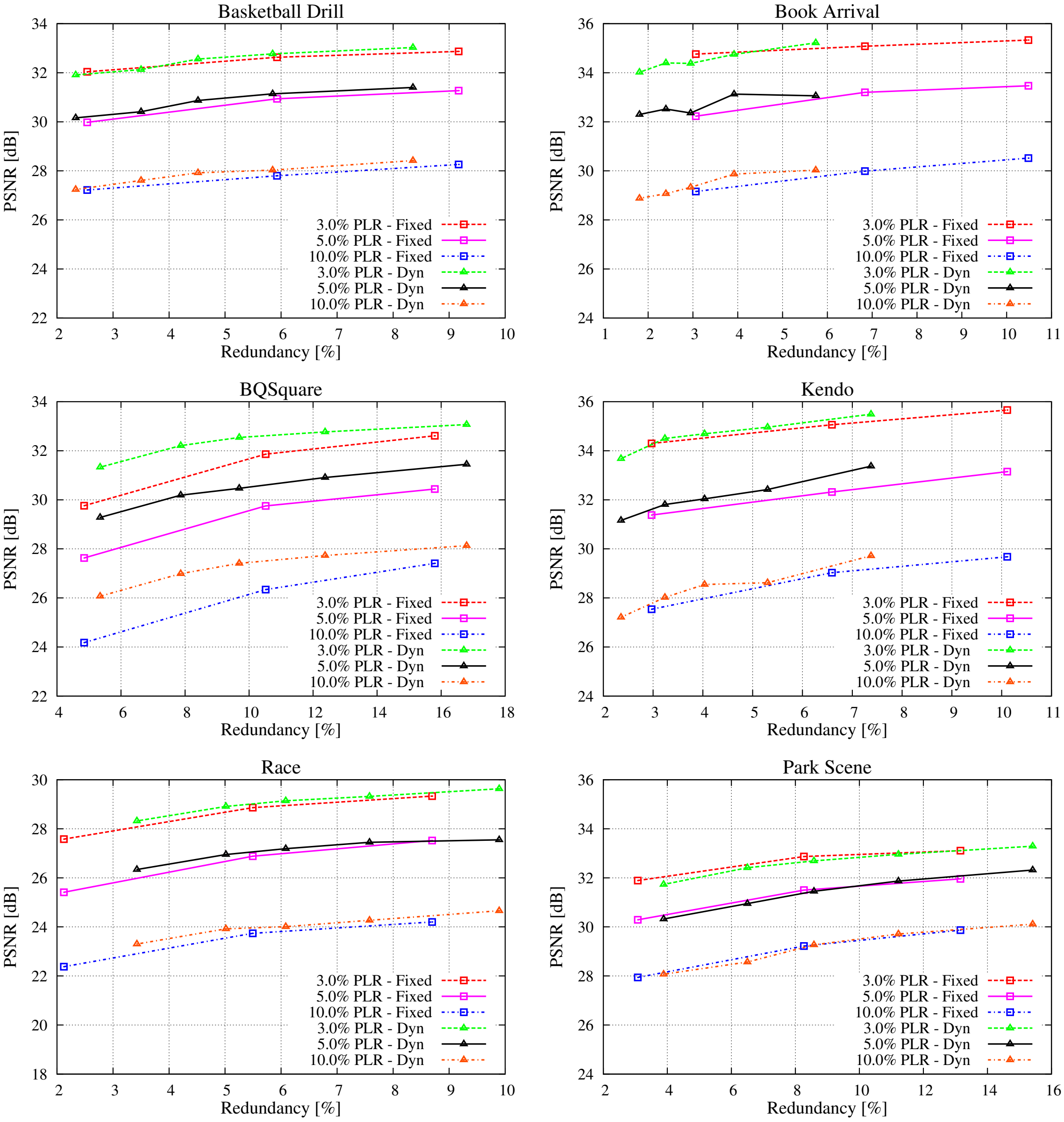


Figure 5.10: Decoded video quality for different amounts of redundancy, selected by using a fixed approach and the proposed dynamic method (*Dyn*).

not due to mismatched MV predictions, and thus cannot be reduced by transmitting extra MVs as side-information. The use of the proposed method also achieves higher quality for sequences with higher motion activity that make use of inter-prediction more often, i.e., BQSquare and Traffic sequences (see Inter ratio in Table 4.1), leading to a maximum gain of 3.93 dB for IDR_{Loss} (7.82 dB for IDR_{NoLoss}).

In summary, the proposed method is able to select the most relevant motion MV information and when combined with a reference picture selection scheme the problem of mismatch MV predictions is mitigated. Moreover, the dynamic approach devised to select the best MVs also contributes to increase the overall performance. Thus,

Table 5.5: Average PSNR (dB) using redundant MV selected by the proposed approach.

Sequence	**Method**	**Overhead ratio (%)**	**Packet Loss Ratio**				
			IDR_{Loss}		IDR_{NoLoss}		
			1%	**5%**	**1%**	**5%**	**10%**
Basketball Drill	*Without MVs*	–	30.41	23.31	34.76	28.56	25.89
	Dyn (α=1.5)	2.33	+0.68	+0.41	+0.58	+1.60	+1.35
	Dyn (α=0.8)	4.52	+0.96	+0.96	+0.95	+2.31	+2.03
	Dyn (α=0.6)	5.85	+1.12	+1.16	+0.93	+2.58	+2.14
Book Arrival	*Without MVs*	–	33.63	25.59	37.47	31.60	28.27
	Dyn (α=1.5)	1.81	+0.38	+0.14	+0.02	+0.70	+0.61
	Dyn (α=0.8)	2.95	+0.23	+1.15	+0.49	+0.76	+1.06
	Dyn (α=0.6)	3.92	+0.44	+1.80	+0.34	+1.53	+1.60
BQSquare	*Without MVs*	–	28.66	20.22	31.00	23.09	20.21
	Dyn (α=1.5)	5.35	+3.04	+3.94	+4.05	+6.19	+5.86
	Dyn (α=0.8)	9.69	+3.23	+3.80	+5.03	+7.38	+7.20
	Dyn (α=0.6)	12.38	+3.54	+3.93	+5.13	+7.82	+7.52
Kendo	*Without MVs*	–	36.21	28.38	38.23	29.58	25.52
	Dyn (α=1.5)	2.36	+0.78	+1.16	+0.36	+1.58	+1.69
	Dyn (α=0.8)	4.03	+0.94	+1.32	+0.79	+2.46	+3.03
	Dyn (α=0.6)	5.30	+0.86	+1.08	+0.97	+2.84	+3.10
Park Scene	*Without MVs*	–	28.60	23.85	33.02	28.26	25.98
	Dyn (α=1.5)	3.88	+0.53	+0.72	+1.05	+2.06	+2.09
	Dyn (α=0.8)	8.58	+0.79	+0.96	+1.55	+3.19	+3.29
	Dyn (α=0.6)	11.22	+0.73	+1.05	+1.58	+3.61	+3.73
Race Horses	*Without MVs*	–	26.87	21.78	29.67	23.23	20.48
	Dyn (α=1.5)	3.43	+2.46	+1.45	+2.28	+3.11	+2.82
	Dyn (α=0.8)	6.08	+2.76	+2.23	+2.88	+3.96	+3.53
	Dyn (α=0.6)	7.58	+2.93	+2.27	+2.89	+4.22	+3.79
Tennis	*Without MVs*	–	31.27	24.81	34.33	27.09	23.96
	Dyn (α=1.5)	1.63	+0.92	+0.69	+0.87	+1.42	+1.20
	Dyn (α=0.8)	3.12	+1.26	+0.76	+1.33	+1.93	+1.73
	Dyn (α=0.6)	4.10	+0.97	+1.00	+1.45	+2.17	+2.07
Traffic	*Without MVs*	–	27.38	22.22	33.92	27.92	24.86
	Dyn (α=1.5)	2.09	+0.30	+0.48	+0.95	+1.61	+1.66
	Dyn (α=0.8)	5.67	+0.82	+0.95	+1.82	+3.59	+3.67
	Dyn (α=0.6)	7.98	+0.96	+0.87	+2.07	+3.89	+4.12

using the proposed two-stage approach to recovering temporal predictions consistently improves the error robustness of HEVC.

5.7 Summary

In this chapter a two-stage approach was proposed to increase the error robustness of HEVC streaming and reduce the error propagation in case of packet loss. A constrained coding approach was devised to select reference frames and dynamically distribute temporal dependencies. This is jointly used with a controlled amount of side information in coded streams, comprising a small set of the most relevant MVs, in order to minimise MV mismatch at the decoder in the presence of frame loss. The use of MVs as side information was firstly proposed and evaluated in the publication (C2) and the reference frame selection was further discussed in the publication (C4). Subsequently, a wider analysis was carried out and published in (J1). As can be concluded from the results, this method contributes to reduce the temporal dependencies with consistent quality gains for different PLRs and coding configurations, only incurring a small drop of coding efficiency. Overall, the proposed approach is an effective method to cope with video transmission over error-prone networks.

Chapter 6

Error concealment-aware video encoding

6.1 Introduction

In this chapter an error concealment-aware encoding scheme is proposed to improve the quality of video delivered over networks prone to errors and data loss. The method aims to enhance the reconstruction of recovered lost frames by optimising the efficiency of the EC. The proposed scheme is based on a scalable coding approach where the best EC methods to be used at the decoder are optimally determined at the encoder and signalled through SEI messages. Such optimal EC modes are found by simulating transmission losses followed by a Lagrangian optimisation of the *signalling rate - EC distortion* cost. A generalised saliency-weighted distortion is used and the residue between coded frames and their EC substitutes is encoded using a rate-controlled enhancement layer. When data loss occurs the decoder uses the signalling information to improve the reconstruction quality.

This chapter is organized as follows. Section 6.2 introduces the problem definition and motivation for the method proposed in this chapter. Section 6.3 presents an initial evaluation of different EC schemes in order to find the relevant ones to be used in the proposed EC-aware coding scheme. Section 6.4 provides an overview of the proposed method architecture followed by a detailed description in Section 6.5. Sections 6.6 and 6.7 present the performance results and discussion of the proposed EC optimisation and signalling. Finally, Section 6.8 concludes the chapter.

6.2 Problem definition

In Chapter 3 different EC algorithms that can effectively reconstruct lost video were described, though not all of them exhibit the same accuracy for every lost frame or slice. Moreover, it was also shown that combining different EC approaches at the block level for improving the reconstruction accuracy is not straightforward. In the past, hybrid EC schemes were proposed to improve the quality of recovered frames by either combining spatial and temporal EC algorithms [171] or by optimising the MVE at the pixel and block levels [151, 172]. Regarding the HEVC, existing algorithms also aim to improve the reconstruction accuracy of motion-based recovery approaches by using the information of correctly received neighbouring frames, e.g., block partitions [25] and prediction residue [152]. Although such methods are able to reduce the reconstruction artefacts and improve the overall video quality, decoder-based decisions are still not fully accurate, as they are based on limited information to be used as reference in the decision process. Therefore, higher reconstructed quality can be achieved by estimating the best EC techniques at the encoding-side, based on the original video signal.. Different approaches were proposed in the past to address the problem of EC-aware video coding [47, 49, 50]. Extra information was also used to aid the EC operation at the decoder-side, either by sending redundant information [153, 154] or by embedding additional coded data into the bitstream [51]. Although these methods facilitate the EC performance either through slice reordering or based on MVs, they always rely on the assumption that a fixed EC method is used at the decoder.

To advance one step further, this chapter presents an investigation of EC-aware video coding and proposes a new scheme to achieve high decoding flexibility in terms of the EC approaches, leading to better video quality. This is accomplished by selecting, at the encoder, among several EC algorithms with different characteristics, the one that yields the lowest reconstruction distortion at the decoder and also minimises the R-D cost considering the necessary signalling rate. Thus, the proposed approach differs from similar previous methods [47, 50] by selecting the best EC method without affecting the encoder optimisation. To further improve the quality of recovered frames, an enhancement layer is encoded with the mismatch residue computed as a difference between the transmitted and recovered frames. This is a dynamic optimisation process, which is more efficient than conventional approaches based on redundant pictures [120, 121], which transmit the whole frame as redundancy. In contrast with the previous work presented in Chapter 5, which addressed the error resilience problem by optimising the coding dependencies, the approach described in this chapter considers both, EC performance and the coded rate required for signalling the best EC mode to the decoder.

6.3 Error concealment algorithms

In order to propose a novel EC-aware technique different algorithms were taken into consideration. This section presents a description and evaluation of the most relevant EC algorithms used in the proposed method. Firstly, a technical description of each method is provided followed by a performance comparison between them.

6.3.1 EC algorithms based on motion-compensation

The most common algorithms to recover missing frames take advantage of the correctly received MVs from the neighbouring frames. This has been widely used in the past for H.264/AVC standard [140, 146, 147] and also in the HEVC standard [25, 152, 173].

Let's consider the example of Figure 6.1, where a missing frame (f_0) at time instant t_0 is reconstructed based on the information present in the temporally adjacent frames, f_{-1} and f_{-2}. One of the common approaches is MC where the missing frame is reconstructed through motion compensation directly using the co-located vectors in the closest neighbour frame (f_{-1}). The MVs used for the motion compensation (v_0) are obtained by copying the MVs v_{-1} from the same spatial position in f_{-1}, as illustrated in the figure by the MV v_0^1 which is obtained by copying v_{-1}^1.

An alternative approach is based on the reconstruction of the missing frame by using motion compensation through extrapolation of the received MVs, of frame f_{-1} at time instant t_{-1}, according to the following procedure:

1. the MVs (v_{-1}) of the neighbouring frame (f_{-1}) are extrapolated to compute the set of MVs (v_0) for the missing frame as follows:

$$v_0 = \frac{t_0 - t_{-1}}{t_{-1} - t_R} \times v_{-1}, \tag{6.1}$$

   where t_R is the time instant of the reference frame pointed by the original MVs v_{-1}, which in the example of Figure 6.1 corresponds to t_{-2}.

2. the MV associated to the block $b_{-1}(x, y)$ at spatial position (x, y) in f_{-1} may be used to recover the block $b_0'(x', y')$ in f_0, at position (x', y') obtained from v_0

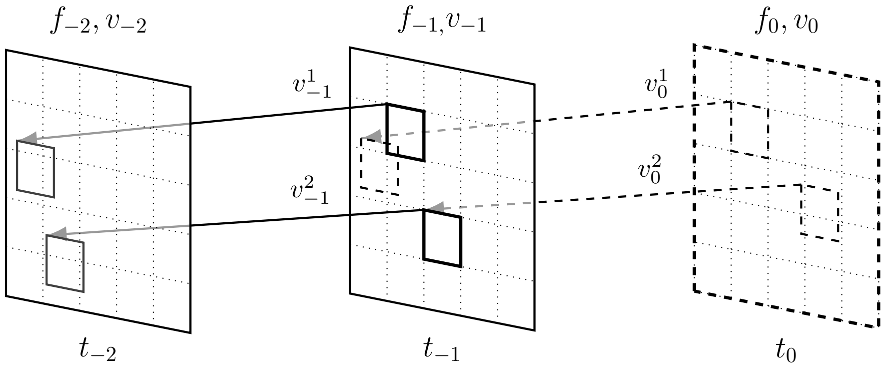


Figure 6.1: Representation of the video frame with associated original and extrapolated motion vectors represented using full and dashed arrows, respectively.

components as follows:

$$x^{'} = x - v_{0x}, \tag{6.2}$$

$$y^{'} = y - v_{0y}. \tag{6.3}$$

This method, referred to as Motion Vector Extrapolation (MVE), is illustrated in Figure 6.1, where the MV v_0^2 results from the extrapolation of v_{-1}^2. One should note that these methods may not result in a complete motion field for the entire frame, resulting in an incomplete reconstruction. For the regions without MVs the average of the three top-left neighbouring MVs are used to complete the motion field.

Finally, as a base for comparison, the default EC in the reference HEVC implementation the FC algorithm is also evaluated. This method consists in replacing the lost frame with a copy of the previously decoded one, which can also be described as a motion compensated reconstruction using only zero MVs.

6.3.2 Algorithms based on motion estimation

In this sub-section a different method to estimate a motion-field is described. This can be used to reconstruct the missing motion information of a lost frame. Then, the recovered motion-field is used in combination with the MVE method to recover the lost frame.

Typically, optical flow estimation algorithms are used to estimate a 2-D motion field, comprising one MV $v = (v_x, v_y)$, for each pixel, which minimises the following energy function:

$$E(x,y) = \sum_x \sum_y \left| f_0\Big(x,y\Big) - f_{-1}\Big(x + v_x(x,y), y + v_y(x,y)\Big) \right|. \tag{6.4}$$

Besides a full search algorithm, where a block-based approach is used to search for the best MVs that minimise the sum of absolute errors, three other approaches were tested to evaluate their effectiveness in EC algorithms. These three methods are as follows:

1. a method based on the duality between total pixel difference and regularisation term [174];

2. a method which approximates the neighbourhood of each pixel by quadratic polynomials on both frames [175];

3. a method based on a combination of feature matching and total pixel difference [176].

Motion estimation based on a regularisation term (*DualFlow*)

In this method the motion field is estimated based on the minimization of the total difference of pixel values measured using the L-1 norm [174,177], taking into account a regularisation term obtained from the Horn and Schunck smoothness condition [178]. This condition forces the estimated field to be regular across neighbouring pixels by penalising deviations in a quadratic sense. Using this method, for each pixel $\mathbf{p}$, the motion field $v(\mathbf{p})$ is obtained by minimising the following energy function,

$$E(v) = \sum_{\mathbf{p}\in\Omega} \left(\Big| f_{-1}(\mathbf{p}) - f_{-2}(\mathbf{p} + v(\mathbf{p})) \Big| + \Big| \nabla v(\mathbf{p}) \Big| \right) \tag{6.5}$$

where f_{-1} and f_{-2} are the two previous frames at time instants t_{-1} and t_{-2}, as represented in Figure 6.1. In (6.5), Ω defines the sets of pixels of those frames, and $|\nabla v(\mathbf{p})|$ is a regularisation term that penalises high variations of the optical flow which effectively disallow discontinuities. This is based on the assumption that neighbouring regions belong to the same object, thus such regions should have similar motion.

Motion estimation based on polynomial expansion (*PolyFlow*)

In this method the motion field is estimated based on a polynomial expansion of each pixel neighbourhood [175]. In this method only quadratic polynomials are considered, expressed as:

$$f(\mathbf{p}) \sim \mathbf{p}\mathbf{A}_1\mathbf{p}^T + \mathbf{b}_1\mathbf{p}^T + C_1 \tag{6.6}$$

where $\mathbf{p} = [x\ y]$, $\mathbf{A_1}$ is a symmetric matrix, $\mathbf{b_1}$ is a row vector and C_1 is a scalar. The coefficients are estimated using weighted least squares fit to the signal values in the neighbourhood. The use of quadratic polynomials was chosen primarily due to their low computational complexity and acceptable estimation accuracy. Nevertheless, as stated in [179] the extension to higher degrees is straightforward.

Since the result of a polynomial expansion is such that each neighbourhood is approximated by a polynomial, it is necessary to analyse what happens if a polynomial undergoes an ideal translation. Let's consider the signal $f_1(\mathbf{p})$ represented by (6.6) and a signal f_2 which corresponds to a displacement of f_1, given by $\mathbf{d} = [v_x\ v_y]$. This can be expressed by the following:

$$\begin{aligned}
f_2(\mathbf{p}) &= f_1(\mathbf{p} - \mathbf{d}) \\
&= (\mathbf{p} - \mathbf{d})\mathbf{A}_1(\mathbf{p} - \mathbf{d})^T + \mathbf{b}_1(\mathbf{p} - \mathbf{d})^T + C_1 && (6.7) \\
&= \mathbf{p}\mathbf{A}_1\mathbf{p}^T - 2\mathbf{d}\mathbf{A}_1\mathbf{p}^T + \mathbf{d}\mathbf{A}_1\mathbf{d}^T + \mathbf{b}_1\mathbf{p}^T - \mathbf{b}_1\mathbf{d}^T + C_1 \\
&= \mathbf{p}\mathbf{A}_1\mathbf{p}^T + (\mathbf{b}_1 - 2\mathbf{d}\mathbf{A}_1)\mathbf{p}^T + \mathbf{d}\mathbf{A}_1\mathbf{d}^T - \mathbf{b}_1\mathbf{d}^T + C_1 && (6.8) \\
&= \mathbf{p}A_2\mathbf{p}^T + \mathbf{b}_2\mathbf{p}^T + C_2,
\end{aligned}$$

then, by equating the coefficients of the quadratic polynomials results in the following relations:

$$\begin{aligned}
\mathbf{A}_2 &= \mathbf{A}_1 \\
\mathbf{b}_2 &= \mathbf{b}_1 - 2\mathbf{d}\mathbf{A}_1, \\
C_2 &= \mathbf{d}\mathbf{A}_1\mathbf{d}^T + \mathbf{b}_1\mathbf{d}^T + C_1.
\end{aligned} \tag{6.9}$$

This allows to find the displacement $\mathbf{d}$ for any signal dimensions by solving the above equations. Detailed implementations of the polynomial expansion and respecting solving can be found in [179, 180].

Large displacement optical flow with deep matching (*DeepFlow*)

In this method the motion field is estimated based on extension of the total pixel difference methods using descriptor (feature) matching. This allows the optical flow estimation to accurately find large displacements, because it is not fully bound by any constrains, such as the Horn & Such presented in (6.5) [181]. Moreover, the correspon-

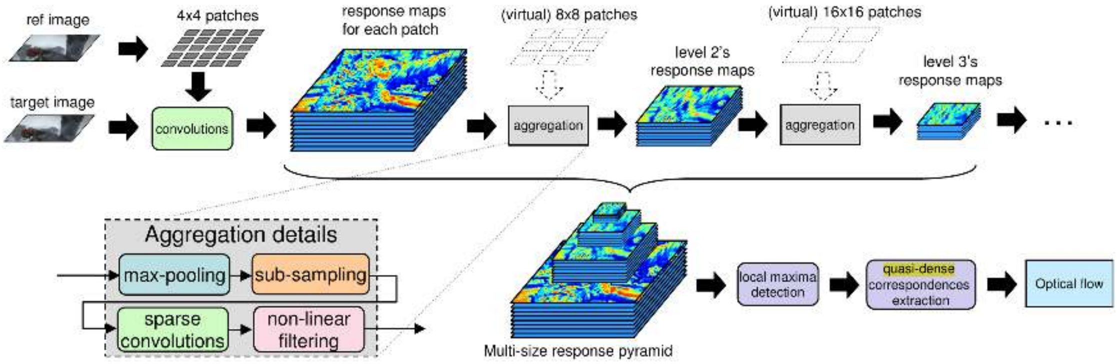


Figure 6.2: Block diagram of the deep matching algorithm [176].

dences between descriptions are used to apply a coarse-to-fine wrapping, which avoids local minimum values. The combination of both approaches increases the matching accuracy.

This approach was also exploited in [176] with the following extensions: (i) use of description matching in the minimisation problem; (ii) introduction of a normalisation parameter to decrease the impact of regions with high spatial changes; (iii) use of different weights for each layer optimisation. Figure 6.2 shows the blocks that are included in the description matching process. The matching algorithm is built upon a multi-stage architecture using six layers with different patch sizes, interleaving convolutions and max-pooling, which is similar to deep convolutional networks [182]. As shown in Figure 6.2, multi-size response levels are generated from the reference and target images using convolutions with different patch sizes. This method uses a bottom-up approach, i.e., starting at a fine level and moving towards the coarser levels (larger patches), which are built as an aggregation of responses of smaller patches. This results in a multi-size response pyramid. Subsequently, a local maximum is extracted from each level to achieve a more accurate motion field (quasi-dense correspondence). Then, an initial motion field is extracted, referred to as v' . The final optical flow is obtained by minimising an energy function, which can be described by the following equation:

$$E(v) = \sum_{\mathbf{p}\in\Omega} \Big(E_D(\mathbf{p}) + \alpha E_S(\mathbf{p}) + \beta E_M(\mathbf{p}) \Big), \tag{6.10}$$

where E_D is the data term, E_S is a smoothing term and E_M is the matching term. The data term is obtained from the difference between matching pixels in both video frames. The smoothness term is a robust penalisation of the gradient of the flow, similarly to (6.5), i.e., $E_S(\mathbf{p}) = |\nabla v(\mathbf{p})|$. Finally, the matching term (E_M) is used to encourage the flow estimation to be similar to the precomputed motion field (v').

Table 6.1: Average decoded video quality (PSNR in dB) for seven error concealment algorithms under 10% of random packet loss.

Sequence	*FC*	*MC*	*MVE*	*Full Search*	*Dual Flow*	*Poly Flow*	*Deep Flow*
Basketball Drill	28.22	+2.43	+2.56	+2.61	+2.71	+3.08	+3.33
Book Arrival	31.01	+2.50	+2.49	+2.31	+2.81	+3.05	+3.39
BQSquare	31.36	+2.28	+2.32	+0.43	+1.38	+1.93	+1.88
Cactus	31.93	+0.91	+0.96	+0.77	+0.96	+1.05	+1.08
Four People	37.12	+1.24	+1.26	+0.33	+1.38	+1.44	+1.44
Kendo	31.57	+2.34	+2.36	+2.04	+2.80	+2.78	+3.05
Kimono	32.24	+0.12	+0.12	+0.10	+0.13	+0.13	+0.14
Park Scene	30.42	+1.30	+1.38	+1.21	+1.23	+1.50	+1.57
People on Street	25.05	+0.40	+0.43	+0.36	+0.56	+0.61	+0.61
Race Horses	24.19	+0.68	+0.67	+0.66	+0.66	+0.76	+0.81
Tennis	27.08	+1.34	+1.56	+1.17	+1.65	+0.74	+2.37
Average	–	**+1.41**	**+1.46**	**+1.09**	**+1.48**	**+1.55**	**+1.79**

6.3.3 Experimental evaluation

This section presents a simulation study of the different EC schemes previously described. This evaluation aims at finding the suitable methods to be used for efficient reconstruction of the missing frames in case of data loss. The two EC methods based on the correctly received MVs are tested, i.e., MC and MVE, while the FC method is used as reference for comparison. Moreover, the four different approaches to estimating the motion field for the missing frame are compared: (i) full search based on the minimisation of the sum of absolute errors (*FullSearch*), (ii) optical flow based on the duality between total difference and regularisation term (*DualFlow*), (iii) optical flow based on polynomial approximations (*PolyFlow*) and (iv) optical flow based on deep matching (*DeepFlow*).

Table 6.1 shows the average quality (PSNR) obtained for the different methods under investigation. The absolute value of PSNR is shown for the *FC* method, while the differential values are shown for the remaining methods. The results shown in the table reveal that methods *MC* and *MVE* are both able to overcome *FC*, achieving average quality gains of up to 1.41 dB and 1.46 dB, respectively. Moreover, by extrapolating the MVs from the correctly receiving frame to the lost one (*MVE*), it is also possible to outperform motion-copy (*MC*). This can be explained by the objects' trajectories of objects which are preserved in the *MVE* method (assuming constant translational motion), in contrast with the *MC* method, which assumes that neighbouring objects have parallel motion which may result in inaccurate estimations.

According to the results of Table 6.1 (columns 5 to 8) the EC methods based on motion estimation achieve an overall superior performance when compared with methods which rely on correctly received MVs (i.e., *MC* and *MVE*). Results for *FullSearch* method reveal that only minimising the sum of absolute errors between two consecutive frames might not always be directly related with objects' motions, which results in inaccurate motion field to be used in the reconstruction of lost frames. Finally, results from Table 6.1 reveal that the *DeepFlow* algorithm leads to the most accurate motion field, in comparison with the other methods. It is able to achieve an average quality gain of 1.79 dB, and up to 3.39 dB for Book Arrival sequence, where *PolyFlow* only achieves 3.05 dB. Since *PolyFlow* combines different techniques also used by other algorithms in study, e.g., total difference of pixel values also used by *DualFlow*, it is expected to achieve accurate estimations and consequently higher EC reconstruction quality. Moreover, using feature matching in combination with smoothing conditions, the *DeepFlow* algorithm is able to correctly estimate large displacements, which results in improved performance for video signals with higher motion, e.g., Tennis sequence.

Remarks

Summarising, the analysis of the previous results indicates that all EC methods are able to cope with lost frames and manage to recover the missing data with reasonable quality. Nevertheless, it is noticeable that lost frames recovered with an estimated motion field result in higher average quality, especially when using the *DeepFlow* method. However, since the other methods are also able to achieve acceptable video quality, they shall not be neglected, as they provide alternative approaches for lost frame recovery with a lower computational complexity (i.e., they not require motion estimation). Thus, in order to achieve higher reconstruction quality, different approaches shall be considered, aiming at balancing the overall complexity with the quality of the reconstructed frames.

6.4 Architecture for the EC-aware video encoding

This section presents the general architecture proposed for EC-aware video encoding, describing its main functional blocks, and providing an introduction for the subsequent detailed descriptions. Figure 6.3 illustrates the architecture of the proposed EC-aware encoding scheme. After the video encoder, it includes data loss simulation at the NAL for every slice. Subsequently, EC is applied to the simulated losses in order to find the best methods that should be used to reconstruct the lost slices in case of errors and packet loss. Saliency-weighted EC mode decision is used to select the optimal EC

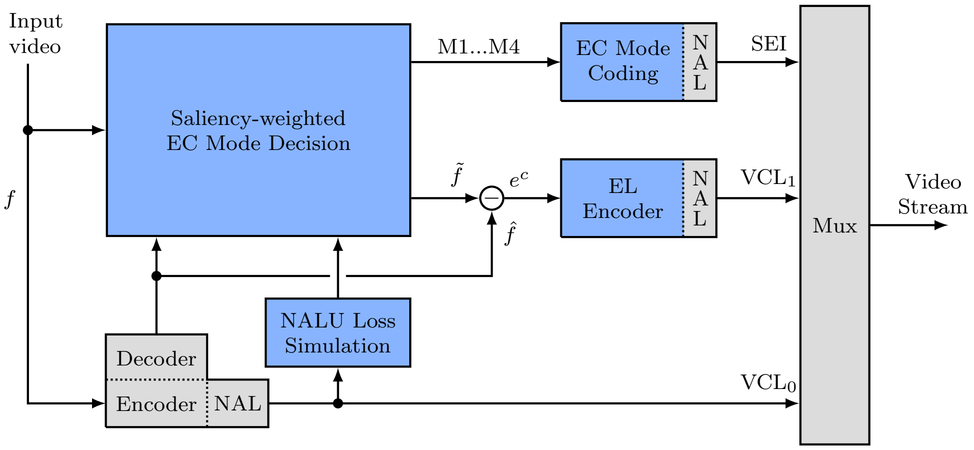


Figure 6.3: Architecture of the error concealment-aware encoding scheme.

mode from the best trade-off between EC distortion, using the original input video (f) as reference, and the number of bits required for signalling. Then, the optimal EC mode is sent to the decoder along with the mismatch residue computed as a difference between the transmitted and recovered frames. Finally, the scalable extension of the HEVC standard is used to encode this residue information.

For the EC mode decision optimization the missing slices are reconstructed, as they would be in the decoder, i.e., they are recovered at the encoder by simulating the decoder EC, but using several EC methods, based on motion estimation and extrapolation techniques, to select the best one at the encoder-side. In this process, each CTU is partitioned to form a quadtree structure of Error Concealment Units (ECU) that are used to test the different EC candidate modes. Then, the best trade-off between ECU partition and signalling is found through R-D optimisation. In order to achieve an efficient use of the signalling overhead, a saliency-weighted procedure is embedded in the R-D optimisation based on a spatio-temporal saliency, where more signalling bits are allocated to regions with higher saliency values. Finally, the signalling information is transmitted to the decoder by using the SEI NAL units [9]. Since, in general, the optimal EC mode is still not able to perfectly reconstruct all the missing regions, extra residual information is multiplexed in the stream encoded as a scalable EL. This residual information contains the difference between the compressed ($\hat{f}_t$) and the EC frames ($\tilde{f}$).

The transmitted information is used to assist the EC operation at the decoder to recover erroneous frames with high accuracy, by using the best EC mode and also residual information to reduce error propagation. The remaining of this section describes in detail the building blocks of the proposed encoding architecture, i.e., the EC candidate modes, the saliency-weighted optimal EC mode decision, signalling and enhancement layer coding.

6.5 Saliency-weighted EC mode decision

In this section the proposed saliency-weighted EC mode decision algorithm is described. Firstly, the list of EC candidates is provided based on the previously described methods. Then, the R-D optimisation is performed using a saliency region in each frame, in order to decrease the overall bitrate without compromising the reconstruction quality. The signalling coding strategy is also described. Finally, coding the EC residue, introduced as auxiliary information to improve the reconstruction quality, is described.

6.5.1 EC candidates

A set of four EC candidate modes (M1...M4) is used in the optimisation process that selects one of them as the best match for each particular ECU. Each EC mode corresponds to a different EC method, as shown in Figure 6.4, namely:

- M1, M2: These two methods are based on extrapolation of an estimated motion field;
- M3: This method is based on MV extrapolation of previously received MVs;
- M4: This method uses the co-located motion information from the closest neighbour;

The first two methods, M1 and M2 rely on an estimated motion field, using the method *DeepFlow* [176], to recover the lost frame. Such motion field is extrapolated

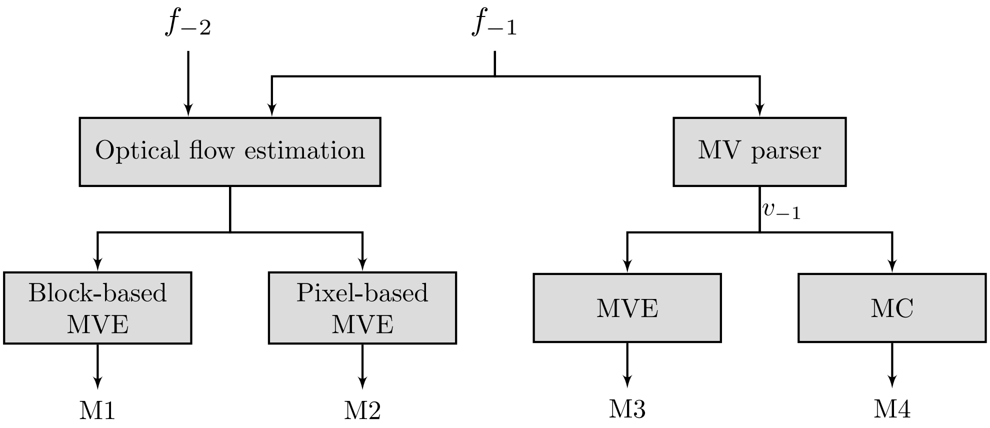


Figure 6.4: EC candidates used for optimization.

at the pixel level, i.e., one MV is assigned to each pixel, and PU level, i.e., one MV is assigned to the entire unit. This results in two different reconstructions of the missing frame, where the former may increase the accuracy of the pixel extrapolation, but the latter has the advantage of keeping the block structure and generate a smoother reconstruction.

The method M3 is based on a conventional MVE, and finally M4 reconstructs the missing frames using the MC algorithm.

6.5.2 Optimal EC mode selection

As previously described in sub-section 6.3, the proposed method allows to choose between four EC candidates, which result in different recovered frames for each lost frame (simulated). The lost frame is partitioned into quadtree structures, each one forming a set of variable size ECUs. Then, the process of finding the optimal EC method for each ECU is formulated as a rate-distortion cost minimisation. To solve such problem, a Lagrangian optimisation is used to achieve the best trade-off between reconstruction quality and signalling overhead. Furthermore, image saliency driven by visual attention modelling is used in the optimisation process to weight the cost of the overhead information of each ECU, according to the visual relevance of its content.

The optimal EC mode $m^*_{l,i}$, for the i_{th} ECU of size defined by the quadtree partitioning level l, is obtained by minimising the following Lagrangian cost:

$$m^*_{l,i} = \underset{\substack{l \in \{0 \ldots 3\} \\ i \in \left\{1 \ldots 4^l\right\}}}{\arg\min} \left\{ D(m_{l,i}) + \left(1 + \frac{t_C - t_I}{T_I}\right) \lambda R(m_{l,i}) \right\}, \tag{6.11}$$

where $R(m_{l,i})$ represents the number of overhead bits required to encode the EC mode, and $D(m_{l,i})$ is the sum of absolute error of the ECU.

In (6.11), besides the parameter λ, which is used to trade-off distortion and overhead rate in the minimisation process, there is also a weighting factor to control the cost of $R(m_{l,i})$, according to the duration of the error propagation. As in the method proposed in the Chapter 5, this method uses the ratio between the temporal distance from the current frame (t_C) to the previous refresh frame (t_I) and the period of refreshing frames (T_I). This behaviour follows the fact that when a lost frame is closer to the next intra-coded frame, the error propagation is lower. Thus, it is not worthwhile to spend too many overhead bits to reduce it. This is accomplished by increasing the cost of such bits according to the temporal distance between the missing frame and the next I frame. Moreover, a dynamic value for λ is also used to adjust the amount of overhead within the frame, according to the visual relevance of each image region, defined by a saliency map (see Section 6.5.3).

The algorithm previously described results in ECUs with variable size, which is defined by the depth of the quad-tree structure, i.e., level l. The partition size is jointly optimised along with the best EC candidate method using (6.11). Alternatively, a fixed ECU size can also be used by defining a fixed depth for all quad-tree structures, i.e., constant l for the entire frame. This can be achieved by using (6.11) to solely optimise the EC methods for each ECU. Although this reduces the overall complexity of the proposed method, it leads to a non-optimal solution which may require additional overhead. In Section 6.6 a comparison between a dynamic and fixed approach is provided.

6.5.3 Saliency detection

The saliency maps are computed based on the well-known Itti's attention model [76] that applies a hierarchical decomposition, based on a set of spatial features, which are combined with a center-surround weighting to obtain the saliency map. Since the Itti's model only produces saliency maps for still images, in this work it is further combined with a temporal feature. This is obtained by estimating the optical flow between two consecutive frames, based on the algorithm presented in [178]. The final saliency map is achieved by combining the spatial and temporal components, S_s and S_t respectively, as follows,

$$S = \rho S_t + (1 - \rho) S_s \tag{6.12}$$

The parameter ρ is used to give higher relevance to the temporal component, since motion is correlated with error propagation, producing longer artefacts and consequently higher impact on visual attention. In this work the temporal component was given a weight slightly higher than 50%, i.e. $\rho = 0.6$.

On the one hand, higher values of saliency (i.e. regions with higher visual relevance) should result in lower values of λ, which tend to increase the overhead rate value on those regions, thus reducing the EC distortion at the decoder, in case of loss. On the other hand, higher values of λ should reduce the overhead, allowing higher EC distortion at the decoder. In order to achieve the aforementioned goal, the following definition for λ is used:

$$\lambda = \begin{cases} \lambda_0 e^{\frac{\bar{S}-s}{\Delta S}}, & \text{if } \Delta S > 0, \\ \lambda_0, & if \Delta S = 0 \text{ (no saliency used)}, \end{cases} \tag{6.13}$$

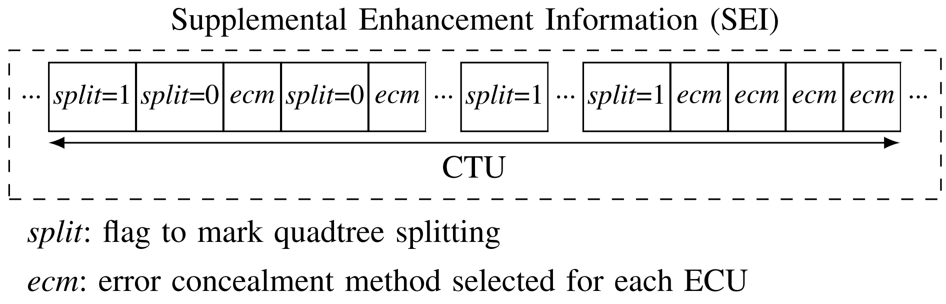

Supplemental Enhancement Information (SEI)
···
split=1
split=0
ecm
split=0
ecm
···
split=1
···
split=1
ecm
ecm
ecm
ecm
···
CTU
split: flag to mark quadtree splitting
ecm: error concealment method selected for each ECU

Figure 6.5: Payload structure of the SEI NALs

where s is the saliency value of the current CTU, $\bar{S}$ is the average saliency value and ΔS is the difference between the maximum and minimum. One should note that this method does not depend on the existence of a saliency map. If the saliency map is constant or non-existent, a fixed value of λ is used (i.e., λ_0) defined at the encoder, which means that no distinction is made across image regions.

6.5.4 Signalling the optimal EC modes

Figure 6.5 illustrates the structure of the signalling information for each CTU comprising a quadtree of ECUs. In this structure, one bit is used for the *split* flag and two bits for the EC mode associated to each ECU ($ecm \in \{M1, M2, M3, M4\}$). The *split* flag indicates whether the ECU is further split into the next level ($split = 1$) or ECU partitioning stops at the current level ($split = 0$) and the corresponding *ecm* modes are inserted. Since no split is necessary at the deepest level, then $split = 1$ followed by the four *ecm* modes for each ECU. This structure allows dynamic selection of different values for the level of the quad-tree partitioning (l) within each frame. Adaptive arithmetic coding [170] is used to efficiently encode both the *split* and *ecm* symbols, further reducing the overhead. Note that the SEI information uses much less bitrate than the video signal, thus it is very unlikely to be hit during transmission.

6.5.5 Enhancement-layer: EC residue coding

As described earlier in this section, in order to allow further improvement of recovered slices, besides the best EC mode, the proposed encoding scheme also allows encoding a controlled amount of EC residue using the scalable extension of HEVC (see Figure 6.3). The enhancement layer module encodes the difference between the coded frame ($\hat{f}$) and its counterpart recovered by using the optimal EC mode ($\tilde{f}$). Note that, the higher the accuracy of the best EC mode, the lower the amount of bits carried by the enhancement layer. At the decoder-side, the reconstruction of a missing frame ($\tilde{f}$) follows three steps:

1. the missing frame is reconstructed using the optimal EC methods for each ECU $\tilde{f}'$ received from the encoder;

2. the enhancement layer residue is decoded (e^c);

3. the recovered frame is obtained as follows:

$$\tilde{f} = \tilde{f}' + e^c. \tag{6.14}$$

The additional residue decoded from the enhancement layer is particularly useful to improve the quality of those CTUs where the EC strategies are not able to produce very accurate reconstruction. Moreover, since the enhancement layer only carries residual information the overhead is affordable.

Note that the generated bitstream is compliant with the standard, as the EC signalling is encapsulated in SEI NALs, while the EC residue is carried in NALs related with the scalable extension. Then, an adapted decoder, capable of decoding both the SEI and the scalable NAL units is also able to take full benefit from the proposed method. However, any standard HEVC decoder is also able to reconstruct the video, because it simply ignores the extra NAL units.

6.6 Error concealment signalling evaluation

This section presents the performance of the proposed EC-aware optimisation scheme using the HEVC reference encoder as the basis of the encoding framework. Two different types of performance evaluation are presented. In the first one, only robust EC of intra-frames is evaluated, while in the second one the proposed EC-aware optimisation is applied to all frames, corresponding to a generic robust video transmission using optimal EC. For comparison, it is used the PSNR obtained from the reference HEVC decoder using motion-copy (MC) and frame-copy (Ref) for EC of the lost frame. Realistic conditions were used for transmission, by allowing independent losses for any type of NAL units, hence there is no guarantee that the decoder conditions are always equal to those at the encoder, which means that sub-optimal EC may be used at the decoder.

6.6.1 Intra-frame EC optimisation

In this section the performance of the proposed method is evaluated only for intra frames, in order to demonstrate its efficiency in reducing error propagation. Since intra CUs result in higher amount of bits, they also require a higher number of packets, making them more prone to errors and data loss. Results indicate that for an intra-period of 16 frames, approximately 42% of the packets correspond to intra coded slices, which is a non-negligible probability of one of them being hit by transmission errors which would lead to packet loss. Moreover, by limiting the proposed method to intra frames, the amount of required overhead is reduced. In this case a constant $\lambda_0 = 10$ was used for all test cases, without taking into consideration saliency estimation. This value results from an experimental study using several sequences and different packet loss ratios with the objective of obtaining the maximum quality gain with low overhead (i.e., average overhead under 1%). This is not a critical value for the performance of the proposed method. It was found that small variations do not lead to significant quality changes.

Since a relevant performance indicator is the error propagation over temporally predicted frames, this was evaluated using a single intra-frame loss. Figures 6.6 and 6.7 show the PSNR results when frame Intra-frame #16 is missing, for LD and RA encoding configurations, respectively. These figures show the results for the proposed method (*Prop*), the motion-copy method (*MC*) and also an alternative proposed method using fixed ECU size of 16 × 16 pixels. These results indicate that the proposed method (*Prop*) outperforms the reference one and it is able to achieve quality gains up to 2 dB for LD configuration (see Figure 6.6), decrease the error propagation and increase the quality of the frames affected from the loss of the intra frame, by up to 3 dB. The use of RA configuration leads to higher quality degradation than LD, due to the higher temporal distance of the reference frame used for EC (Frame #8 is used to recover the lost Frame #16). Note that when Intra-frame #16 is lost, all frames within the range #9 to #31 (i.e., next I-frame) are affected by error propagation. Nevertheless, the proposed method still achieves similar quality gains (approximately 2 dB).

Further tests were performed to evaluate the effectiveness of the proposed method under different loss rates. In these experiments the packets corresponding to intra frames were randomly discarded to simulate loss events, at different PLRs rates with an average burst length of 5 packets. Table 6.2 presents the average quality of the frames affected by errors (the lost frame and dependent ones). This table also shows the PSNR variance value in parentheses and the difference to the PSNR obtained with a

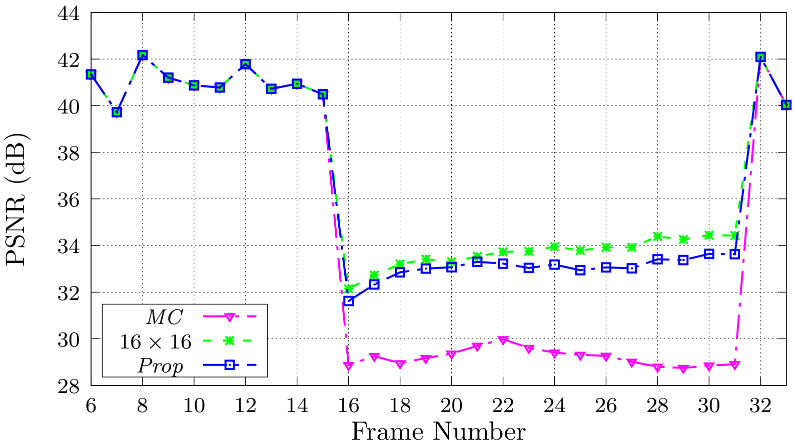


Figure 6.6: Error propagation using LD for the Basketball Drill sequence.

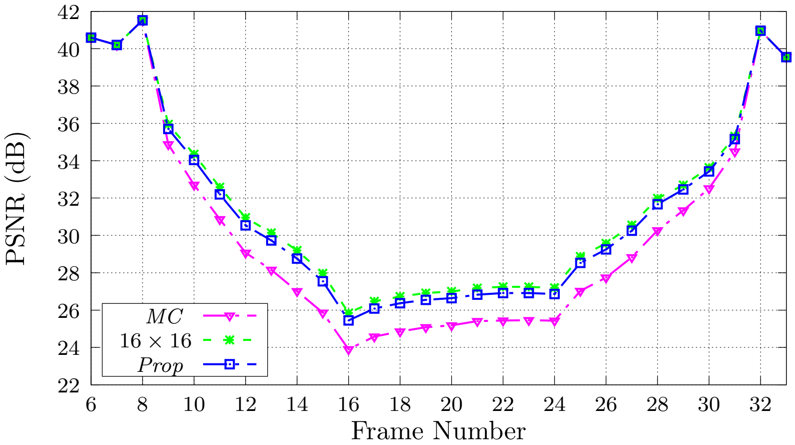


Figure 6.7: Error propagation using RA for the Park Scene sequence.

reference HEVC decoder. The results confirm the superiority of the proposed method, outperforming the reference HEVC for both low and high PLR. Using the LD coding configuration, the proposed method is able to achieve quality gains of 0.88 dB for 3% of PLR and 1.50 dB for 10% of PLR. In case of RA, higher quality gains are achieved, leading to an average improvement of 2.14 dB for 3% of PLR and 3.44 dB for 10% of PLR. It is also worthwhile to notice that the PSNR variance obtained with the *Prop* method is lower, which indicates lower quality variations in the video segments affected by errors.

Subsequently, the percentage of overhead bitrate introduced by the proposed method is compared against fixed ECU sizes of 32×32 and 16×16, i.e., $Level = 1$ and $Level = 2$. These results are presented in Table 6.3 and were obtained for the total bitrate shown in Table 4.1. It is noticeable from the results that reducing the ECU size (see results for 32×32 and 16×16) increases the overhead ratios as more symbols are transmitted (one for each ECU). While the maximum overhead of these fixed partitions is 1.81%

Table 6.2: Average PSNR (dB) of the frames affected by errors and its variance for two different methods under random loss of packets carrying intra-coded frames.

Sequence	**PLR=3%**			**PLR=10%**		
	MC	*Prop*	ΔPSNR	*MC*	*Prop*	ΔPSNR
Low-delay configuration (LD)						
Basketball Drill	33.15 (0.42)	34.26 (0.24)	**+1.11**	32.97 (0.47)	35.90 (0.27)	**+2.93**
Race Horses	30.78 (1.32)	31.55 (0.94)	**+0.77**	30.26 (1.55)	31.12 (1.20)	**+0.86**
Kendo	34.33 (1.28)	36.57 (1.11)	**+2.24**	33.84 (1.26)	37.49 (1.05)	**+3.65**
Park Scene	35.22 (0.12)	35.33 (0.08)	**+0.11**	34.75 (0.11)	34.96 (0.09)	**+0.21**
People On Street	34.36 (0.32)	34.98 (0.22)	**+0.62**	32.17 (0.65)	33.45 (0.50)	**+1.28**
Traffic	36.77 (0.48)	36.88 (0.57)	**+0.11**	36.04 (0.30)	36.31 (0.27)	**+0.27**
Jockey	36.00 (1.14)	37.20 (1.05)	**+1.20**	35.44 (1.21)	36.74 (1.18)	**+1.30**
Average	**34.37** **(0.73)**	**35.25** **(0.60)**	+0.88	**33.64** **(0.79)**	**35.14** **(0.65)**	+1.50
Random-access configuration (RA)						
Basketball Drill	30.22 (8.46)	30.81 (7.66)	**+0.59**	29.19 (7.60)	29.84 (6.88)	**+0.65**
Race Horses	27.35 (8.02)	29.92 (5.89)	**+2.57**	25.75 (7.76)	28.61 (6.14)	**+2.86**
Kendo	27.45 (23.9)	32.38 (18.6)	**+4.93**	25.22 (28.3)	36.98 (22.4)	**+11.8**
Park Scene	33.91 (1.40)	34.46 (1.07)	**+0.55**	31.38 (1.85)	32.50 (1.48)	**+1.12**
People On Street	29.30 (5.12)	31.44 (3.17)	**+2.14**	24.89 (5.37)	27.69 (4.09)	**+2.80**
Traffic	35.64 (1.34)	38.12 (1.24)	**+2.48**	32.85 (1.88)	35.60 (1.40)	**+2.75**
Jockey	32.35 (12.7)	34.06 (8.60)	**+1.71**	29.96 (12.5)	32.11 (8.90)	**+2.15**
Average	**30.89** **(5.64)**	**33.03** **(2.14)**	+2.14	**28.46** **(5.68)**	**31.90** **(4.44)**	+3.44

(Traffic), the proposed dynamic partitioning method is able to reduce the overhead to a maximum of 0.19% (Kendo). Moreover, comparing the results of Table 6.3 with the video quality obtained under packet loss reveals that the proposed method only incurs in a small reduction of the reconstruction quality, with a significantly lower overhead. This indicates that using dynamic partitioning allows to achieve a better trade-off between the amount of overhead and the EC performance. In summary, the proposed technique is able to enhance the performance of the standard decoding of intra coded frames by finding the best EC method to be used, for each ECU, at the encoder using

Table 6.3: Percentage of overhead (%) when using the proposed method solely for intra-coded frames.

Sequence	32×32 **Level: 1**	16×16 **Level: 2**	*Prop* **Level: dynamic**
Basketball Drill	0.08	0.27	0.08
Race Horses	0.16	0.65	0.08
Kendo	0.27	1.05	0.19
Park Scene	0.42	1.72	0.05
People On Street	0.06	0.25	0.11
Traffic	0.50	1.81	0.06

with a small amount of signalling overhead.

6.6.2 Robust video transmission with optimal EC

In this section the proposed EC-aware optimisation scheme is evaluated for a generalised robust video transmission of both intra and inter coded frames. The experimental results presented in this section were obtained by applying the proposed technique to all video frames. As in the previous subsection, a $\lambda_0 = 10$ was used. This section is organised in the following subsections:

- evaluation of the usage distribution of each EC mode;
- performance evaluation of the EC optimal mode decision (without using the enhancement layer);
- evaluation of the error propagation under single loss events;

EC mode usage distribution

The EC mode usage distribution is an indicator of the usefulness of each EC mode. Table 6.4 shows the average *Level* selected for the ECU quadtree and the usage percentage of each EC mode. It is noticeable from these results that the proposed method leads to low average *Level* values (lower than 1), resulting in larger ECUs, which requires a lower amount of ECU to be transmitted than if higher values of *Level* were chosen. This implicitly results in low signalling information. This also justifies the low overhead shown in Table 6.7.

The results in Table 6.4 also show that all EC modes are effectively used, which indicates that all of them are found relevant to recover the missing data. The EC candidate methods based on motion field estimation (i.e., M1, M2) are selected more often because they provide more accurate representation of the scene's motion than the others (i.e., M3, M4).

Table 6.4: Average Level and usage ratio of each EC mode.

Sequence	**Avg. Level**	**Avg. usage ratio (%)**			
		M1	M2	M3	M4
Basketball Drill	0.81	59.61	16.62	14.40	9.37
Book Arrival	0.34	66.77	18.79	7.15	7.29
BQSquare	0.41	15.23	0.64	72.97	11.16
Four People	0.17	85.84	3.22	8.09	2.86
Kimono	0.88	33.76	23.87	26.03	16.35
Park Scene	0.69	39.34	13.58	34.19	12.89
People on Street	1.51	31.17	34.39	16.15	18.29
Race Horses	1.62	34.05	29.89	17.06	19.01
Tennis	1.17	25.27	30.03	23.22	21.48

Evaluation of optimal EC mode decision

The evaluation of the proposed saliency-weighted EC mode decision (*Saliency*) was carried out under random loss events and it was compared with the reference methods. The video quality was evaluated by using two quality metrics: PSNR and a Feature Similarity Index (FSIM) [183], in order to validate the performance of the proposed method in a consolidated manner. The results obtained without saliency weighting are also shown for comparison (*NoSaliency*).

Tables 6.5 and 6.6 show the average quality under different PLRs for the LD configuration. Table 6.5 shows the PSNR and Weighted Peak Signal-to-Noise Ratio (WPSNR) between parentheses. The MSE weights used in the WPSNR are defined by the saliency values [184]. Table 6.6 shows the results obtained for the FSIM quality metric. Note that, the absolute value is shown for the reference case (*Ref*), while the differential values are shown for the remaining methods. From the PSNR results shown in Table 6.5, one may conclude that the proposed method (*Saliency*) consistently outperforms the reference ones for both low and high PLRs. Average PSNR gains up to 2.95 dB for PLR=3% and 2.58 dB for PLR=10% are achieved. It is also worthwhile to notice that higher performance is obtained for sequences with high motion, such as Basketball Drill, which results in PSNR gains up 4.82 dB. Moreover, the results of Table 6.6 confirm the quality improvement achieved by the proposed method, leading to an increase of FSIM up to 4.32, which indicates that the proposed method is able to recover important image features with higher accuracy than the others.

Table 6.7 shows the percentage of overhead in the total rate with and without saliency-weighted optimisation, i.e., *NoSaliency* and *Saliency*, respectively. As shown in this table, the saliency-weighted optimisation reduces the amount of overhead by dynamically assigning unequal number of bits to image regions of different visual importance. These results show that using an adaptive value of λ, based on the saliency

Table 6.5: Average quality (PSNR and WPSNR) of the affected frames for Low-Delay configuration.

Sequence	PLR=3%					PLR=10%				
	Ref	*MC*	*Deep Flow*	*NoSaliency (Constant λ)*	*Saliency (Adaptive λ)*	*Ref*	*MC*	*Deep Flow*	*NoSaliency (Constant λ)*	*Saliency (Adaptive λ)*
Basketball Drill	28.41 (29.25)	+2.93 (+3.22)	+3.25 (+3.24)	+5.28 (+5.02)	+2.64 (+5.01)	24.64 (22.89)	+2.63 (+3.27)	+3.09 (+3.25)	+4.93 (+5.66)	+4.82 (+5.66)
Book Arrival	30.69 (33.53)	+2.31 (+1.56)	+3.11 (+1.48)	+3.80 (+2.61)	+3.73 (+3.17)	28.29 (27.28)	+2.65 (+2.82)	+1.47 (+2.74)	+4.59 (+4.91)	+4.51 (+5.86)
BQSquare	31.82 (35.39)	+1.97 (+1.59)	+1.23 (+0.90)	+2.21 (+1.75)	+2.19 (+2.29)	28.06 (30.91)	+2.59 (+2.52)	+1.92 (+1.43)	+3.14 (+2.97)	+3.11 (+3.82)
Four People	34.54 (37.03)	+1.46 (+0.75)	+1.87 (+0.73)	+1.91 (+0.96)	+1.89 (+1.14)	33.07 (34.04)	+1.85 (+1.86)	+2.70 (+1.88)	+2.59 (+2.52)	+2.56 (+2.83)
Kimono	32.39 (33.86)	+1.66 (+1.76)	+1.88 (+1.87)	+2.32 (+2.43)	+2.30 (+2.65)	27.98 (29.54)	+0.14 (+0.14)	+0.29 (+0.13)	+0.19 (+0.19)	+0.19 (+0.50)
Park Scene	31.66 (31.76)	+2.28 (+2.89)	+2.15 (+2.39)	+2.96 (+4.03)	+2.93 (+4.27)	27.76 (26.82)	+1.32 (+1.53)	+1.25 (+1.31)	+1.93 (+2.42)	+1.90 (+2.81)
People on Street	26.72 (27.24)	+1.59 (+1.78)	+2.15 (+2.36)	+2.96 (+3.21)	+2.88 (+3.35)	22.53 (22.57)	+0.40 (+0.45)	+1.25 (+0.61)	+0.85 (+0.91)	+0.83 (+1.06)
Race Horses	25.97 (27.81)	+2.55 (+2.73)	+2.59 (+2.57)	+3.80 (+3.98)	+3.69 (+4.29)	21.55 (22.63)	+0.68 (+0.76)	+0.76 (+0.72)	+1.12 (+1.23)	+1.07 (+1.50)
Tennis	28.47 (30.79)	+1.66 (+1.27)	+2.02 (+1.53)	+4.39 (+3.52)	+4.27 (+3.62)	23.77 (24.71)	+1.41 (+1.13)	+1.82 (+1.45)	+4.34 (+3.79)	+4.19 (+3.84)
Average	–	**+2.04** **(+1.95)**	**+2.25** **(+1.90)**	**+3.29** **(+3.06)**	**+2.95** **(+3.31)**	–	**+1.52** **(+1.61)**	**+1.62** **(+1.50)**	**+2.63** **(+2.73)**	**+2.58** **(+3.10)**

Table 6.6: Average FSIM under frame loss conditions (scaled between $[0 - 100]$).

Sequence	**PLR=3%**					**PLR=10%**				
	Ref	*MC*	*Deep Flow*	*NoSaliency (Constant λ)*	*Saliency (Adaptive λ)*	*Ref*	*MC*	*Deep Flow*	*NoSaliency (Constant λ)*	*Saliency (Adaptive λ)*
Basketball Drill	96.96	+1.13	+1.16	+1.58	+1.59	92.74	+2.38	+2.50	+3.60	+3.65
Book Arrival	98.61	+0.36	+0.41	+0.57	+0.60	96.66	+1.05	+1.21	+1.74	+1.75
BQSquare	98.03	+0.39	+0.28	+0.41	+0.41	96.31	+1.15	+0.81	+1.30	+1.31
Four People	99.41	+0.09	+0.09	+0.11	+0.11	98.97	+0.31	+0.32	+0.39	+0.60
Kimono	98.54	+0.48	+0.52	+0.59	+0.62	95.79	+0.09	+0.12	+0.12	+0.15
Park Scene	98.74	+0.58	+0.55	+0.69	+0.71	96.41	+0.83	+0.79	+1.03	+1.10
People on Street	98.19	+0.67	+0.73	+0.92	+0.94	94.65	+0.45	+0.51	+0.66	+0.70
Race Horses	95.16	+1.78	+1.80	+2.28	+2.29	88.25	+1.18	+1.21	+1.62	+1.68
Tennis	96.64	+0.98	+0.98	+1.69	+1.64	91.02	+2.20	+2.29	+4.27	+4.32
Average	–	**+0.72**	**+0.73**	**+0.98**	**+0.99**	–	**+1.07**	**+1.09**	**+1.64**	**+1.70**

Table 6.7: Error concealment signalling overhead.

Sequence	**Bits used for signalling (%)**	
	NoSaliency *(Constant λ)*	*Saliency* *(Adaptive λ)*
Basketball Drill	2.88	2.55
Book Arrival	4.03	3.35
BQSquare	0.33	0.24
Four People	3.06	3.18
Kimono	1.75	1.30
Park Scene	1.84	1.37
People on Street	8.39	5.67
Race Horses	1.84	1.21
Tennis	12.15	7.46
Average	**+4.04**	**+2.91 (-28%)**

values, leads to an average overhead reduction of 28% in comparison with the case where no saliency is used, i.e., constant λ. It is worthwhile to note that using the saliency-weighted optimisation to reduce the overhead does not compromise the video quality, as can be seen in the results shown in Tables 6.5 and 6.6. The PSNR values for *NoSaliency* and *Saliency* are very close, but in terms of overhead savings the results are significant for most sequences. The achieved WPSNR and FSIM results further demonstrate the effectiveness of the saliency-based optimisation, because the lower overhead does not affect the most relevant image regions, i.e., those with higher saliency values. Thus, the proposed method, using a adaptive λ (i.e., using saliency), is able to not only achieve higher EC quality than the reference methods, but also higher quality gains than using a constant λ (i.e., without saliency). This confirms the two-fold superiority of adaptive λ, i.e. not only the total overhead is reduced but also higher quality gains are achieved.

Error propagation evaluation

In this section the error propagation is evaluated using single loss events and compared against the reference methods. In this simulation a whole frame loss is enforced in order to assess the impact of error propagation on the worst case scenario. Experimental evaluations were performed for both intra and inter-coded frame loss. Figure 6.8 shows the PSNR over a GOP with an IDR period of 16 frames, where the whole frame #6 is missing in sequences Basketball Drill and People On Street. Figure 6.9 shows the PSNR for the same conditions but with a loss event on the intra-coded Frame #16, in sequences Kimono and Tennis.

These results reveal that the proposed method is able to improve the reconstruction of inter-coded frames and also outperforms the existing techniques, whenever a

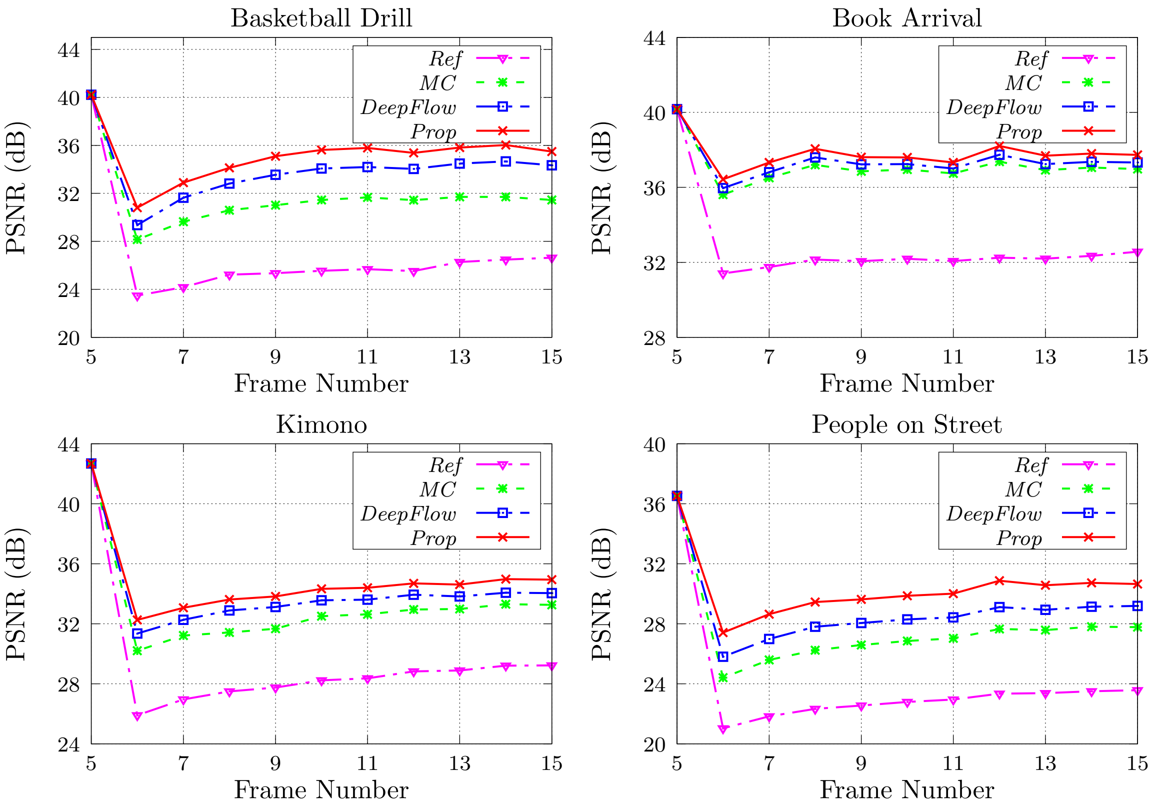


Figure 6.8: Error propagation after losing the inter-coded frame #6 for the proposed EC-aware approach and three different EC reference methods.

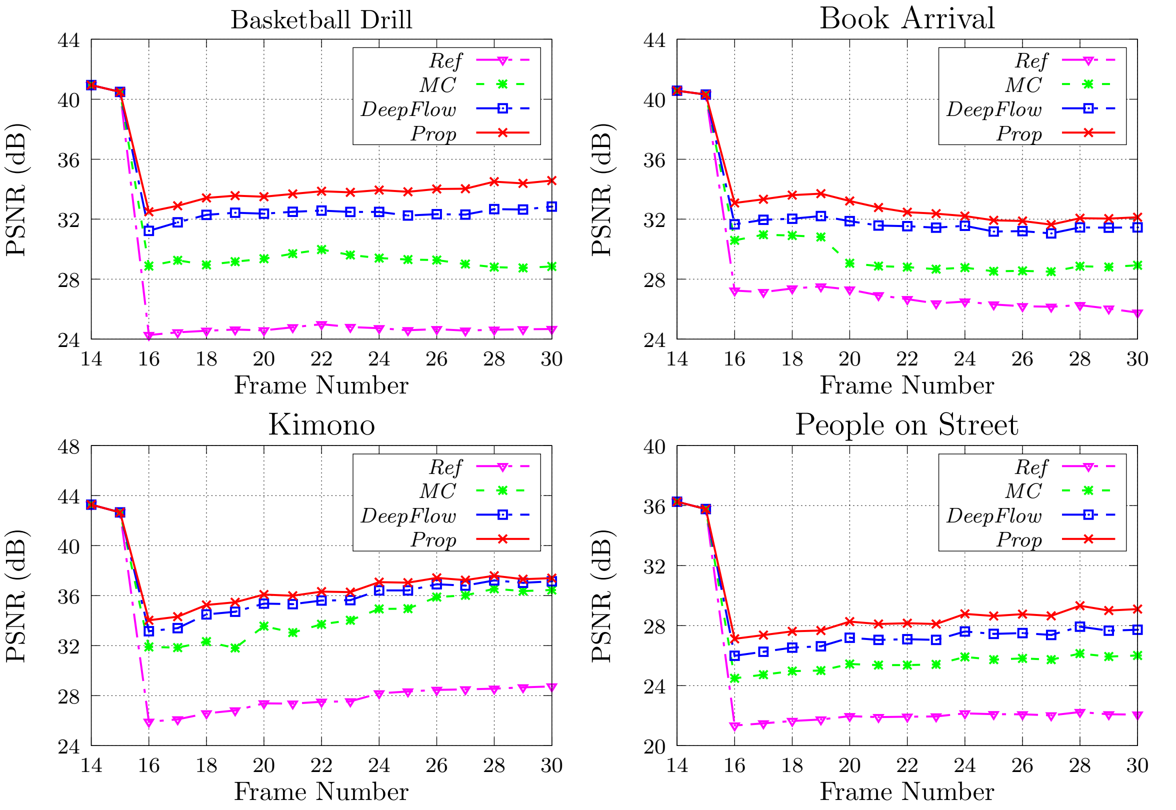
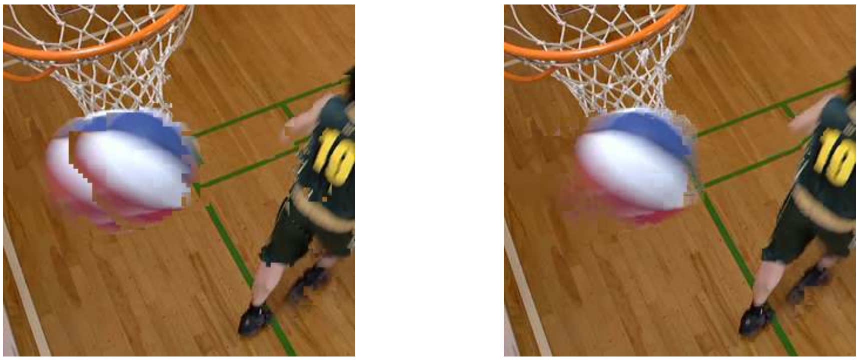


Figure 6.9: Error propagation after losing intra-coded frame #16 for the proposed EC-aware approach and three different EC reference methods.



MC (28.15 dB) *Prop* (30.99 dB)

Figure 6.10: Recovered frame #6 of Basketball Drill

intra-coded frame is missing. These results also indicate that the proposed method outperforms both reference methods right after the time instant when the loss occurs. The lost frame itself is recovered with much higher quality than using the reference EC methods. This is due to the use of the optimal EC mode to recover each ECU, benefiting from the SEI information received from the encoder. In this case, the proposed method is able to achieve quality gains up to 5 dB, decreasing the error propagation and increasing the PSNR of affected frames by up to 6 dB. Comparing the results of Figures 6.8 and 6.9 one may conclude that the loss of intra-frame leads to higher error propagation. Furthermore, as the IDR period increases, the error propagation also lasts longer with greater impact on video quality.

The visual impact of using the proposed EC-aware encoding can be evaluated by observing the recovered frame (i.e., frame #6) of Basketball Drill, as an example. Figure 6.10 shows that higher visual quality and less reconstruction artefacts are obtained in comparison with the *MC* method. It is worthwhile to note that the proposed method is able to reconstruct the object edges with higher accuracy than the reference *MC*.

6.7 Overall performance evaluation

As in the previous simulations, the packets were randomly discarded to simulate loss events, at different PLRs. The quality obtained with the proposed method is compared with the default method used by reference HEVC , i.e., frame-copy (*Ref*), motion-copy (*MC*) [152], and the method proposed in [85], which implements a robust scheme to minimise error propagation. In this experiment both layers, i.e., video stream and enhancement layer with the EC residue, may be affected by errors.

Table 6.8: Average PSNR (dB) under 3% of PLR for different methods and overhead ratios.

Sequence	*Ref*	*MC*	*Ref [85]*	*Proposed*	
				10%	20%
Low-delay configuration (LD)					
Basketball Drill	33.31	+2.10	+3.86	+4.62	+4.56
Book Arrival	36.21	+1.33	+2.07	+2.71	+2.99
BQSquare	35.64	+1.47	+1.18	+2.01	+1.77
Four People	39.16	+0.47	-0.65	+0.40	-0.06
Kimono	36.19	+1.41	-0.54	+3.09	+3.47
Park Scene	34.54	+2.06	+2.01	+3.22	+3.15
People On Street	29.48	+1.50	+1.35	+2.64	+3.08
Race Horses	29.08	+2.34	+4.52	+5.35	+6.02
Tennis	32.71	+1.32	+3.13	+3.81	+4.97
Average	–	**+1.55**	**+1.88**	**+3.10**	**+3.33**
Random-access configuration (RA)					
Basketball Drill	35.40	+0.17	+0.20	+1.39	+1.53
Book Arrival	37.26	+0.05	+0.72	+1.31	+1.28
BQSquare	34.32	-0.06	+0.51	+1.46	+1.59
Four People	39.92	+0.02	-0.39	+0.29	+0.02
Kimono	36.98	+0.01	-2.44	+0.34	+0.00
Park Scene	35.18	+0.06	-1.27	+1.11	+1.13
People on Street	29.21	+0.03	-0.53	+0.58	+0.95
Race Horses	30.89	+0.09	-0.34	+1.38	+1.42
Tennis	33.86	+0.16	+0.18	+0.98	+1.48
Average	–	**+0.06**	**-0.37**	**+0.98**	**+1.05**

6.7.1 Quality evaluation

Tables 6.8 and 6.9 present the average PSNR of decoded video for different PLRs and different percentages of total overhead (i.e., signalling rate plus enhancement layer rate). The absolute PSNR value is shown for the reference case, while the PSNR difference is presented for the remaining methods. These results show that transmitting the EC residue in the enhancement layer further improves the reconstruction quality obtained when only the optimal EC method is used (see previous section). By only using 10% of the total bit rate, the proposed EC-aware coding scheme is able to outperform the reference HEVC, the *MC* and the method proposed in [85] for both LD and RA configurations. Using 20% of the total bitrate the quality gains increase up to 6.90 dB for the Basketball Drill sequence when comparing with the reference HEVC, achieving an average gain of 3.79 dB for the LD configuration (1.33 dB for the RA).

When comparing the results of Tables 6.8 and 6.9 with the spatial and temporal information of each sequence (SI/TI in Table 4.1) it is noticeable that higher quality gains are achieved for sequences with higher motion information, i.e., Basketball Drill

Table 6.9: Average PSNR (dB) under 10% of PLR for different methods and overhead ratios.

Sequence	*Ref*	*MC*	*Ref [85]*	*Proposed* 10%	*Proposed* 20%
Low-delay configuration (LD)					
Basketball Drill	28.22	+2.34	+5.32	+6.26	+6.90
Book Arrival	31.01	+2.48	+4.49	+5.60	+6.31
BQSquare	31.36	+2.50	+2.40	+3.33	+3.55
Four People	37.12	+1.24	-0.99	+1.82	+1.51
Kimono	32.24	+0.12	-1.92	-0.04	-0.21
Park Scene	30.42	+1.22	+1.56	+2.92	+3.64
People On Street	25.05	+0.40	+0.74	+0.97	+2.09
Race Horses	24.19	+0.63	+2.53	+2.75	+3.74
Tennis	27.08	+1.30	+3.51	+4.58	+6.61
Average	–	**+1.36**	**+1.96**	**+3.13**	**+3.79**
Random-access configuration (RA)					
Basketball Drill	29.92	+0.14	+1.45	+1.84	+2.42
Book Arrival	32.27	+0.13	+2.58	+2.74	+2.91
BQSquare	28.13	-0.01	+2.74	+2.06	+2.85
Four People	37.81	+0.09	-0.05	+1.46	+1.68
Kimono	33.09	+0.00	-0.03	-1.15	-2.02
Park Scene	30.93	+0.01	+0.73	+0.83	+1.02
People on Street	24.57	+0.00	+1.70	+0.09	+0.35
Race Horses	25.71	+0.01	+1.30	+0.58	+0.71
Tennis	27.74	+0.07	+1.35	+1.17	+2.09
Average	–	**+0.05**	**+1.31**	**+1.07**	**+1.33**

and Tennis sequences. Moreover, for sequences with higher spatial information (e.g., BQSquare and People on Street) considerable quality gains are also achieved. This indicates the superior performance of the proposed method when dealing with complex sequences, both in temporal and spatial levels.

6.7.2 Visual quality evaluation

To provide an informal subjective quality comparison between the proposed method (using 20% of overhead) and the reference HEVC using *MC* as error concealment, Figure 6.11 shows a region of the decoded frames #8 and #12 when Frame #6 is lost. The visible artefacts demonstrate how the decoder is able to recover from whole frame loss in both methods, and also the impact of error propagation. Figure 6.11 shows that the proposed method clearly improves the reconstructed quality in comparison with *MC*.

A closer look at the Race Horses sequence reveals that the horse rider is not correctly reconstructed by the *MC* method (highlighted by the circles and ellipses). As shown

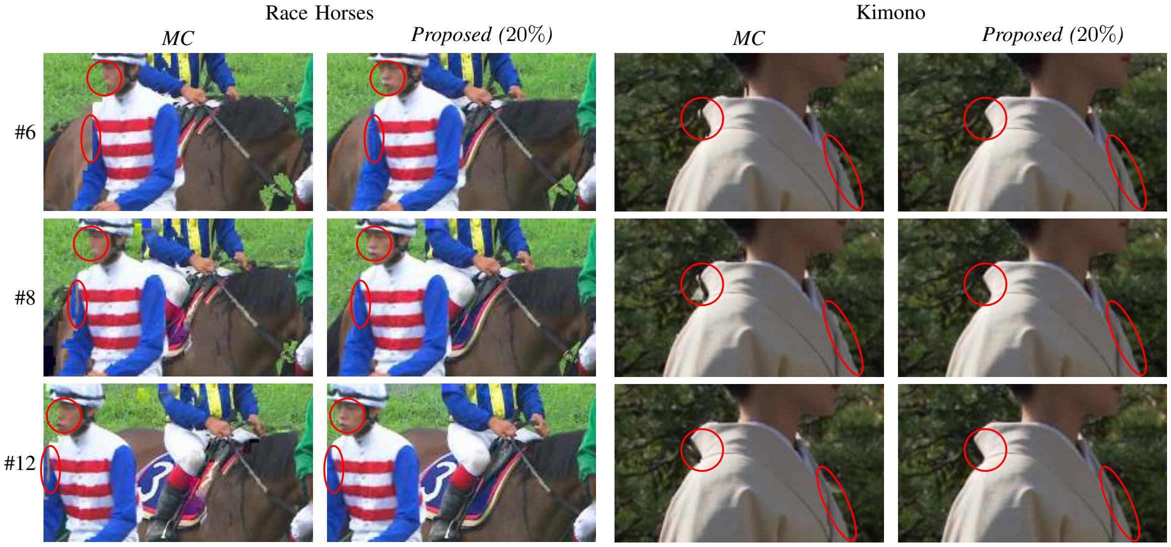


Figure 6.11: Crop of the decoded video frame for the Race Horses and Kimono sequences (Loss event at Frame #6)

in the figure, the proposed method not only results in a more accurate reconstruction of the lost frame, but also increases the visual quality of the subsequent frames. In the Kimono sequence, some artefacts are also noticeable in the *MC* case, but not visible when the proposed method is used (highlighted with circles). Moreover, the edges of the women's kimono (highlighted with ellipses) are also improved by the proposed method, resulting in a smooth reconstruction without error propagation. From the results discussed above, it is clearly observed that the proposed method is able to reconstruct the contours of the moving parts more accurately than traditional EC algorithms.

6.7.3 Comparison with previous works

The achieved results were compared with previous works, where fair comparisons could be made. In [52] a method was proposed to improve the recovery quality by signalling the suitable MV to be used for EC. The proposed EC-aware scheme of Figure 6.3 is able to achieve similar quality gains with less amounts of redundancy, allowing a better trade-off between quality and overhead. In comparison with the work described in [156], the proposed method is able to select among more complex EC methods, which allows for higher quality gains. Both of the above mentioned methods search for the optimal EC method at the encoder-side, which results in similar complexity as the proposed method.

The methods in [48, 79] both use a feedback channel and intra-refresh to reduce the error propagation. The results presented in [79] revealed an average quality gain of approximately 1.68 dB. Since the proposed EC-aware scheme is able to achieve higher quality without requiring any feedback channel, it presents better performance and functional advantages. An advantage of using a feedback channel is the low extra complexity required to implement robust coding because the encoder receives specific information about transmission errors. This allows straightforward spatio-temporal localization of the lost data without computational effort.

Table 6.10 shows the average quality differences (Δ **PSNR**) between the reference method (*Ref*) and the proposed one plus three other previous works. Two different test conditions are shown in the table: no packet loss (No-Loss) and PLR=10%. Moreover, an efficiency measure R is defined, as the ratio between the quality differences with and without packet loss, i.e.,

$$R = \left| \frac{\Delta PSNR(\text{PLR} = 10\%)}{\Delta PSNR(No\text{-}Loss)} \right| . \tag{6.15}$$

The results shown in Table 6.10, demonstrate that the effectiveness of the EC-aware coding scheme is higher than all other methods under comparison. In fact, not only the absolute PSNR difference (i.e., PSNR gain) is higher for the same PLR, but also the efficiency of the overhead is also higher. This means that the overhead bits spent to encode the signalling information and the enhancement residue carry more useful information than the other methods. For instance, in comparison with the method in [87] the proposed method leads to higher quality degradation in those cases without errors. However, for PLR=10% it clearly outperforms the method in [87], achieving an higher R, indicating a superior efficiency. Results in Table 6.10 also show that the small loss of coding efficiency in packet loss free transmission is clearly surpassed by the quality gains under error-prone transmission conditions. The methods in [85, 87] are based on modified predictions, while the method in [121] uses redundant data without any further optimisation. Both of these approaches require low additional computational complexity. However, as observed in the results shown in Table 6.10, the lower additional complexity also corresponds to lower quality gains in case of packet loss.

The proposed method uses EC techniques with different levels of computational complexity for each ECU. Since an adaptive approach is used, the average complexity tends to be lower in comparison with other methods that rely on a fixed approach. Thus, less complex video decoding is obtained when using the proposed method in comparison with other straightforward methods that always use the same EC technique

Table 6.10: Comparison of the proposed EC-aware robust coding architecture against existing techniques in environments with and without packet loss.

Method	**Δ PSNR (dB)**		**R**
	No-Loss	PLR=10%	
Ref [85]	-0.88	1.64	1.9
Ref [121]	-0.37	2.03	5.5
Ref [87]	-0.24	1.42	5.9
Prop	-0.35	2.10	6.0

for all CTUs.

6.8 Summary

In this chapter a novel EC-aware encoding scheme was proposed by incorporating loss simulation at the encoder-side to enhance the EC efficiency at the decoder. The novelty of this method is to include optimal EC mode decision at the encoder, to be used by the decoder in case of packet loss, and also to allow encoding of the EC residue using the enhancement layer of the HEVC standard. The use of EC signalling was firstly proposed for the use on intra-coded frames in (C5), and then improved and proposed for a generic robust video transmission, resulting in a more detailed analysis published in (J2). The overall results demonstrate that the proposed method is able to efficiently reconstruct lost frames and reduce the error propagation compared with other reference methods. A consistent quality improvement achieved at the cost of low overhead allows the use of this method in broadcasting video services and network applications where packet loss probability is not negligible.

Chapter 7

Adaptive error robustness for highly efficient video coding

7.1 Introduction

The effectiveness of different error resilience techniques is rather variable depending on the video content, the transmitted bitrate, network conditions, among others. Therefore, it is not straightforward to select a single technique capable of guaranteeing the highest efficiency for every possible scenario and also for the whole range of networking conditions. This chapter addresses this problem by proposing an adaptive error robustness scheme that optimises the efficiency of the techniques previously described in Chapters 5 and 6. At the encoder-side, a pre-trained neural network predicts the video stream that is able to achieve the highest decoded quality for particular loss cases, among various video streams encoded with different levels of compression efficiency and error robustness. The stream selection is based on different input parameters that include information from the video signal, the coded stream and the transmission network. Using this approach the overall performance of the proposed adaptive method significantly increases, achieving a better trade-off between compression efficiency and error robustness.

This chapter is organised as follows. Section 7.2 presents the problem addressed in this chapter and experimental evidence, as motivation for the method investigated. Section 7.3 provides an overview of architecture that supports the implementation of the proposed method and Section 7.4 describes the technical details. Section 7.5 presents the performance evaluation and a discussion of results. Finally, Section 7.6 concludes the chapter.

7.2 The problem of adaptive error robustness

In the scope of this thesis different robust coding techniques were investigated and described in previous chapters to increase the overall performance of video transmission under error-prone conditions. Different algorithms were devised in the proposed methods, such as, reference frame selection, motion vector redundancies and error concealment signalling. In the research work described in Chapters 3 to 6, these methods were evaluated against existing techniques, revealing superior performance and better efficiency to deal with error robustness problems in high efficient video coding.

Overall, the performance results of these methods have demonstrated that they can efficiently increase the robustness of video streams against errors and data loss, leading to higher decoded video quality. However, each method on its own does not adapt to different conditions. For instance, while the reference frame selection algorithm presented in Chapter 5 introduces low overhead in comparison with existing techniques, the method proposed in Chapter 6 achieves higher quality gains at cost of slightly higher bitrate increase. Therefore, one may conclude that each method may be further optimised for different applications, video characteristics and network conditions. This is the motivation for the research presented in this chapter, which advances one step forward by investigating a joint robust coding method using different techniques to optimise the video coding process across different conditions, such as, the input content and the network information. This scheme can be used in transmission systems where different video bitstreams can be made available at the encoder-side to be selected based on the dynamic conditions.

7.2.1 Comparison of the proposed methods

In order to address the above problem of adaptive robust coding, a simulation study was carried out to compare the error robustness performance of the previously proposed methods for different conditions. This study extends the previous evaluations presented in Chapters 5 and 6, as they target a wider range of test material and network conditions. Moreover, it aims at comparing the performance of different techniques across a variety of video content, bitrates and loss conditions, in order to find relevant insights on how the decoded video quality varies with each parameter. Consequently, these relevant insights may be used to develop an adaptive robust coding method.

The following methods were evaluated in the simulation study, namely:

- the default HEVC decoder with frame-copy EC (*Ref*), as reference method;

Table 7.1: Test material used in the experimental simulation and its characteristics.

Sequence	**SI**	**TI**	**Inter (%)**	**MV Magnitude**	**Main bitrate (Mbps)**
Basketball Drill	33.4	14.4	88.1	6.40	4.50
Basketball Drive	33.0	15.2	85.0	27.4	15.00
Book Arrival	28.4	21.7	90.7	4.67	1.50
BQSquare	63.2	11.5	93.7	1.51	4.00
Cactus	30.5	11.4	89.7	5.90	20.00
Four People	31.3	6.90	93.3	0.76	2.00
Kendo	19.6	16.1	88.0	9.85	0.60
Kimono	23.4	32.5	83.0	14.8	8.00
Kristen and Sara	25.6	6.30	93.1	1.75	15.00
Park Scene	31.3	11.6	92.0	7.26	10.00
Party Scene	52.6	11.4	89.8	3.42	4.50
People on Street	40.0	25.4	87.4	7.80	12.00
Race Horses	43.7	24.4	79.7	22.5	8.00
Tennis	20.3	45.3	76.2	91.0	5.00

- the default HEVC encoder without robust coding-techniques and motion vector extrapolation EC (*M1*);
- the proposed reference frame selection scheme (*M2*);
- the method *M2* in combination with an EC optimisation with 10% of bitrate allocated for the scalable enhancement layer (*M3*);
- the same as above, with 20% of bitrate allocated for the scalable enhancement layer (*M4*).

These methods were selected because they are the main techniques proposed within this research work and cover two elements of a robust video communication chain: (i) robust video encoding by dynamically select the reference frames and (ii) robust video decoding by improving the EC performance.

Table 7.1 presents the list of test sequences used. These sequences were selected to cover a wide range of video content, with different levels of spatial and temporal complexity, as demonstrated by the values of SI and TI [159], respectively. Moreover, from the bitstream encoded with the reference implementation of the HEVC encoder, the percentage of the inter-coded blocks and the average MV magnitude was extracted and presented in the table (columns 4 and 5, respectively). These parameters were considered due to their possible relation with the performance obtained by the proposed methods. The remaining configuration was kept as described in Section 4.2.

Preliminary simulation results and discussion

Figures 7.1 and 7.2 show the average decoded video quality (PSNR) obtained for different PLRs. Figure 7.1 shows the results obtained with the test sequence encoded using a constant target bitrate listed in Table 7.1, referred to as the main bitrate. Results on Figure 7.2 were obtained by reducing the main bitrate by 40%. This is used to evaluate the impact of the network bandwidth on the decoded video quality. The results shown in the figures correspond to the same test sequences, in order to allow an accurate and fair comparison.

Firstly, comparing the results for different bitrates, it is noticeable that lower quality is obtained for the lower bitrates, which obviously results from the higher compression ratios used. It is also noticed that lower quality gains, when comparing with the reference HEVC encoder (*Ref*), are obtained for the video streams encoded with a 40% reduction of bitrate. This indicates that for lower bitrates the performance of the proposed methods decreases, especially for the *M4* case, where 20% of the bitrate is allocated for the enhancement layer. Therefore, for lower bitrates, using methods that have a lower impact on the coding efficiency (e.g., *M2*) or even disable the robust coding techniques (*M1*) may lead to higher performance.

Secondly, comparing the results across different video content, i.e., different test sequences, one can notice that the performance of the different methods is not consistent. For instance, in the results of Figure 7.1, for Tennis sequence the method *M4* clearly outperforms the remaining ones. However for the BQSquare sequence, higher performance is achieved by using less bitrate (i.e., 10%) allocated for the EL (*M3*). This is a consequence of the higher quality degradation due to higher motion in Tennis sequence (see results for TI and MV size in Table 7.1) which requires higher redundant information to increase the overall quality.

Finally, by comparing the results for all test sequences and both bitrates it can be clearly noticed that the PLR has a significant impact on the performance of the error robustness techniques. In case of the proposed reference frame selection method (*Exp*), quality gains in comparison with the *Ref* case are achieved for lower loss ratio, i.e., PLRs as low as 2% for all test conditions. For method *M4*, higher PLRs are required to be able to outperform both the reference HEVC and the MVE method.

Summarising, the performance of the proposed methods does not exhibit a consistent behaviour across different video content and bitrates, opening a new research goal to adaptively choose the best method that should be applied to achieve the highest decoded video quality. In order to achieve increased efficiency and higher video quality,

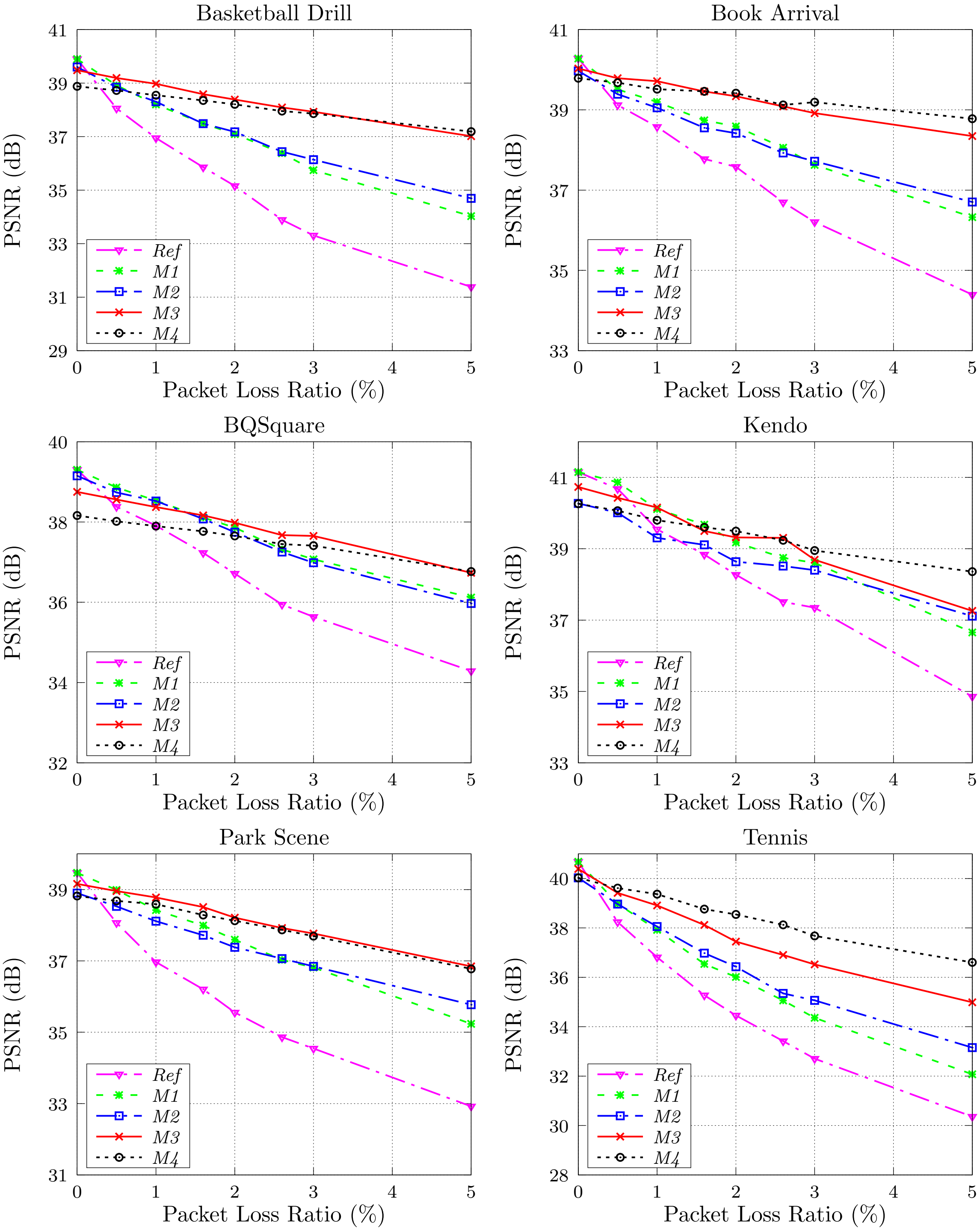


Figure 7.1: Average decoded video quality for the bitstreams encoded with the main bitrate (bps).

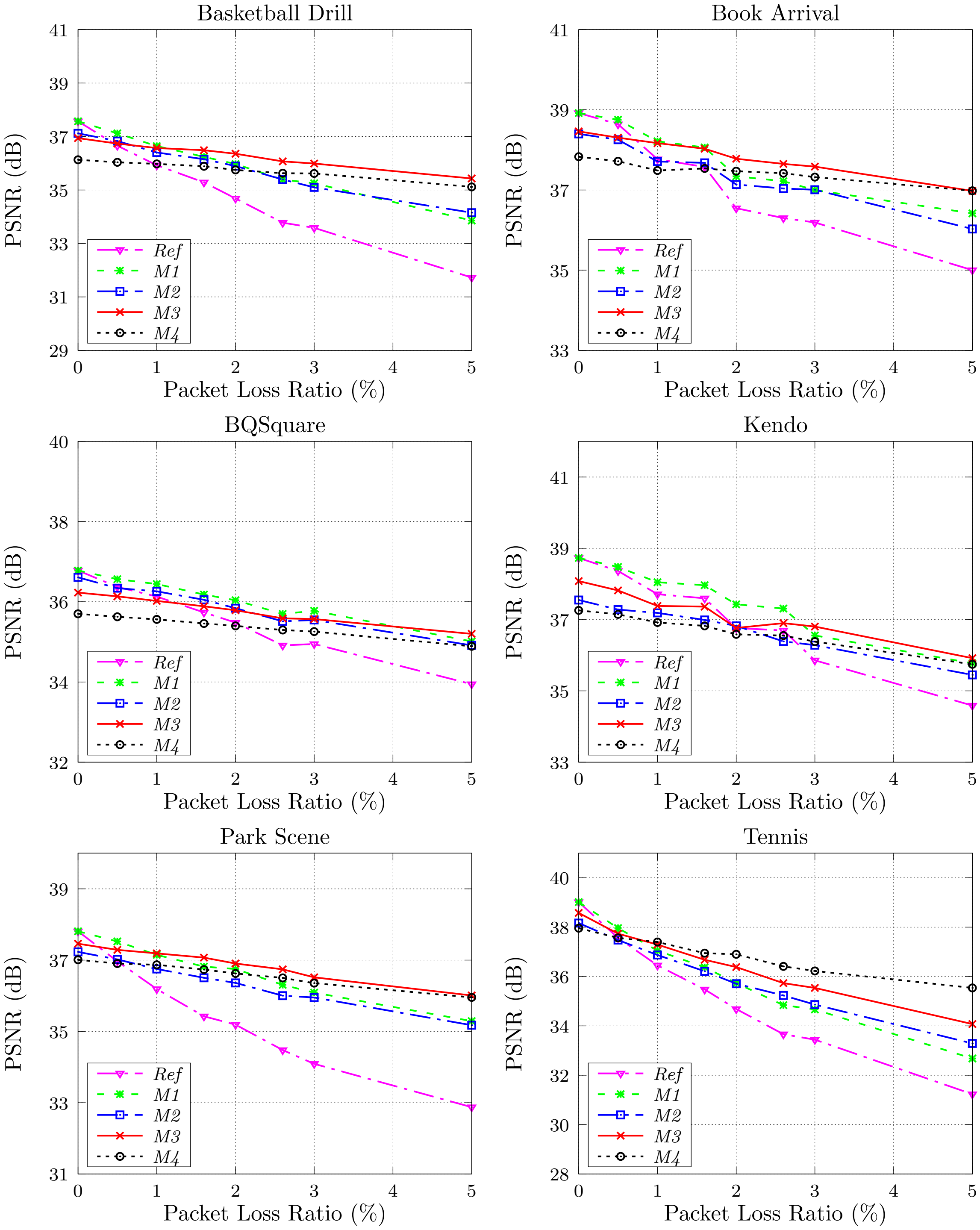


Figure 7.2: Average decoded video quality using a 40% reduction of the main bitrate (bps).

the best trade-off between auxiliary information and video bitrate must be found by selecting the optimal robust coding algorithm.

7.3 Architecture for adaptive error robustness

This section describes the method proposed to attain the research goal defined above, which aims at obtaining optimal error robustness for different streaming scenarios, characterised by different video content and network conditions. Figure 7.3 illustrates the proposed architecture, whose main objective is to achieve an optimal trade-off between error robustness tools and coding efficiency.

Firstly, the input signals are coded using different levels of error protection. Then, an optimisation algorithm is used to decide which bitstream should be transmitted at any given instant, based on different parameters. The selected stream is expected to obtain the highest quality at the decoder-side for the given conditions. A reference video coder is used to compress the video signal using highly efficient coding techniques, obtaining the best R-D performance for packet loss free environments. This is used as the reference bitstream. Subsequently, a transcoding operation is applied to this reference bitstream aided by the input video signal to efficiently select the reference frames used for inter prediction. The reference frame selection mechanism proposed in Chapter 5 is used for this purpose. Moreover, loss simulation and a saliency-based EC optimisation are also used in the robust coding process, which enables the selection of the optimal EC algorithm to be applied at the decoder. Finally, since the optimal EC mode might not be able to correctly reconstruct all the missing regions, extra residual information is multiplexed in the stream as a scalable EL using two different bitrates. In this case 10% and 20% of the total bitrate were chosen, providing two different levels of error robustness. The EC optimisation and EL coding is performed as described in Chapter 6. The proposed approach results in four different bitstreams with different characteristics (referred to as S1 to S4, as shown in Figure 7.3): (i) S1: reference stream without error robust coding; (ii) S2: robust video coding using reference frame selection; (iii) S3: robust video coding using selective reference frames and optimal EC method signalling plus 10% of bitrate allocated for the scalable enhancement layer EC residue; (iv) S4: same as the previous bitstream but with 20% bitrate allocated for the scalable EL.

The video stream with the optimal error robustness is selected using an NN, which receives different parameters extracted from the input video, the compressed bitstream and information from the network conditions. The relevant video characteristics that

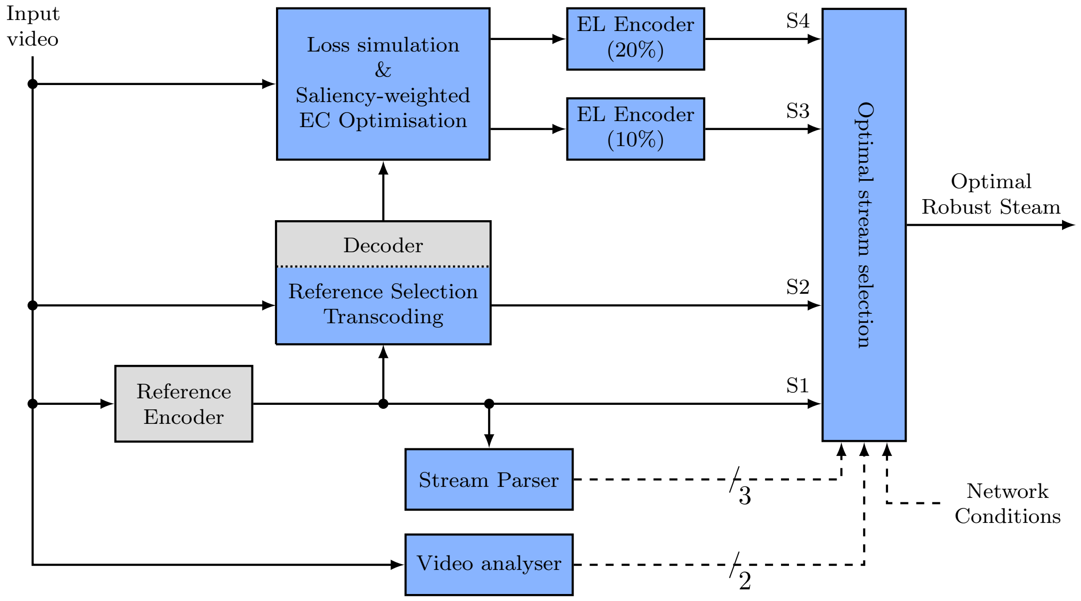


Figure 7.3: Adaptive error robustness architecture for 4 streams (S1...S4).

are useful for this algorithm are known measures of spatial and temporal complexity [159], defined in Section 7.4.2. These are computed by the functional block defined as ”Video analyser” in Figure 7.3. Moreover, further information is also extracted from the coded bitstream using a stream parser. Finally, the packet loss ratio of the transmission channel is also used. These parameters are fed into the optimisation block to select the best bitstream for transmission. The optimisation process is described in detail in the following section. The main purpose of this architecture is to allow dynamic stream switching during real time transmission. To accomplish this without causing drift at the decoder, the switching procedure should take place at refresh points, i.e., IDR and CRA frames, which correspond to clean random access points. In the proposed scheme, IDR frames are introduced every 16 frames, resulting in between 1 to 3 switching points every second, depending on the frame rate of the video signal.

7.4 Robust stream selection method

In this section the optimisation method used to select a robust video stream is presented. The selection method is based on a shallow neural network which predicts the optimal stream based on different parameters. This section describes both the input parameters and the neural network used in the optimal stream selection process. Moreover, a description data set used to train and evaluate the neural network is also provided.

7.4.1 Data set

The proposed method includes an optimisation stage that uses a neural network classifier, in order to select the optimal stream to be used for transmission. The stream that should be selected is the one which obtains the highest quality for a particular transmission scenario. Therefore, several test cases were simulated in order to obtain results for a wide range of conditions. These tests are mainly characterised by three elements:

- Video signal;
- Target bitrate;
- Packet loss ratio.

To obtain the data set, 14 different test sequences with different scene content were used, as listed in Table 7.1. Those sequences were encoded with the main bitrate (see Table 7.1), four different others, obtained by reducing the main bitrate by 10%, 20%, 30% and 40%, and one with 10% of bitrate increase. Then, the following PLRs were used to cover a wide range of loss ratios giving more detail to low PLRs: 0%, 0.5%, 1.0%, 1.2%, 1.6%, 2.0%, 2.2%, 2.4%, 2.6%, 2.8%, 3.0%, 5.0% and 10.0%. This results in a total of 1092 test conditions.

7.4.2 Input parameters

A set of different parameters is used to predict the optimal stream that achieves the highest quality in specific network conditions. Based on these inputs the proposed scheme should be able to select the appropriate robust coding stream that should be transmitted. The parameter set characterises three different aspects of the video transmission system; (i) input video (two parameters), (ii) video stream (three parameters) and (iii) transmission network (one parameter). From the input video the two parameters are measured in a frame by frame basis, while for the video stream information a stream parser is applied to the reference bitstream that was compressed only based optimal R-D optimisation without considering any robust coding techniques, as shown in Figure 7.3. Finally, the parameter related with the transmission network can be obtained using feedback information from the decoder or the network itself.

The characterisation of these three aspects of the video transmission results in a total of six parameters:

1. *Spatial Information (SI)*: parameter expressing the amount of spatial detail, with impact on the error recovery performance [159].

2. *Temporal Information (TI)*: parameter expressing the amount of difference between temporally adjacent frames [159]. This has a direct relation to the impact of errors and performance of the error resilience techniques, as shown in the results of Figures 7.1 and 7.2, which were discussed in Section 7.2.

3. *Inter (percentage)*: the percentage of the image that is compressed using inter-frame prediction techniques. This is an indication of the distribution of spatial and temporal dependencies. This is calculated as the ratio between inter-coded blocks and intra-coded ones. Normally, coded streams with higher number of temporal dependencies are more prone to error propagation, thus higher number of bits should be allocated for error protection [113].

4. *AVG_MV*: for the inter-coded regions, the motion information is analysed and the average MV magnitude is calculated to provide information about the motion intensity. As shown in previous studies [148,149,172], motion-based EC algorithms are less efficient for video signals with higher motion intensity, as MVs are less predictable using algorithms, such as MC or MVE.

5. *Bitrate*: information regarding the average bit allocation per frames is calculated, since it will affect the video quality and the amount of bits that can be available for auxiliary information. This value is expressed in Mbps, as illustrated in Table 7.1.

6. *Packet Loss Ratio (PLR):* feedback information from the network to characterise the transmission losses, expressed by the percentage of packets loss. This is used since for higher loss ratio there is a clear increase in the quality degradation, which affects the performance of the robust coding techniques. Transmission networks with higher loss ratios require a higher bit allocation for robust techniques, therefore the error resilience technique should be able to adapt based on the value of the PLR.

These parameters were compared against the optimal stream, selected from S1 to S4. This aims at finding the correlation between the input parameters and the output variable, i.e., target stream. In this study the streams are characterised by a number between 1 and 4. Table 7.2 shows the absolute Person Correlation coefficient [185] obtained across all sequences, bitrates and PLRs. The table shows the results for the

Table 7.2: Pearson Correlation coefficient between the different parameters and the optimal robust coding techniques (optimal stream).

Parameter	Absolute Pearson Correlation
SI	0.3953
TI	0.6079
Inter (percentage)	0.6487
OPT_FLOW	0.1553
AVG_MV	0.4647
Bitrate	0.3580
PLR	0.5755

parameters previously described, as well as, a new parameter obtained from the average magnitude of the optical flow measured using the method in [176] (*OPT_FLOW*). Results in Table 7.2 show that the selected parameters are all related with the optimal streams, with the ratio of inter-coded blocks having the highest correlation. Moreover, one can also notice that the parameters related with the motion activity, i.e., TI and *AVG_MV*, are also correlated with the output variable. Since both *AVG_MV* and *OPT_FLOW* represent similar information (motion insensitive), only *AVG_MV* was used in the proposed method, due to its higher Person Correlation coefficient.

7.4.3 Neural network

A NN is a computational model based on the structure and functions of biological neural networks. Mainly, a NN consists of three different layers: the input layer which receives data from an external environment, the hidden layer which processes the inputs and an output layer [186]. In the past NNs have been widely used in video processing applications such as, quality evaluation [187], video analysis [188] and video coding [65,189]. Neural networks are used in the proposed method since they are universal estimators, which can fit arbitrary complex data not easily approximated by analytical models [190]. Therefore, the complex relationship that describes the dependency of video quality from the input video, compressed stream and network parameters is adequate to be modelled by a NN, which can capture multi-dimensional statistical nature of the problem through the training process. In this work a feed-forward network is used, which connects every neuron in a given layer to the neurons of the subsequent layer.

Figure 7.4 shows the diagram of the NN used in the proposed method. The network

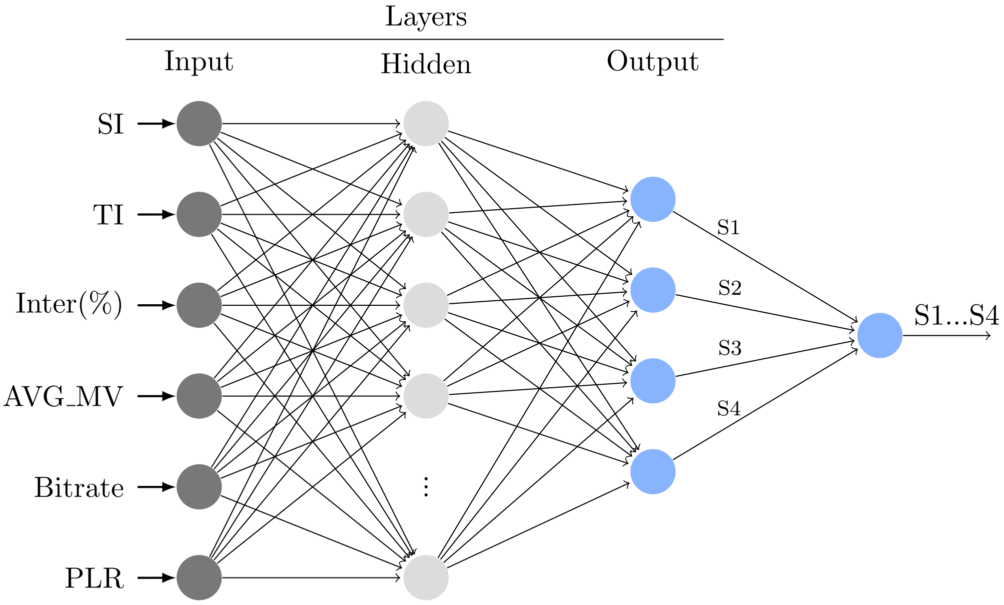


Figure 7.4: Representation of the neural network used in the proposed method.

contains an input and output layer, as well as, a single hidden layer. The input layer receives the six inputs, as described in the previous sub-section. These inputs are correlated with the optimisation variable being maximised, i.e., the output video quality. Subsequently, the network transforms the input data through a series of neurons in the hidden layer. The output layer comprises four neurons corresponding to each available video stream from S1 to S4 (different error resilience techniques). Finally, the output with the highest value is selected and the corresponding stream is used for transmission, since it is the one expected to achieve the highest video quality for a particular scenario.

In the proposed approach a shallow NN is used for simplicity and to avoid over-fitting, due to the limited test cases [191]. As shown in Figure 7.4, the NN used in this method has three layer: an input layer with 6 nodes, an hidden layer with 32 nodes and an output layer with 4 nodes. By avoiding over-fitting, it is guaranteed that the proposed method would continue to work in future test conditions, not used for the NN training. The number of nodes in the hidden layer was determined by comparing the performance obtained for different cases using the cross-entropy performance metric [192].

Training and validation

To effectively train and validate the NN, firstly the cases corresponding to three video sequence were left out, to be used in the performance evaluation of the proposed

Table 7.3: Confusion matrix of the training set.

		Target			
		S1	S2	S3	S4
Selected	S1	132	32	6	6
	S2	16	57	13	0
	S3	4	10	69	18
	S4	9	10	17	158
Accuracy		82.0%	52.3%	65.7%	86.8%

method. The remaining data set was randomly partitioned into three different groups: (i) 65% of the test cases are used for training, (ii) 15% for validation and (iii) 20% for testing the NN performance. The choice of these percentages follow the commonly used approach where about 70% to 80% of the dataset is used for training and validation while the remaining 30% to 20% is used for testing [193]. In this work the NN supervised training uses 557 test conditions and a scaled conjugate gradient algorithm [194] to calculate the backpropagation. Then, 129 test conditions were used for NN validation and 172 conditions o to evaluate the accuracy of the NN after training with different conditions. The 11 test sequences, encoded at different bitrates and subject to a wide range of PLRs, were used to generate the three groups of the dataset. Since these sequences have quiet different characteristics (e.g, motion, texture and resolution) each group is considered representative of any type of video content. The training was stopped when the cross-entropy obtained for the validation samples stop decreasing, which reveals that the NN starts to over-fit the training samples and it is not able to generalise to new data.

Tables 7.3, 7.4 and 7.5 show the confusion matrices obtained for training, validation and test sets, respectively. In these matrix each column corresponds to the target result, i.e., the cases where the coded streams, S1 to S4, would lead to the highest decoded video quality. Similarly, each row corresponds to the NN selected by the trained network. Thus, those cases where the horizontal and vertical labels match, indicate that the network produced the correct output. The remaining positions correspond to the erroneous outputs. The last line shows the accuracy for each target output. These results correspond to the dataset used to train the NN.

Results in Table 7.3 shows that streams S1 (i.e., without robust coding techniques) and S4 (i.e., using optimal EC with 20% of bitrate in the EL) lead to the highest quality in the majority of the test cases (are more often selected). The NN is able to achieve a high accuracy for S1 and S4 streams, correctly selecting the output stream

Table 7.4: Confusion matrix of the validation set.

		Target			
		S1	S2	S3	S4
Selected	S1	28	4	1	2
	S2	4	12	6	0
	S3	1	7	20	5
	S4	1	0	6	32
Accuracy		82.4%	52.2%	60.6%	82.1%

Table 7.5: Confusion matrix of the test set.

		Target			
		S1	S2	S3	S4
Selected	S1	41	10	5	1
	S2	5	14	2	0
	S3	2	0	23	7
	S4	3	4	10	45
Accuracy		80.4%	50.0%	57.5%	84.9%

in more than 80% of the cases. Regarding the remaining robust streams (i.e., S2 and S3), an acceptable accuracy of 52.2% and 60.6% is achieved, respectively. This reveals that the NN used in the proposed method is able to predict the coded stream which leads to the highest video quality across different conditions with acceptable accuracy. The results from the validation set, shown in Table 7.4, reveal that the NN training did not over-fit, since it shows similar accuracy as the results shown in Table 7.3.

To further evaluate the performance of the NN, the confusion matrix of the data set used for testing (test set) is shown in the Table 7.5. This table shows the same results as Table 7.3, but for different test cases, which were not used in the training phase. Results show that the proposed NN is able to achieve an accuracy of at least 50% with a maximum value of 84.9% when the S4 is the target stream. Although a lower accuracy is obtained using the test set, the overall accuracy is still acceptable, since it allows to select the optimal robust coding technique (i.e., S1 to S4) for the majority of the conditions. In summary, the results presented in this sub-section reveal that the input parameters used in the proposed NN are correlated with the decoded video quality, as they can effectively be used to perform data categorisation used in the selection of the optimal error robustness coding technique for the HEVC standard.

7.5 Performance evaluation

In this section the performance of the proposed method is evaluated and discussed. The overall goal of the proposed method is to achieve improved quality using a combination of different robust coding techniques adaptively selected using a feed-forward NN. The error resilience techniques comprising the proposed method are used in the context of the HEVC standard and compared with the reference implementation (*Ref*). The quality obtained with the proposed optimal error robustness scheme (*Prop*) is compared against the reference HEVC (*Ref*) and the individual methods used to generate the streams S1 to S4.

Figure 7.5 shows the quality obtained for the proposed method and the *Ref* case for the main target bitrate shown in Table 4.1. Moreover, the quality obtained with the target stream (*Target*) is also shown. One should note that the *Target* case corresponds to the optimal selection, which can only be performed after quality evaluation at the decoder, therefore it is not suitable for real time applications. It is only used to establish the optimal case for comparison purposes. The results in the figure allow the comparison between three relevant cases: proposed stream selection method, the optimal case and the reference method.

As shown in Figure 7.5, the proposed method clearly outperforms the reference HEVC for all PLRs. Moreover, in the packet loss free case (PLR=0), it leads to imperceptible quality reduction, since the proposed approach is able to select the optimal stream, thus disabling any robust coding technique when the transmission network does not have packet losses. In case of data loss, the results show that the proposed method is able to match the quality obtained by the *Target* case in most cases, revealing that the NN is able to efficiently select the optimal stream for transmission.

Subsequently, a comparison between the different methods is performed, in order to evaluate the quality gains obtained with the proposed method in comparison with each of the different techniques. Table 7.6 shows the average quality gains in comparison with the reference HEVC for the test set. These were obtained with the proposed scheme and the four different techniques used in selection method, i.e., methods used in the streams S1 to S4. The average quality gains are measured across all test cases, i.e., test set with different bitrates and different packet loss ratios.

Results in Table 7.6 show that the robust coded streams S1 to S4 are able to achieve acceptable average quality gains up to 1.52 dB (see results for S3). However, since the implementation of robust coding techniques decreases the coding efficiency, it also leads to lower quality than the reference HEVC. This is noticeable in the results obtained

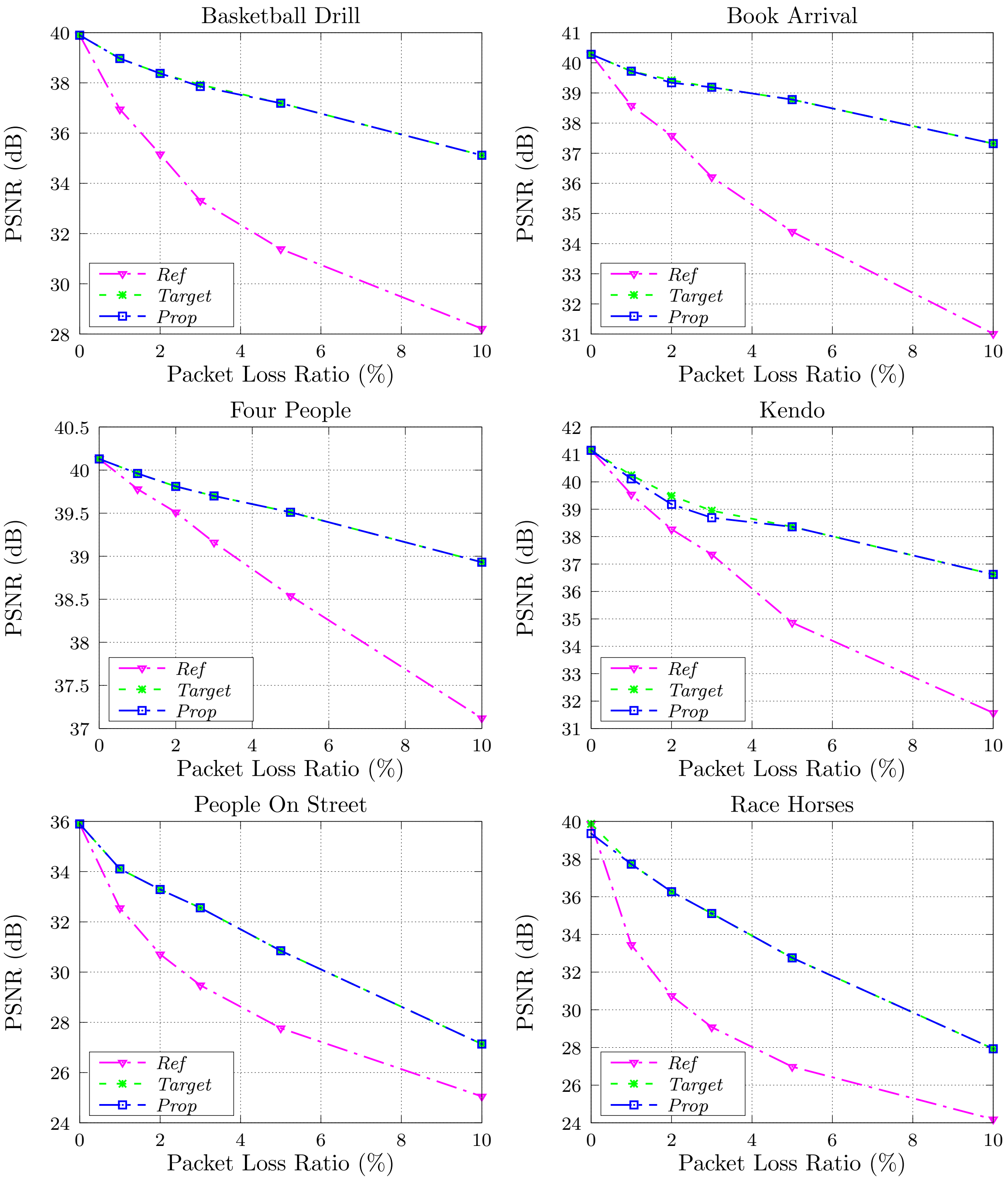


Figure 7.5: Average video quality obtained for the reference HEVC (*Ref*), the optimal stream (*Target*) and the one selected by the proposed approach (*Prop*).

for the Four People sequence in the streams S2 to S4 and for the Kristen and Sara sequence in the streams S3 and S4 (see red ellipses). The overall performance of such methods reveals that they are not the best options for transmission conditions with low PLRs (higher number of conditions have PLRs lower than 3%). On the contrary, the proposed method is always capable of outperforming the reference HEVC and achieve

Table 7.6: Average quality gains (ΔPSNR in dB) in comparison with the reference HEVC across different bitrates and packet loss ratios.

Sequence	**Robust coded stream**				
	S1	S2	S3	S4	*Prop*
Basketball Drill	+1.68	+2.13	+2.81	+2.62	+2.81
Book Arrival	+1.00	+1.51	+1.79	+1.74	+1.87
BQSquare	+0.98	+1.14	+1.06	+0.73	+1.20
Cactus	+1.22	+1.36	+1.95	+1.85	+1.88
Four People	+0.30	-1.19	-0.12	-0.61	+0.28
Kimono	+1.55	+2.68	+2.86	+3.06	+3.07
Kristen Sara	+0.38	-0.52	+0.20	-0.38	+0.30
Park Scene	+1.65	+1.73	+2.24	+2.10	+1.83
Party Scene	+0.21	+0.11	+0.17	+0.44	+0.30
People on Street	+1.01	+1.13	+1.38	+1.62	+1.76
Tennis	+1.18	+2.11	+2.43	+3.28	+3.28
Average	**+1.01**	**+1.11**	**+1.52**	**+1.50**	**+1.69**

quality gains from 0.28 dB to 3.28 dB. For the Kimono and Tennis sequences, it is able to achieve the highest quality gains, of up 3.07 dB and 3.28 dB, respectively (see blue ellipses). Summarising, by adaptively selecting the video stream, and implicitly the robust coding technique, the proposed method is able to achieve a better trade-off between coding efficiency and error robustness, achieving an average quality gain across different bitrates and PLRs of 1.69 dB.

Finally, the performance of the proposed method is evaluated for video sequences, which was not fed into the NN training phase. This study aims at finding the accuracy of the optimal stream selection in case of a new video content. For this purpose the Basketball Drive sequence was randomly selected, and not used in previous simulations. Table 7.7 shows the average quality results obtained with these sequences, similarly to Table 7.6. These results also confirm that the proposed method is able to select the optimal robust stream to be used for transmission, even when new video content is transmitted. The proposed method achieves a quality gain up to 4.46 dB in comparison with the reference HEVC, outperforming the quality achieved with the stream S4 (i.e., 4.19 dB). This confirms the advantage of the proposed method to adaptively select among different streams with different levels of error protections.

Table 7.7: Average quality gain (ΔPSNR in dB) in comparison with the reference HEVC for video sequences not used in the NN training.

Sequence	**Robust coded stream**				
	S1	S2	S3	S4	*Prop*
Basketball Drive	+2.10	+2.68	+3.36	+3.45	+3.55
Kendo	+1.92	+3.24	+3.94	+4.19	+4.46
Race Horses	+0.85	+1.02	+0.88	+0.75	+1.06

7.6 Summary

In this chapter an adaptive error robustness method is proposed, based on reference frame selection and EC-aware optimisation. By adaptively selecting a different error robustness technique at the encoder-side using a pre-trained NN, higher quality gains are achieved when comparing against the individual reference techniques. The overall results demonstrate that the proposed method is able to increase the reconstruction video quality in case of errors. Also, by adapting the transmitted stream to the network conditions, a better trade-off between coding efficiency and error robustness can be achieved. The consistent quality improvement, allows the use of this method in dynamic broadcasting video services where network feedback is available. This architecture is suitable for dynamic streaming technologies, e.g., HTTP streaming based on MPEG-DASH [18] or content delivery over heterogeneous networks using the MMT [20].

Chapter 8

Conclusion and future work

This thesis investigated new algorithms and techniques for efficient error robustness in highly efficiency video coders. The importance of this research work is justified by the widespread of digital video content, including high resolution formats, multiview and 360-degree video formats, which significantly increase the network bandwidth requirements. Consequently, errors and data loss may occur along the transmission system, despite the fact that video coding standards have been improved, in order to meet the network bandwidth constrains. The use of advanced predictions techniques, results in increased bitstream dependencies with negative impact in the effect of errors, leading to higher quality degradation. In this thesis different methods were proposed to address this problem, by investigating robust coding and efficient EC.

Firstly, an introduction of the main motivation for this research work was presented, enumerating its main objectives and contributions. Then, the main coding tools of the state-of-the-art HEVC were described. This description covered the high-level features, such as the new picture types and the slice partitioning schemes, the new block partitioning scheme, which allow multi-level hierarchical structures with size-independent representation, the main predictions techniques, including the newly introduced advanced motion vector prediction technique. This provided the necessary knowledge about the novel techniques in HEVC, which may have a relevant impact on the error robustness of the coded streams. The main algorithms proposed in the past to increase the error robustness of video streams were also presented, describing techniques based on error resilience R-D optimisation, redundant based approaches and EC-based methods. After a simulation study it was concluded that, although previous works provide efficient techniques to reduce the bitstream dependencies and ease the error recovery, vulnerabilities of the HEVC standard could not be fully addressed with existing techniques. The main problems range from new MV coding techniques to the lack of robust coding mechanisms, such as FMO or data partitioning. Moreover, there were also open issues regarding robust EC-based encoding, as existing methods do not fully exploit the use of different techniques with different levels of complexity, especially the use of advanced EC-based approaches.

To investigate new techniques for robust video coding an evaluation study was firstly carried out, with the aim of finding the influence of the new coding tools on the error resilience of the HEVC streams. Existing studies were limited to generic comparisons between HEVC and previous standards (e.g., H.264/AVC). This study extended existing ones by considering different coding configurations, which allowed to evaluate the impact of new coding tools. The study revealed that some features, such as increased block size, do not have significant impact on the error robustness performance. However, slice partitioning and intra-refreshing have a noticeable impact on both coding efficiency and error robustness. This led to the insight that a small compromise in coding efficiency should be accepted, in order to achieve an higher quality in case of errors. Finally, one could also conclude that the use of temporal predictions in MV coding significantly increases temporal dependencies, leading to higher quality degradation and compromising the performance of HEVC-based video transmission systems.

Based on the above findings, a new MV coding scheme was devised, aiming at reducing the temporal dependencies. In this scheme the use of the temporal MV candidate is selectively used in some PUs, preventing temporal dependencies across several frames. As this method significantly reduces the temporal dependencies, the number of mismatched MVs is decreases. Moreover, as the temporal candidate is not fully disabled the impact on coding efficiency was shown to be low. Experimental results demonstrated that a clear improvement on decoded video quality is achieved, across different video content, with a small bitrate increase.

Another investigation carried out in this thesis was focused on the reference frame distribution in highly efficient video coders. Since in general, the reference frame usage is unbalanced, a new method based on constrained R-D optimisation was proposed to equally distribute the use of multiple reference frames across neighbouring PUs. Moreover, the use of selective MV redundancies was also investigated, in order to recover from erroneous motion predictions. This resulted in a two-fold approach to reducing error propagation. As part of the proposed method can be applied at the streaming-stage, using a low complexity parser, it can be used to adapt pre-encoded streams to error-prone transmission channels. The proposed method was compared against existing techniques, and the simulation results revealed a consistently superior performance of using the two-fold approach for different packet loss ratios.

To address the problem of robust video decoding and increase the quality of the reconstructed frames, EC-aware resilience techniques were also investigated. A saliency-based EC optimisation was proposed to optimally select the EC method that should be applied at the decoder in case of frame loss. This information is transmitted to the decoder as side-information. To keep the signalling overhead affordable, saliency estimation and a parameter based on the temporal duration of the error propagation Then, a limited amount of overhead was used to encode a residual information in an EL, which is used to improve the reconstruction quality, in those regions where the optimal EC was not accurate. Experimental results revealed that using different EC algorithms contributes to increase the performance of the EC-aware coding scheme, achieving superior performance when compared with existing methods.

Finally, the problem of adaptive robust coding was investigated, in order to achieve superior quality when different error robustness techniques are available and dynamic conditions are taken into account. Simulation results showed that the robust coding technique which obtains the higher quality depends on different factors, such as video content, bitrate and network conditions. Therefore, an adaptive scheme based on six input parameters and a pre-trained NN was proposed, to predicts the video stream that is able to achieve the highest decoded quality, under packet loss environments. Results revealed that using the proposed adaptive scheme, an efficient trade-off between coding efficiency and error resilience is achieved for the majority of test conditions. Using this architecture, the problem of robust video coding and decoding was addressed using a combination of different methods, which leads to a superior performance in comparison with individual techniques.

As a final remark, the novel methods proposed in this thesis are also suitable for emerging formats such as, multi-view and 360-degree video. Since not all visual information is equally important at the same time thus, dynamic error resilient methods like the ones investigated in this thesis are particularly suitable for these formats. Overall, one may conclude that the use of the novel methods proposed in this thesis consistently contributes to increase the robustness of video communications using state of the art high efficiency video encoding tools in error prone networks. The best performance is achieved by combining robust video coding techniques with efficient error concealment algorithms optimised at the encoder, using an optimised dynamic approach.

Future work

In the context of this thesis there are some open issues for future research directions which can be investigated to further strength the overall performance.

In the context of robust coding, one can further combine existing error resilience techniques, e.g., intra-refreshing constrains, in the Lagrangian optimisation used in the reference picture selection proposed in Chapter 4. This would possibly allow stronger limitation of error propagation by introducing extra refresh points. Moreover, one can also use more advanced constrains, such as, end-to-end based optimisations, to devise more efficient methods for distribution of the use of reference frames.

Regarding the EC-aware method proposed in Chapter 6 there are also some possible improvements to be addressed in future research. Firstly, more advanced algorithms can be developed on a CTU basis, in order to improve the reconstruction accuracy. This can be implemented either by replacing the existing methods, or by extending beyond the four candidates that were used in this work. Alternatively, to reduce the computational complexity overhead at the encoder, instead of measuring the R-D cost for each method, pre-trained classification algorithms can be investigated to predict the optimal EC mode for each block. As the complexity of the decoder is also a relevant issue, due to the increasing of mobile and battery-based devices, the optimal EC selection should also take into account the end devices, thus penalising high complexity algorithms (e.g., algorithms requiring an estimated motion field) in case of constrained devices. The performance of the adaptive selection of the robust coding technique can also be improved by combining different classification algorithms, based on more complex neural network architectures.

Furthermore, the error robustness schemes proposed in this research work can be extended to cope with robust transmission of multi-view video, where errors propagate in the temporal and inter-layer domains. Finally, as a new coding standard, entitled Versatile Video Coding, is currently being developed, targeting the deployment of higher-quality video services and emerging applications such as, 360-degree multimedia. This opens new research directions to the study the impact of errors in such video applications and the development of novel efficient techniques to guarantee high quality visual experiences, to users of such emerging video formats.

Appendix A

Test video sequences

This appendix illustrates the original test signals used in simulations throughout the thesis. This includes a partial set of the HEVC test material and other external test material.

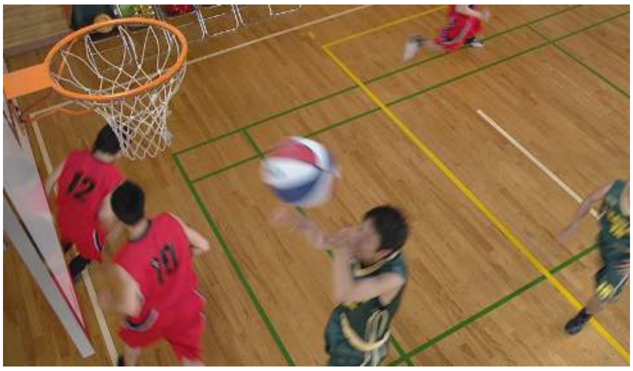


Figure A.1: Basketball Drill, 832 × 480, 50 fps, 240 frames.

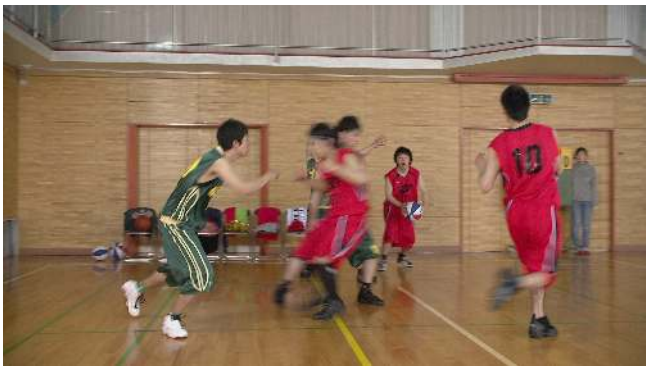


Figure A.2: BasketballDrive, 1920 × 1080, 50 fps, 240 frames.

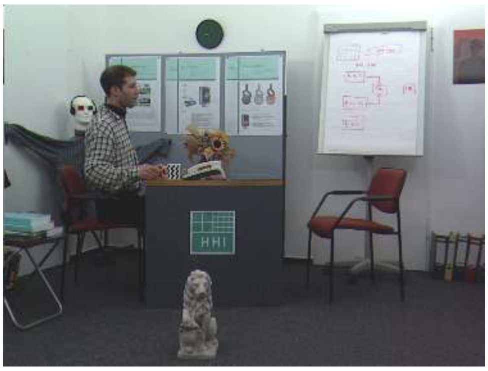


Figure A.3: Book Arrival, 1024 × 768, 30 fps, 240 frames.

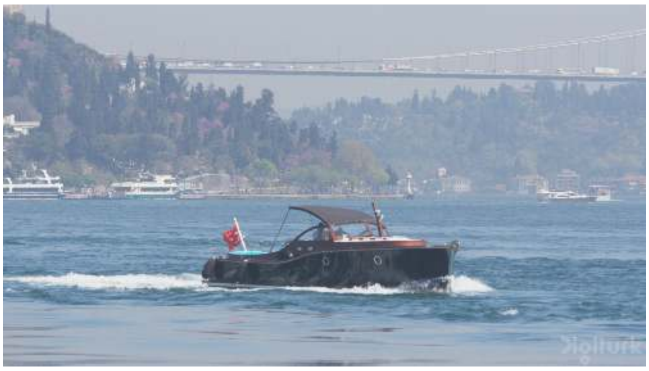


Figure A.4: Bosphorus, 3840 × 2160, 120 fps, 240 frames.

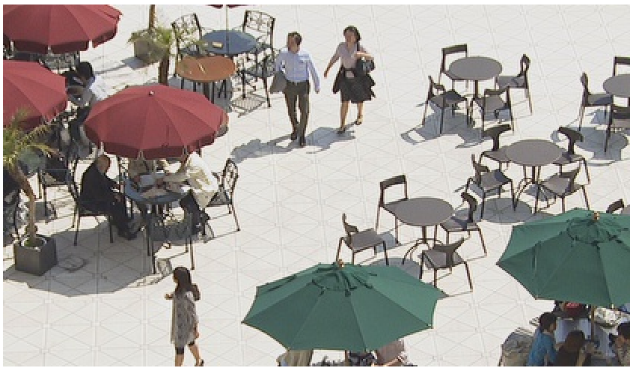


Figure A.5: BQSquare, 416 × 240, 60 fps, 240 frames.

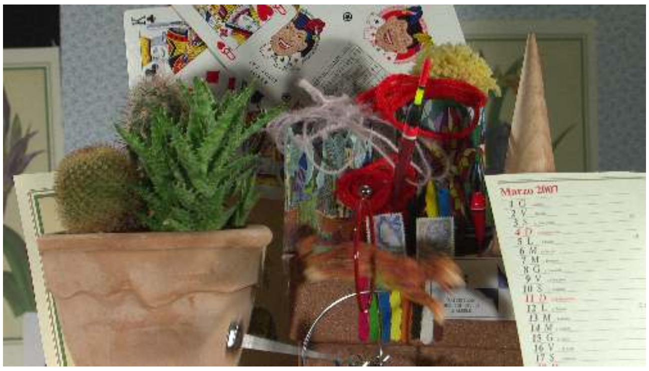


Figure A.6: Cactus, 1920 × 1080, 50 fps, 500 frames.

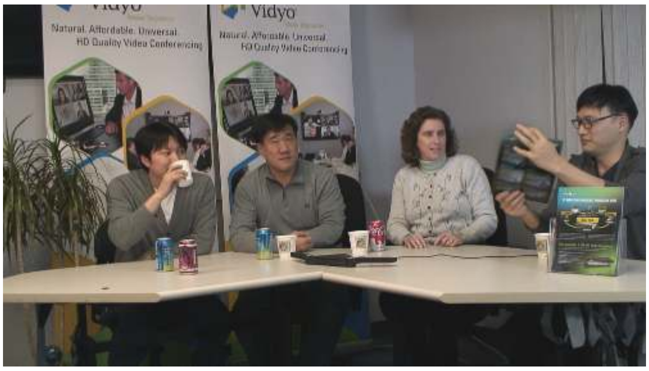

Vidyo
Natural. Affordable. Universal.
HD Quality Video Conferencing
Vidyo
Natural. Affordable. Universal.
HD Quality Video Conferencing

Figure A.7: Four People, 1280 × 720, 60 fps, 240 frames.

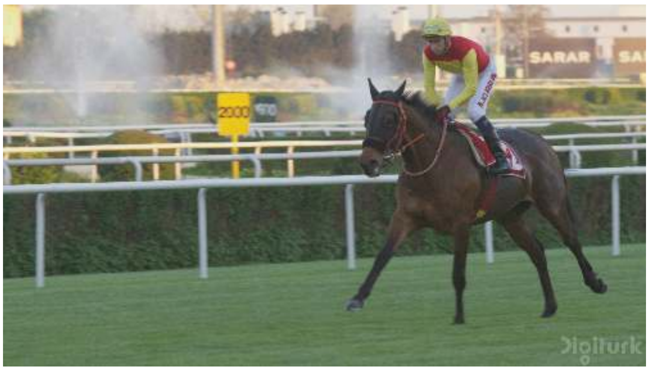


Figure A.8: Jockey, 3840 × 2160, 120 fps, 240 frames.

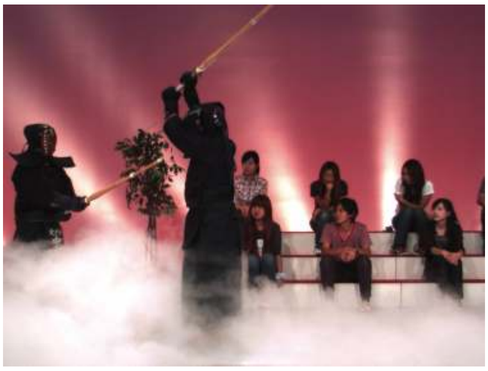


Figure A.9: Kendo, 1024 × 768, 30 fps, 240 frames.

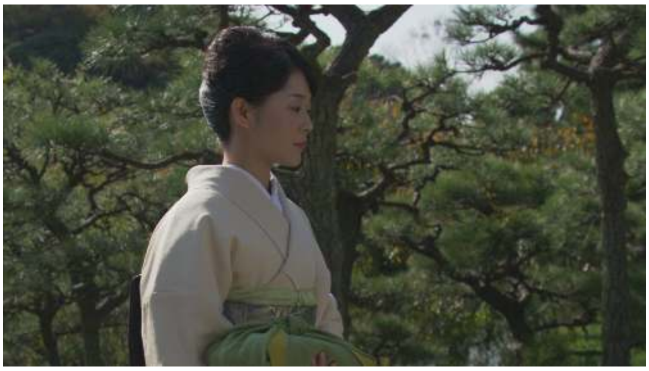


Figure A.10: Kimono, 1920 × 1080, 24 fps, 240 frames.

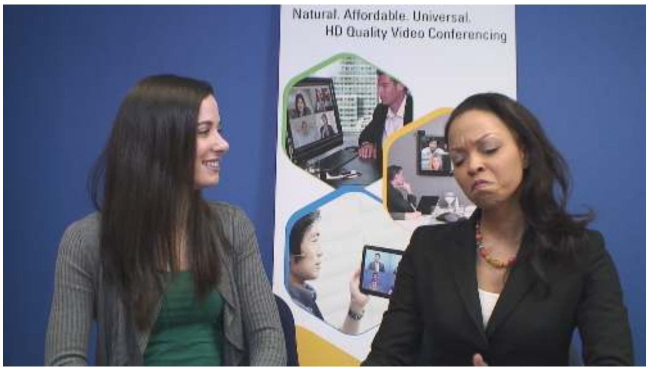

Natural. Affordable. Universal.
HD Quality Video Conferencing

Figure A.11: Kristen and Sara, 1280 × 720, 60 fps, 240 frames.

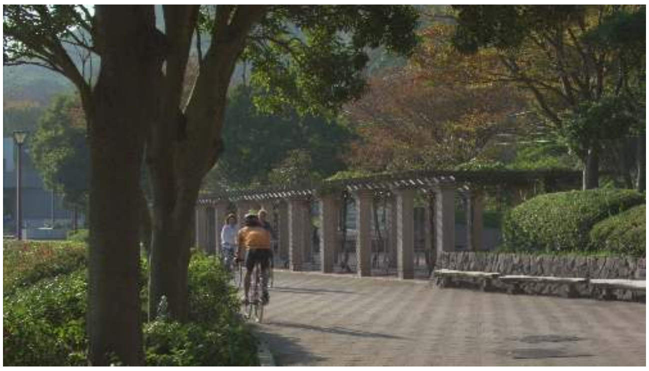


Figure A.12: Park Scene, 1920 × 1080, 24 fps, 240 frames.

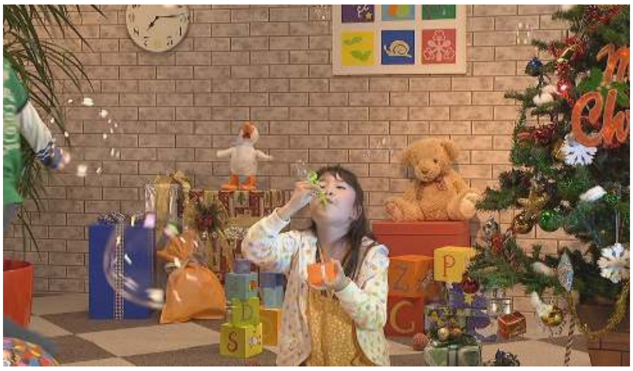


Figure A.13: Party Scene, 832 × 480, 50 fps, 240 frames.

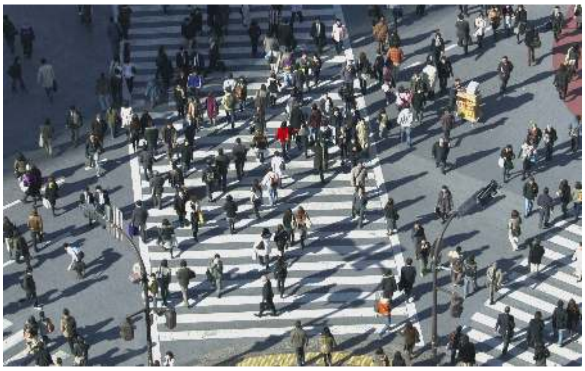


Figure A.14: People On Street, 2560 × 1600, 24 fps, 150 frames.

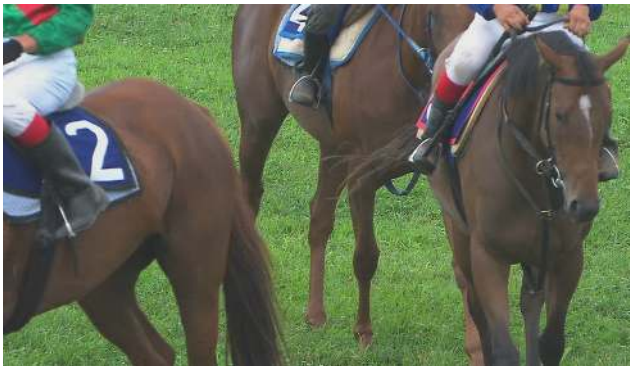


Figure A.15: Race Horses, 832 × 480, 30 fps, 240 frames.

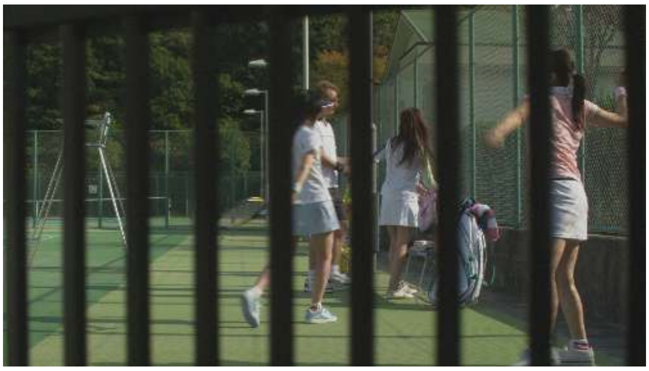


Figure A.16: Tennis, 1920 × 1080, 30 fps, 240 frames.

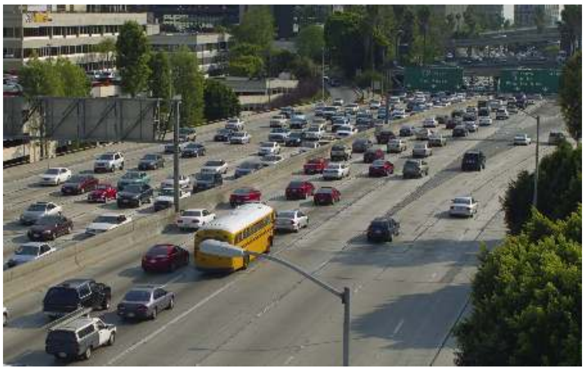


Figure A.17: Traffic, 2560 × 1600, 30 fps, 150 frames.

Appendix B

Published papers

This appendix presents a list of the published papers that resulted from the research work done during this thesis.

B.1 Journal publications

J1 J. Carreira, P. Assuncao, S. Faria, E. Ekmekcioglu, and A. Kondoz, "A two-stage approach for robust HEVC coding and streaming," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1960-1973, Aug. 2018.
DOI:10.1109/TCSVT.2017.2691471

J2 J. Carreira, P. Assuncao, S. Faria, E. Ekmekcioglu, and A. Kondoz, "Error concealment-aware encoding for robust video transmission," *IEEE Transactions on Broadcasting* (available in IEEEXplore since Aug. 2018).
DOI:10.1109/TBC.2018.2865644

B.2 Conference publications

C1 J. Carreira, V. D. Silva, E. Ekmekcioglu, A. Kondoz, P. Assuncao, and S. Faria, "Dynamic motion vector refreshing for enhanced error resilience in HEVC," in *22nd European Signal Processing Conference (EUSIPCO)*, Sep. 2014, pp. 281–285.
ISBN:978-0-9928-6261-9

C2 J. Carreira, E. Ekmekcioglu, A. Kondoz, P. Assuncao, S. Faria, and V. D. Silva, "Selective motion vector redundancies for improved error resilience in HEVC," in *IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 2457–2461.
DOI:10.1109/ICIP.2014.7025497

C3 J. Carreira, S. Faria, P. Assuncao, E. Ekmekcioglu, and A. Kondoz, "Error resilience analysis of motion vector prediction in HEVC," in *Conference on Telecommunications (Conftele)*, Oct. 2015, pp. 1–4.
URL: `https://www.it.pt/Publications/DownloadPaperConference/20422`

C4 J. Carreira, P. Assuncao, S. Faria, E. Ekmekcioglu, A. Kondoz, and H.Lim, "Reference picture selection using checkerboard pattern for resilient video coding," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*, Dec. 2015, pp. 1–5.
DOI:10.1109/VCIP.2015.7457852

C5 J. Carreira, P. Assuncao, S. Faria, E. Ekmekcioglu, and A. Kondoz, "A robust video encoding scheme to enhance error concealment of intra frames," in *IEEE International Conference on Circuits and Systems (ISCAS)*, May 2017, pp. 1–5.
DOI:10.1109/ISCAS.2017.8050576

Bibliography

[1] ITU-T, ISO/IEC JTC1, "Advanced video coding for generic audiovisual services. ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC)," ITU-T/ISO, London, UK, Standard, May 2003.

[2] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[3] "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021," Cisco, Tech. Rep., Feb. 2017.

[4] ISO/IEC JTC1 ITU-T, "High efficiency video coding. ITU-T recommendation H.265 and ISO/IEC 23008-2," ITU-T/ISO, Standard, Feb. 2018.

[5] G. Sullivan, J. Ohm, W.-J. 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.

[6] Y. Yuan, I.-K. Kim, X. Zheng, L. Liu, X. Cao, S. Lee, M.-S. Cheon, T. Lee, Y. He, and J.-H. Park, "Quadtree based nonsquare block structure for inter frame coding in high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1707–1719, Dec. 2012.

[7] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1792–1801, 2012.

[8] I.-K. Kim, S. Lee, M.-S. Cheon, T. Lee, and J. Park, "Coding efficiency improvement of HEVC using asymmetric motion partitioning," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on*, Jun. 2012, pp. 1–4.

[9] R. Sjoberg, Y. Chen, A. Fujibayashi, M. Hannuksela, J. Samuelsson, T. K. Tan, Y.-K. Wang, and S. Wenger, "Overview of HEVC high-level syntax and reference picture management," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1858–1870, Dec. 2012.

[10] J. Ohm, G. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards - including high efficiency video coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[11] G. Correa, P. Assuncao, L. Agostini, and L. S. da Cruz, "Performance and computational complexity assessment of high-efficiency video encoders," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1899–1909, Dec. 2012.

[12] B. Oztas, M. Pourazad, P. Nasiopoulos, and V. Leung, "A study on the HEVC performance over lossy networks," in *19th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Dec. 2012, pp. 785–788.

[13] J. Nightingale, Q. Wang, C. Grecos, and S. Goma, "The impact of network impairment on quality of experience (QoE) in H.265/HEVC video streaming," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 2, pp. 242–250, May 2014.

[14] ——, "Subjective evaluation of the effects of packet loss on HEVC encoded video streams," in *IEEE Third International Conference on Consumer Electronics (ICCE)*, Sep. 2013, pp. 358–359.

[15] Y. Zhang, W. Gao, Y. Lu, Q. Huang, and D. Zhao, "Joint source-channel rate-distortion optimization for h.264 video coding over error-prone networks," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 445–454, April 2007.

[16] J. Nightingale, Q. Wang, and C. Grecos, "HEVStream: a framework for streaming and evaluation of high efficiency video coding (HEVC) content in loss-prone networks," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 404–412, May 2012.

[17] T. Schierl, M. Hannuksela, Y.-K. Wang, and S. Wenger, "System layer integration of high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1871–1884, Dec. 2012.

[18] D. Schroeder, A. Ilangovan, M. Reisslein, and E. Steinbach, “Efficient multi-rate video encoding for HEVC-based adaptive HTTP streaming,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, Aug. 2016.

[19] J. Gorostegui, A. Martin, M. Zorrilla, I. Alvaro, and J. Montalban, “Broadcast delivery system for broadband media content,” in *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Jun. 2017, pp. 1–9.

[20] K. Park, Y. Lim, and D. Y. Suh, “Delivery of ATSC 3.0 Services With MPEG Media Transport Standard Considering Redistribution in MPEG-2 TS Format,” *IEEE Transactions on Broadcasting*, vol. 62, no. 1, pp. 338–351, Mar. 2016.

[21] J. Wu, B. Cheng, M. Wang, and J. Chen, “Delivering high-frame-rate video to mobile devices in heterogeneous wireless networks,” *IEEE Transactions on Communications*, pp. 1–1, 2016.

[22] J. Wu, C. Yuen, M. Wang, and J. Chen, “Content-aware concurrent multipath transfer for high-definition video streaming over heterogeneous wireless networks,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 3, pp. 710–723, Mar. 2016.

[23] K. Park, N. Kim, and B. D. Lee, “Performance evaluation of the emerging media-transport technologies for the next-generation digital broadcasting systems,” *IEEE Access*, vol. 5, pp. 17 597–17 606, 2017.

[24] J.-T. Chien, G.-L. Li, and M.-J. Chen, “Effective error concealment algorithm of whole frame loss for H.264 video coding standard by recursive motion vector refinement,” *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1689–1695, Aug. 2010.

[25] T.-L. Lin, N.-C. Yang, R.-H. Syu, C.-C. Liao, and W.-L. Tsai, “Error concealment algorithm for HEVC coded video using block partition decisions,” in *IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, Aug. 2013, pp. 1–5.

[26] R. Hoffman, *Data Compression in Digital Systems*, 1st ed. London, UK: Chapman & Hall, Ltd., 1997.

[27] L. F. R. Lucas, E. A. B. da Silva, S. M. M. de Faria, N. M. M. Rodrigues, and C. L. Pagliari, *Efficient Predictive Algorithms for Image Compression.* Springer Publishing Company, 2017.

[28] B. Bing, *Video Coding Fundamentals.* Wiley-Blackwell, 2015, ch. 2, pp. 29–82.

[29] M. Pourazad, C. Doutre, M. Azimi, and P. Nasiopoulos, "HEVC: The new gold standard for video compression: How does HEVC compare with H.264/AVC?" *IEEE Consumer Electronics Magazine*, vol. 1, no. 3, pp. 36–46, Jul. 2012.

[30] V. Sze, M. Budagavi, and G. J. Sullivan, *High Efficiency Video Coding (HEVC): Algorithms and Architectures.* Springer Publishing Company, 2014.

[31] S. Wenger, "H.264/AVC over IP," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 645–656, Jul. 2003.

[32] A. Fujibayashi and T. Tan, "Random access support for HEVC, document JCTVC-D234," JCT-VC, Daegu, Korea, Tech. Rep., Jan. 2011.

[33] K. Misra, A. Segall, M. Horowitz, S. Xu, A. Fuldseth, and M. Zhou, "An overview of tiles in HEVC," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 6, pp. 969–977, Dec. 2013.

[34] R. Rodrguez-Snchez and E. S. Quintana-Ort, "Tiles-and WPP-based HEVC decoding on asymmetric multi-core processors," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, Apr. 2017, pp. 299–302.

[35] K. Chen, J. Sun, Y. Duan, and Z. Guo, "A novel wavefront-based high parallel solution for HEVC encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 181–194, Jan. 2016.

[36] I.-K. Kim, J. Min, T. Lee, W.-J. Han, and J. Park, "Block partitioning structure in the HEVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1697–1706, Dec. 2012.

[37] T. Nguyen and D. Marpe, "Performance analysis of HEVC-based intra coding for still image compression," in *Picture Coding Symposium (PCS)*, May 2012, pp. 233–236.

[38] N. Purnachand, L. N. Alves, and A. Navarro, "Improvements to TZ search motion estimation algorithm for multiview video coding," in *19th International*

Conference on Systems, Signals and Image Processing (IWSSIP), Apr. 2012, pp. 388–391.

[39] T. K. Lee, Y. L. Chan, and W. C. Siu, "Adaptive search range for HEVC motion estimation based on depth information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 2216–2230, Oct. 2017.

[40] S. H. Yang, J. Z. Jiang, and H. J. Yang, "Fast motion estimation for HEVC with directional search," *Electronics Letters*, vol. 50, no. 9, pp. 673–675, Apr. 2014.

[41] J.-L. Lin, Y.-W. Chen, Y.-W. Huang, and S.-M. Lei, "Motion vector coding in the HEVC standard," *IEEE Journal of Selected Topics in Signal Processing*, 2013.

[42] A. Saxena and F. C. Fernandes, "DCT/DST-based transform coding for intra prediction in image/video coding," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3974–3981, Oct. 2013.

[43] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 620–636, Jul. 2003.

[44] Y. Wang, S. Wenger, J. Wen, and A. Katsaggelos, "Error resilient video coding techniques," *IEEE Signal Processing Magazine*, vol. 17, no. 4, pp. 61–82, Jul. 2000.

[45] B. Yan, H. Gharavi, and B. Hu, "Pixel interlacing based video transmission for low-complexity intra-frame error concealment," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 253–257, Jun. 2011.

[46] A. Vetro, J. Xin, and H. Sun, "Error resilience video transcoding for wireless communications," *IEEE Wireless Communications*, vol. 12, no. 4, pp. 14–21, Aug. 2005.

[47] G. Kulupana, D. S. Talagala, H. K. Arachchi, and A. Fernando, "Error resilience aware motion estimation and mode selection for HEVC video transmission," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2016, pp. 85–86.

[48] H. M. Maung, S. Aramvith, and Y. Miyanaga, "Error resilience aware rate control and mode selection for HEVC video transmission," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2017, pp. 374–375.

[49] E. Baccaglini, T. Tillo, and G. Olmo, "Concealment driven smart slice reordering for robust video transmission," in *IEEE International Conference on Multimedia and Expo*, Jun. 2008, pp. 1173–1176.

[50] H. Yang and J. Boyce, "Concealment-aware motion estimation and mode selection for error resilient video coding," in *International Conference on Image Processing (ICIP)*, Oct. 2006, pp. 2229–2232.

[51] S. Chen and H. Leung, "A temporal approach for improving intra-frame concealment performance in H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 422–426, Mar. 2009.

[52] J. Y. Pyun, "Error concealment aware streaming video system over packet-based mobile networks," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 1705–1713, Nov. 2008.

[53] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[54] D. Hawkins, "Constrained optimization and lagrange multiplier methods," *Athena Scientific, Belmont*, 1982.

[55] Y. Zhang, W. Gao, H. Sun, Q. Huang, and Y. Lu, "Error resilience video coding in H.264 encoder with potential distortion tracking," in *International Conference on Image Processing (ICIP)*, vol. 1, Oct. 2004, pp. 163–166.

[56] R. Zhang, S. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, Jun. 2000.

[57] H. Yang and K. Rose, "Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 7, pp. 845–856, Jul. 2007.

[58] Y. Liao and J. D. Gibson, "Enhanced error resilience of video communications for burst losses using an extended ROPE algorithm," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 1853–1856.

[59] G. Peng, Y. Liu, Y. Hu, S. Ci, and H. Tang, "End-to-end distortion optimized error control for real-time wireless video streaming," in *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on.* IEEE, 2011, pp. 1–5.

[60] Q. Cheng and D. Agrafiotis, "End to end video distortion estimation with advanced error concealment considerations," in *IEEE Visual Communications and Image Processing Conference*, Dec. 2014, pp. 303–306.

[61] Q. Chen and D. Wu, "Delay-rate-distortion model for real-time video communication," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1376–1394, Aug. 2015.

[62] Q. Peng, L. Zhang, X. Wu, and Q. Wang, "Modeling of SSIM-based end-to-end distortion for error-resilient video coding," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 45, 2014.

[63] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment using structural distortion measurement," in *IEEE International Conference on Image Processing*, vol. 3, 2002, pp. 5–68.

[64] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[65] B. Xu, X. Pan, Y. Zhou, Y. Li, D. Yang, and Z. Chen, "CNN-based rate-distortion modeling for H.265/HEVC," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Dec. 2017, pp. 1–4.

[66] B. Girod and N. Farber, "Feedback-based error control for mobile video transmission," *Proceedings of the IEEE*, vol. 87, no. 10, pp. 1707–1723, Oct. 1999.

[67] Q. Chen, Z. Chen, X. Gu, and C. Wang, "Attention-based adaptive intra refresh for error-prone video transmission," *IEEE Communications Magazine*, vol. 45, no. 1, pp. 52–60, Jan. 2007.

[68] Y. Zhou, W. Xu, and Y. Chen, "A network-aware error-resilient video coding using adaptive intra and reference selection refresh," in *International Symposium on Computer Network and Multimedia Technology (CNMT)*, Jan. 2009, pp. 1–4.

[69] P. Nunes, L. D. Soares, and F. Pereira, "Automatic and adaptive network-aware macroblock intra refresh for error-resilient H.264/AVC video coding," in *16th IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 3073–3076.

[70] S. Moiron and M. Ghanbari, "Enhanced error resiliency for video with cyclic intra-refresh lines," in *2011 18th IEEE International Conference on Image Processing*, Sep. 2011, pp. 3229–3232.

[71] A. R. dela Cruz and R. D. Cajote, "Error resilient joint source-channel adaptive intra-refresh rate for wireless video transmission," in *19th International Conference on Digital Signal Processing*, Aug. 2014, pp. 509–514.

[72] Y. R. Zhou, G. Q. Li, and S. S. Ning, "A new feedback-based intra refresh method for robust video coding," in *International Conference on Computer Science and Applications (CSA)*, Nov. 2015, pp. 218–221.

[73] Z. Wang, Z. Hou, J. Xiao, R. Wang, and R. Zhu, "A hybrid intra refresh and reference selection scheme for mobile video coding," in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 357–362.

[74] P. Nunes, L. D. Soares, and F. Pereira, "Error resilient macroblock rate control for H.264/AVC video coding," in *15th IEEE International Conference on Image Processing*, Oct. 2008, pp. 2132–2135.

[75] B. Li, T. Nanjundaswamy, and K. Rose, "An error-resilient video coding framework with soft reset and end-to-end distortion optimization," in *IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 1910–1914.

[76] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[77] V. A. Nguyen, Z. Chen, and Y. P. Tan, "Perceptually optimized error resilient transcoding using attention-based intra refresh," in *Proceedings of IEEE International Symposium on Circuits and Systems*, May 2010, pp. 4217–4220.

[78] H.-J. Chiou, Y.-R. Lee, and C.-W. Lin, "Content-aware error-resilient transcoding using prioritized intra-refresh for video streaming," *Journal of Visual Communication and Image Representation*, vol. 16, pp. 563–588, Aug. 2005.

[79] H. M. Maung, S. Aramvith, and Y. Miyanaga, “Region-of-interest based error resilient method for HEVC video transmission,” in *15th International Symposium on Communications and Information Technologies (ISCIT)*, Oct. 2015, pp. 241–244.

[80] H. Maung, S. Aramvith, and Y. Miyanaga, “Improved region-of-interest based rate control for error resilient HEVC framework,” in *IEEE International Conference on Digital Signal Processing (DSP)*, Oct. 2016, pp. 286–290.

[81] U. Celikcan and E. Tuncel, “Bimodal leaky prediction for error resilient video streaming,” in *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, Nov. 2009, pp. 583–587.

[82] Q. Chen, X. Yang, L. Song, and W. Zhang, “Robust video region-of-interest coding based on leaky prediction,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 9, pp. 1389–1394, Sept 2009.

[83] H.-C. Huang, C.-N. Wang, and T. Chiang, “A robust fine granularity scalability using trellis-based predictive leak,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 372–385, Jun. 2002.

[84] D. Connor, “Techniques for reducing the visibility of transmission errors in digitally encoded video signals,” *IEEE Transactions on Communications*, vol. 21, no. 6, pp. 695–706, Jun. 1973.

[85] H. Yang and K. Rose, “Optimizing motion compensated prediction for error resilient video coding,” *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 108–118, Jan. 2010.

[86] J. M. Boyce, “Weighted prediction in the H.264/MPEG AVC video coding standard,” in *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512)*, vol. 3, May 2004, pp. III–789–92 Vol.3.

[87] D. Liu, L. Wang, C. Li, Y. Hao, and F. Yin, “Error protection with extended dual frame motion compensation,” in *IEEE International Symposium on Multimedia (ISM)*, Dec. 2015, pp. 331–334.

[88] A. Leontaris and P. Cosman, “Video compression for lossy packet networks with mode switching and a dual-frame buffer,” *Image Processing, IEEE Transactions on*, vol. 13, no. 7, pp. 885–897, Jul. 2004.

[89] Y.-L. Chan, H.-K. Cheung, and W.-C. Siu, “Compressed-domain techniques for error-resilient video transcoding using RPS,” *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 357–370, Feb. 2009.

[90] W. Tu and E. Steinbach, “Proxy-based reference picture selection for error resilient conversational video in mobile networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 2, pp. 151–164, Feb. 2009.

[91] H. Hadizadeh and I. Bajic, “Rate-distortion optimized pixel-based motion vector concatenation for reference picture selection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 8, pp. 1139–1151, Aug. 2011.

[92] V. Chellappa, P. Cosman, and G. Voelker, “Error concealment for dual frame video coding with uneven quality,” in *Proceedings of Data Compression Conference*, Mar. 2005, pp. 319–328.

[93] Y. J. Liang, E. Setton, and B. Girod, “Channel-adaptive video streaming using packet path diversity and rate-distortion optimized reference picture selection,” in *IEEE Workshop on Multimedia Signal Processing*, Dec. 2002, pp. 420–423.

[94] Q. Qu, Y. Pei, and J. W. Modestino, “An adaptive motion-based unequal error protection approach for real-time video transport over wireless IP networks,” *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 1033–1044, Oct. 2006.

[95] P. Lambert, W. D. Neve, Y. Dhondt, and R. V. de Walle, “Flexible macroblock ordering in H.264/AVC,” *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 358–375, 2006.

[96] S. K. Im and A. J. Pearmain, “Error resilient video coding with priority data classification using H.264 flexible macroblock ordering,” *IET Image Processing*, vol. 1, no. 2, pp. 197–204, June 2007.

[97] K. Tan and A. Pearmain, “An improved FMO slice grouping method for error resilience in H.264/AVC,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 1442–1445.

[98] ——, “A new error resilience scheme based on FMO and error concealment in H.264/AVC,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 1057–1060.

[99] J. Panyavaraporn and R. D. Cajote, “Flexible macroblock ordering based on region of interest for H.264/AVC wireless video transmission,” in *19th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Apr. 2012, pp. 384–387.

[100] H. D. Pham and S. Vafi, “Motion-energy-based unequal error protection for H.264/AVC video bitstreams,” *Signal, Image and Video Processing*, vol. 9, no. 8, pp. 1759–1766, Nov. 2015.

[101] B. Katz, S. Greenberg, N. Yarkoni, N. Blaunstien, and R. Giladi, “New error-resilient scheme based on FMO and dynamic redundant slices allocation for wireless video transmission,” *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 308–319, Mar. 2007.

[102] Y. Wang, M. O’Neill, and F. Kurugollu, “New FMO type to flag ROI in H.264/AVC,” in *5th European Workshop on Visual Information Processing (EUVIP)*, Dec. 2014, pp. 1–5.

[103] X. Yang, C. Zhu, Z. G. Li, X. Lin, and N. Ling, “An unequal packet loss resilience scheme for video over the internet,” *IEEE Transactions on Multimedia*, vol. 7, no. 4, pp. 753–765, Aug. 2005.

[104] T. Zhang and Y. Xu, “Unequal packet loss protection for layered video transmission,” *IEEE Transactions on Broadcasting*, vol. 45, no. 2, pp. 243–252, Jun. 1999.

[105] P. Perez and N. Garcia, “Lightweight multimedia packet prioritization model for unequal error protection,” *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, pp. 132–138, Feb. 2011.

[106] H. D. Pham and S. Vafi, “An adaptive unequal error protection based on motion energy of H.264/AVC video frames,” in *IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2013, pp. 4594–4599.

[107] H. D. Pham and V. Sina, “Unequal error protection of H.264/AVC video bitstreams based on data partitioning and motion information of slices,” in *IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, Aug. 2012, pp. 634–639.

[108] H. D. Pham and S. Vafi, “Unequal error protection of H.264/AVC video bitstreams based on the motion energy estimation for group of pictures,” in *7th*

International Conference on Signal Processing and Communication Systems (ICSPCS), Dec. 2013, pp. 1–6.

[109] Y. Zhang, S. Qin, B. Li, and Z. He, “Rate-distortion optimized unequal loss protection for video transmission over packet erasure channels,” *Signal Processing: Image Communication*, vol. 28, no. 10, pp. 1390–1404, 2013.

[110] Y. Zhang, S. Qin, and Z. He, “Fine-granularity transmission distortion modelling for video packet scheduling over mesh networks,” *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 1–12, Jan. 2010.

[111] H. Ha, J. Park, S. Lee, and A. C. Bovik, “Perceptually unequal packet loss protection by weighting saliency and error propagation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 9, pp. 1187–1199, Sep. 2010.

[112] H. Ha and C. Yim, “Layer-weighted unequal error protection for scalable video coding extension of H.264/AVC,” *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 736–744, May 2008.

[113] E. Maani and A. K. Katsaggelos, “Unequal error protection for robust streaming of scalable video over packet lossy networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 3, pp. 407–416, Mar. 2010.

[114] Y. Huo, M. El-Hajjar, R. G. Maunder, and L. Hanzo, “Layered wireless video relying on minimum-distortion inter-layer fec coding,” *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 697–710, Apr. 2014.

[115] J. Jose and S. M. Sameer, “A new unequal error protection technique for scalable video transmission over multimedia wireless networks,” in *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Feb. 2015, pp. 1–5.

[116] R. Perera, A. Fernando, T. Mallikarachchi, H. K. Arachchi, and M. Pourazad, “Qoe aware resource allocation for video communications over LTE based mobile networks,” in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, Aug. 2014, pp. 63–69.

[117] R. Perera, A. Fernando, H. K. Arachchi, and M. A. Imran, “Adaptive modulation and coding based error resilience for transmission of compressed video,”

in *International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug. 2015, pp. 1127–1132.

[118] P. Tovstogan and H. F. Hsiao, "Video streaming optimization using degradation estimation with unequal error protection," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2017, pp. 1–4.

[119] X. Song, X. Peng, J. Xu, G. Shi, and F. Wu, "Unequal error protection for scalable video storage in the cloud," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 699–710, Mar. 2018.

[120] C. Zhu, Y.-K. Wang, M. Hannuksela, and H. Li, "Error resilient video coding using redundant pictures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 3–14, Jan. 2009.

[121] D. Chen, N. Gadgil, and E. J. Delp, "VPx video coding for lossy transmission channels using error resilience packets," in *2016 Picture Coding Symposium (PCS)*, Dec 2016, pp. 1–5.

[122] M. Dissanayake, C. T. E. R. Hewage, S. Worrall, W. A. C. Fernando, and A. Kondoz, "Redundant motion vectors for improved error resilience in H.264/AVC coded video," in *IEEE International Conference on Multimedia and Expo*, Jun. 2008, pp. 25–28.

[123] N. Gadgil and E. J. Delp, "VPx error resilient video coding using duplicated prediction information," *Electronic Imaging*, vol. 2016, no. 2, pp. 1–6, 2016.

[124] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–70, Jan. 2005.

[125] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 821–834, May 1993.

[126] I. Radulovic, P. Frossard, Y.-K. Wang, M. Hannuksela, and A. Hallapuro, "Multiple description video coding with H.264/AVC redundant pictures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 1, pp. 144–148, Jan. 2010.

[127] W.-J. Tsai and H.-Y. You, "Multiple description video coding based on hierarchical b pictures using unequal redundancy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 2, pp. 309–320, Feb. 2012.

[128] P. Correia, P. Assuncao, and V. Silva, “Multiple description of coded video for path diversity streaming adaptation,” *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 923–935, Jun. 2012.

[129] C. Lin, T. Tillo, Y. Zhao, and B. Jeon, “Multiple description coding for H.264/AVC with redundancy allocation at macro block level,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 589–600, May 2011.

[130] Y. Wang and Q.-F. Zhu, “Error control and concealment for video communication: a review,” *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, May 1998.

[131] P. Salama, N. Shroff, E. Coyle, and E. Delp, “Error concealment techniques for encoded video streams,” in *International Conference on Image Processing*, vol. 1, Oct. 1995, pp. 9–12.

[132] S. Valente, C. Dufour, F. Groliere, and D. Snook, “An efficient error concealment implementation for MPEG-4 video streams,” *IEEE Transactions on Consumer Electronics*, vol. 47, no. 3, pp. 568–578, Aug. 2001.

[133] J. W. Woods, *Multidimensional Signal, Image, and Video Processing and Coding*, 2nd ed. Orlando, FL, USA: Academic Press, Inc., 2012.

[134] R. Bernardini, L. Celetto, G. Gennari, M. Petrani, and R. Rinaldo, “Error concealment of INTRA coded video frames,” in *International Workshop on Image Analysis for Multimedia Interactive Services*, Apr. 2010, pp. 1–4.

[135] W. Kim, J. Koo, and J. Jeong, “Fine directional interpolation for spatial error concealment,” *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 1050–1056, Aug. 2006.

[136] S.-C. Hsia, “An edge-oriented spatial interpolation for consecutive block error concealment,” *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 577–580, Jun. 2004.

[137] J. Chen, J. Liu, X. Wang, and G. Chen, “Modified edge-oriented spatial interpolation for consecutive blocks error concealment,” in *IEEE International Conference on Image Processing,*, vol. 3, Sep. 2005, pp. 904–913.

[138] H. Gharavi and S. Gao, "Spatial interpolation algorithm for error concealment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008, pp. 1153–1156.

[139] H. Senel, "Gradient estimation using wide support operators," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 867–878, Apr. 2009.

[140] S. K. Bandyopadhyay, Z. Wu, P. Pandit, and J. M. Boyce, "An error concealment scheme for entire frame losses for H.264/AVC," in *Sarnoff Symposium, IEEE*, Mar. 2006, pp. 1–4.

[141] H. Liu, D. Wang, W. Li, and O. Issa, "New method for concealing entirely lost frames in H.264 video transmission over wireless networks," in *IEEE International Symposium on Consumer Electronics (ISCE)*, Jun. 2011, pp. 112–116.

[142] Y. Chen, Y. Hu, O. Au, H. Li, and C. W. Chen, "Video error concealment using spatio-temporal boundary matching and partial differential equation," *IEEE Transactions on Multimedia*, vol. 10, no. 1, pp. 2–15, Jan. 2008.

[143] S. Garg and S. Merchant, "Interpolated candidate motion vectors for boundary matching error concealment technique in video," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 53, no. 10, pp. 1039–1043, Oct. 2006.

[144] T. Thaipanich, P.-H. Wu, and C.-C. Kuo, "Low-complexity video error concealment for mobile applications using OBMA," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 753–761, May 2008.

[145] J. Zhang, J. Arnold, and M. Frater, "A cell-loss concealment technique for MPEG-2 coded video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 4, pp. 659–665, Jun. 2000.

[146] Q. Peng, T. Yang, and C. Zhu, "Block-based temporal error concealment for video packet using motion vector extrapolation," in *IEEE International Conference on Communications, Circuits and Systems and West Sino Expositions*, vol. 1, Jul. 2002, pp. 10–14.

[147] H. Liu, W. Li, and O. Issa, "New algorithm for motion vector extrapolation for concealing entire frame loss in the H.264 video receiver," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2011, pp. 525–526.

[148] Y. Chen, K. Yu, J. Li, and S. Li, "An error concealment algorithm for entire frame loss in video transmission," in *Picture Coding Symposium (PCS)*, 2004.

[149] K. Song, T. Chung, Y. Kim, Y. Oh, and C.-S. Kim, "Error concealment of H.264/AVC video frames for mobile video broadcasting," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 704–711, May 2007.

[150] S. Belfiore, M. Grangetto, E. Magli, and G. Olmo, "An error concealment algorithm for streaming video," in *International Conference on Image Processing*, vol. 3, Sep. 2003, pp. 49–52.

[151] B. Yan and H. Gharavi, "A hybrid frame concealment algorithm for H.264/AVC," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 98–107, Jan. 2010.

[152] Y.-L. Chang, Y. Reznik, Z. Chen, and P. Cosman, "Motion compensated error concealment for HEVC based on block-merging and residual energy," in *20th International Packet Video Workshop (PV)*, Dec. 2013, pp. 1–6.

[153] C. Yeo, W. T. Tan, and D. Mukherjee, "Receiver error concealment using acknowledge preview (RECAP) - an approach to resilient video streaming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 785–788.

[154] H. Hadizadeh, I. V. Baji, and G. Cheung, "Video error concealment using a computation-efficient low saliency prior," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2099–2113, Dec. 2013.

[155] J. Y. Pyun and H. J. Choi, "Error concealment aware error resilient video coding over wireless burst-packet-loss network," in *5th IEEE Consumer Communications and Networking Conference*, Jan. 2008, pp. 824–828.

[156] E. S. Ryu, Y. Ryu, H. J. Roh, J. Kim, and B. G. Lee, "Towards robust UHD video streaming systems using scalable high efficiency video coding," in *International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2015, pp. 1356–1361.

[157] K. E. Psannis, "HEVC in wireless environments," *Journal of Real-Time Image Processing*, vol. 12, no. 2, pp. 509–516, Aug. 2016.

[158] B. Li, J. Xu, and H. Li, "Parsing robustness in high efficiency video coding - analysis and improvement," in *IEEE Visual Communications and Image Processing*, Sep. 2011, pp. 1–4.

[159] ITU-T, "Recommendation P.910, Subjective video quality assessment methods for multimedia applications," Apr. 2008.

[160] JCT-VC, "HM 16.2 reference software," Oct. 2014.

[161] F. Bossen, "Common test conditions and software reference configurations, document JCTVC-H1100," JCT-VC, San Jose, USA, Tech. Rep., Feb. 2012.

[162] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[163] G. Kokkonis, K. E. Psannis, M. Roumeliotis, and Y. Ishibashi, "Efficient algorithm for transferring a real-time HEVC stream with haptic data through the internet," *Journal of Real-Time Image Processing*, vol. 12, no. 2, pp. 343–355, Aug. 2016.

[164] E. N. Gilbert, "Capacity of a burst-noise channel," *The Bell System Technical Journal*, vol. 39, no. 5, pp. 1253–1265, Sep. 1960.

[165] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *The Bell System Technical Journal*, vol. 42, no. 5, pp. 1977–1997, Sep. 1963.

[166] G. Hasslinger and O. Hohlfeld, "The gilbert-elliott model for packet loss in real time services on the internet," in *14th GI/ITG Conference - Measurement, Modelling and Evalutation of Computer and Communication Systems*, Mar. 2008, pp. 1–15.

[167] S. Wenger, "Nal unit loss software JCTVC-H0072," JCT-VC, San Jose, USA, Tech. Rep., Feb. 2012.

[168] J.-L. Lin, Y.-W. Chen, Y.-P. Tsai, Y.-W. Huang, and S. Lei, "Motion vector coding techniques for HEVC," in *IEEE 13th International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2011, pp. 1–6.

[169] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves, VCEG contribution VCEG-M33," Austin, Apr. 2001.

[170] P. Howard and J. Vitter, "Arithmetic coding for data compression," *Proceedings of the IEEE*, vol. 82, no. 6, pp. 857–865, Jun. 1994.

[171] S. C. Huang and S. Y. Kuo, "Optimization of hybridized error concealment for H.264," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 499–516, Sep. 2008.

[172] M. C. Hwang, J. H. Kim, D. T. Duong, and S. J. Ko, "Hybrid temporal error concealment methods for block-based compressed video transmission," *IEEE Transactions on Broadcasting*, vol. 54, no. 2, pp. 198–207, Jun. 2008.

[173] K. H. Choi and D. H. Kim, "A downhill simplex approach for HEVC error concealment in wireless IP networks," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2016, pp. 143–146.

[174] J. S. Pérez, E. Meinhardt-Llopis, and G. Facciolo, "TV-L1 optical flow estimation," *Image Processing On Line*, vol. 3, pp. 137–150, Jul. 2013.

[175] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 363–370.

[176] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *IEEE Intenational Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013.

[177] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *Pattern Recognition*, F. A. Hamprecht, C. Schnörr, and B. Jähne, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 214–223.

[178] B. K. Horn and B. G. Schunck, "Determining optical flow: a retrospective," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, Aug. 1981.

[179] G. Farnebäck, "Polynomial expansion for orientation and motion estimation," Ph.D. dissertation, Linköping University Electronic Press, 2002.

[180] H. Knutsson and C.-F. Westin, "Normalized and differential convolution," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* IEEE, 1993, pp. 515–523.

[181] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[182] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[183] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[184] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency inspired full-reference quality metrics for packet-loss-impaired video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 81–88, Mar. 2011.

[185] G. Hall. (2015, Feb.) Pearsons correlation coefficient.

[186] F. L., *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*, 1994.

[187] P. L. Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.

[188] I. Aljarrah and D. Mohammad, "Video content analysis using convolutional neural networks," in *9th International Conference on Information and Communication Systems (ICICS)*, Apr. 2018, pp. 122–126.

[189] M. Xu, T. Li, Z. Wang, X. Deng, R. Yang, and Z. Guan, "Reducing complexity of HEVC: A deep learning approach," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5044–5059, Oct. 2018.

[190] S. Samarasinghe, *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition.* Auerbach publications, 2016.

[191] M. Hagan, H. Demuth, and M. Beale, *Neural network design.* Martin Hagan, 2014.

[192] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[193] S. Haykin, *Neural networks: a comprehensive foundation*, 2nd ed. Prentice Hall PTR, 199.

[194] M. F. Mller, “A scaled conjugate gradient algorithm for fast supervised learning,” *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.