Résolution Spatio-temporelle Adaptative pour un Codage à Faible Complexité des Formats Vidéo Émergents

Glenn Herrou

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

INSA RENNES

THÈSE DE DOCTORAT DE

L'INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE RENNES
COMUE UNIVERSITÉ BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Signal, Image et Vision

Par
Glenn HERROU

Adaptive Spatio-temporal Resolution for Lightweight Coding of Emerging Video Formats

Thèse présentée et soutenue à Rennes, le 26 Novembre 2019
Unité de recherche : IETR - UMR CNRS 6164
Thèse N° 19 ISAR 24 / D19 - 24

Composition du jury :

Président :	Frédéric DUFAUX	Directeur de Recherche, CNRS, Paris
Rapporteurs :	David BULL	Professeur, University of Bristol
	Marco CAGNAZZO	Professeur, Telecom ParisTech
Examinateurs :	Vincent RICORDEL	Maître de Conférences, Université de Nantes
	Jarno VANNE	Professeur Assistant, Tampere University
Directeur :	Luce MORIN	Professeur, INSA de Rennes
Encadrant :	Wassim HAMIDOUCHE	Maître de Conférences, INSA de Rennes

Intitulé de la thèse :

Résolution Spatio-temporelle Adaptative pour un Codage à Faible Complexité des Formats Vidéo Émergents

-

Adaptive Spatio-temporal Resolution for Lightweight Coding of Emerging Video Formats

Glenn HERROU

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

INSA RENNES

En partenariat avec :

b com

This work has been achieved within the Institute of Research and Technology b<>com, dedicated to digital technologies. It has been funded by the French government through the National Research Agency (ANR) Investment referenced ANR-A0-AIRT-07.

Acknowledgements

I would like to thank my PhD supervisors, Wassim Hamidouche and Luce Morin, for their guidance throughout my thesis. They helped me to be more insightful and ambitious with my research. I am particularly grateful to Wassim for his deep involvement in my research, I feel lucky to have been supervised by him and I am convinced that I could not have achieved all this work without his valuable advice. I also feel grateful to my lab and project managers, Jean-Yves Aubié and Danièle Cloatre, for their confidence, support and continuous interest in my work during these three years.

I also would like to thank my thesis committee: David Bull, Marco Cagnazzo, Frédéric Dufaux, Vincent Ricordel and Jarno Vanne for accepting to evaluate my work. Their constructive feedback helped me improve the manuscript and point out new perspectives.

Many thanks to all my co-workers and friends from b<>com, particularly Antonin, David, Nicolas, Pierrick and Aude from my lab, and my fellow PhD candidates Anas, Nour, Mohammed, Charles, François and Fatemeh for all the interesting talks in or outside of work, and also for the mutual support during the challenging but interesting and rewarding endeavor that is a PhD.

Last but not least, I would like to thank my parents Monique and Louis, my sister Anne-Tifen and my brother Maël, without whom I could not have achieved this work. I cannot thank them enough for their invaluable support and encouragement throughout my thesis.

Résumé en Français

1 Introduction

La définition du dernier format vidéo standardisé en date, appelé TV Ultra Haute Définition (UHDTV) [1], a pour but d'améliorer le format standard actuellement déployé, appelé TV Haute Définition (HDTV) [2], via l'ajout de nouvelles caractéristiques, telles que la résolution spatiale 4K, une dynamique étendue, un gamut de couleur plus large et une fréquence image plus importante, au signal vidéo [3, 4]. La définition technique du standard UHDTV est disponible dans la recommandation BT.2020 fournie par l'Union Internationale des Télécommunications (ITU) [1]. Le déploiement du standard UHDTV se faisant en trois phases, la phase courante UHD-1 Phase 2 permet d'augmenter la résolution de la HD (1920x1080 pixels) à la UHD (3840x2160 pixels), ainsi que la fréquence image de 50/60 images par secondes (ips) à 100/120 ips. La quantité de données à traiter avant la transmission de ce nouveau signal à l'utilisateur final est donc multipliée par un facteur 8 par rapport au signal actuel.

Ce nouveau format vidéo standard devrait bientôt être déployé par les diffuseurs traditionnels ainsi que les fournisseurs de services de *streaming* tels que Netflix ou Youtube. Ceux-ci sont en effet poussés par la demande liée à la mise à disposition du grand public par les fabricants de téléviseur d'équipements capables d'afficher ces nouveaux formats. La quantité de données à transiter sur les réseaux internets et mobiles devraient donc exploser dans les années à venir. Cisco, l'un des leaders mondiaux en équipements de gestion de réseaux, prévoit même que la vidéo représentera d'ici 2022 79% du volume total de données transitant sur les réseaux mobiles mondiaux [5].

En plus de cette demande accrue de bande-passante, de nombreux utilisateurs regardent le même contenu vidéo sur une large gamme de systèmes ayant des capacités d'affichage très hétérogènes ainsi que des bande-passantes disponible très variables. Les fournisseurs de contenu doivent donc toujours considérer ces marchés en encodant leurs vidéos dans plusieurs formats différents (résolution spatiale, fréquence image) et à plusieurs débits. Ainsi, une importante contrainte de compatibilité avec les systèmes existants doit être respectée, ce qui

est un vrai défi pour les diffuseurs TV et fournisseurs de services de *streaming*.

High Efficiency Video Coding (HEVC) [6] est la dernière norme de compression vidéo mise au point de façon collaborative par l'ITU, l'Organisation Internationale de Normalisation (ISO) et la Commission Électrotechnique Internationale (IEC) au travers leur groupes respectifs MPEG et VCEG réunis dans un groupe commun appelé JCT-VC. Une extension scalable du standard HEVC, appelé HEVC scalable (SHVC), a également été développée pour prendre en compte des cas d'usage de codage scalable dans lesquels la compatibilité avec les infrastructures existantes et la présence de plusieurs formats vidéos dans le même flux sont nécessaire.

Cette thèse a pour but de développer de nouvelles solutions de codage scalable afin de réduire la complexité offerte par l'état de l'art SHVC. En effet, bien que ce standard soit une solution prometteuse pour résoudre les problèmes de compatibilité rencontrés par les diffuseurs et fournisseurs de contenu lors de l'introduction de nouveaux formats vidéo sur le marché, son architecture très gourmande en termes de calculs atteint ses limites avec l'encodage de vidéos respectant le format UHDTV.

Le travail effectué dans le cadre de cette thèse se focalise donc sur la proposition de nouveaux algorithmes permettant d'effectuer un codage scalable à faible complexité des vidéos ayant un format émergent tel que l'UHDTV, tout en conservant une compatibilité avec les services HDTV existants. Le but de ces solutions est de faciliter le déploiement des services UHDTV pour les fournisseurs de services de *streaming* et les diffuseurs de contenu. En effet, cette thèse s'inscrit dans un contexte collaboratif entre l'équipe Vaader du laboratoire académique IETR et l'institut de recherche b<>com, dont l'objectif est de transférer ses solutions, entre autres, aux industriels du secteur de l'audio-visuel.

2 Contributions

Les contributions de cette thèse sont divisées en trois parties. La première a pour but d'offrir la scalabilité spatiale avec seulement des algorithmes de pré et post-traitement associés à une seule instance d'encodeur HEVC. Pour cela, un nouveau schéma scalable basse complexité est introduit. Celui-ci est basé sur la décomposition du signal vidéo avant l'encodage de telle manière à permettre la scalabilité en utilisant uniquement des outils compris dans le standard HEVC et une étape de reconstruction réalisée après décodage. Deux façons différentes de décomposer le signal sont étudiées dans cette thèse:

- **Décomposition basée polyphase:** cette méthode de décomposition est basée sur l'état de l'art qui consiste à décomposer chaque image de la vidéo d'entrée en quatre sous-images différentes ayant une résolution divisée par deux horizontalement et verticalement. Ces sous-images sont ensuite arrangées séquentiellement pour former une vidéo qui est ensuite amenée à être encodée par une unique instance d'encodeur HEVC. La scalabilité spatiale est atteinte en ne décodant que la partie du bitstream obtenu cor-

respondant aux sous-images souhaitées, et ce en profitant du mécanisme de couches temporelles inhérentes à la structure de codage hiérarchique des images d'une vidéo utilisée par HEVC. Cependant, la décomposition polyphase comporte certains défauts dus à sa simplicité, notamment l'introduction d'un décalage de phase entre les pixels de luma et les pixels de chroma de chaque type de sous-image. Un processus de filtrage est proposé dans cette étude pour compenser ce décalage de phase et ainsi obtenir de meilleurs résultats lors des prédictions inter-images faites par HEVC pour les pixels de chroma. Les résultats expérimentaux montrent que des gains moyens en débit de l'ordre de 3.6% sont obtenus par rapport à la décomposition polyphase de l'état de l'art. Ainsi, comparée à SHVC, la solution proposée atteint une réduction de complexité moyenne de 55% pour un surcoût moyen de débit de l'ordre de 6.6%.

- **Décomposition basée ondelettes:** cette approche remplace la décomposition polyphase du schéma précédent par une décomposition basée sur des transformées en ondelettes. Un modification du schéma d'ondelettes traditionnelle est proposée dans cette thèse pour permettre l'utilisation de la vidéo transformée dans le même schéma de compression scalable basse complexité qu'avec la décomposition polyphase. Plusieurs ondelettes utilisant uniquement des opérations sur les entiers ont été considérées pour leur compatibilité avec un encodage HEVC. Une modification du calcul du coût débit-distortion en fonction du type de sous-bande a également été proposé pour les ondelettes bi-orthogonales afin de prendre en compte la non conservation d'énergie due à leur non-orthogonalité. Les résultats expérimentaux montrent que l'ondelette de Haar offre les meilleurs résultats, avec réduction moyenne de la complexité de 54% par rapport à SHVC pour un surcoût en débit de 1.9%. Cette méthode dépasse donc les performances de la proposition précédente basée sur la décomposition polyphase.

La seconde partie de cette thèse se focalise sur le design d'un schéma scalable à deux couches composé d'un encodeur HEVC standard pour la couche de base et d'une proposition d'encodeur très basse complexité pour la couche d'amélioration. Cet encodeur spécifique pour la couche d'amélioration est basé sur la proposition d'un algorithme d'adaptation locale de la résolution spatiale. La spécification de cet algorithme peut être séparée en deux contributions complémentaires:

- **Résolution spatiale adaptative pour les prédictions inter-couches:** Cette méthode a pour but d'adapter la résolution spatiale des images de la couche d'amélioration à un niveau bloc afin de ne traiter que la quantité minimale de données permettant de correctement représenter les détails des images de la couche d'amélioration. Pour chaque bloc en entrée, l'algorithme d'adaptation de la résolution spatiale proposé détermine ainsi la résolution spatiale critique parmi un ensemble fixe de valeurs possibles. Le choix de résolution spatiale se base sur l'analyse du contenu du bloc ainsi que des

blocs du voisinage disponibles. Ensuite, une fois la résolution critique sélectionnée, la prédiction et le reste des opérations constituant le cœur de l'encodage sont effectuées dans la résolution spatiale choisie. Cette méthode, couplée à la réutilisation des décisions d'encodage de la couche de base, permet d'atteindre une réduction moyenne de la complexité d'encodage de l'ordre de 72% par rapport à SHVC pour un surcoût en débit de 4.8%

- **Extension de l'outil d'adaptation de la résolution spatiale aux prédictions inter-images:** Cette solution permet d'améliorer significativement les performances de la proposition précédente en supprimant l'obligation de signaler la résolution spatiale choisie pour un bloc de la couche d'amélioration lorsque son bloc co-localisé dans la couche de base a bénéficié de l'utilisation de prédictions inter-images. En effet, après une mise à l'échelle appropriée des vecteurs de mouvements de la couche de base, la résolution des blocs de la couche d'amélioration peut être prédite à partir des images précédemment encodées via une solution proposée de dérivation de la résolution spatiale. Les résultats expérimentaux montrent que le surcoût en débit moyen observé tombe à 0.5% avec cette méthode comparée à SHVC, pour des réductions de complexité toujours importantes, respectivement de l'ordre de 96% et 53% pour l'encodeur de la couche d'amélioration seul et les deux couches combinées.

La dernière partie de cette thèse explore l'adaptation de la résolution temporelle pour des contenus à fréquence d'images élevées (HFR), c'est à dire des vidéos ayant une fréquence image égale à 120 ips. Dans ce but, un nouvel algorithme est proposé:

- **Fréquence Image Variable (VFR) :** cette méthode a pour but de réduire la fréquence image d'une vidéo quand les mouvements qu'elle contient ne nécessitent pas 120 ips pour être correctement représentés. Pour atteindre ce but, un modèle VFR, capable de déterminer la fréquence image critique de façon dynamique, choisie parmi l'ensemble fixe de valeurs $\{30ips, 60ips, 120ips\}$, pour chaque groupe de 4 images d'entrée consécutives. Ce modèle est basé sur des algorithmes d'apprentissage automatique, entraînées à réaliser cette tâche spécifique en utilisant un ensemble de données construit à partir de vidéos HFR sélectionnées spécialement pour cette étude. Les vidéos ont ainsi été préalablement annotées avec les fréquences image critiques correspondantes pour chaque groupe de 4 images consécutives. Les résultats expérimentaux montrent que le modèle est capable de générer des vidéos à fréquence image variable ayant une qualité perçue identique à la vidéo HFR originale correspondante, mais avec moins d'images. Concernant la compression vidéo, les vidéos VFR obtenues à partir du modèle proposé permettent de réduire le débit moyen de 4.3% par rapport à l'encodage HEVC de la vidéo originale HFR, en plus de réduire la complexité moyenne de 28%.

- **Résolution spatio-temporelle adaptative pour un codage scalable à faible complexité:** cette méthode à pour but de combiner la solution basée sur l'adaptation de la résolution spatiale et la solution basée sur la fréquence image variable VFR, afin de proposer un schéma global de codage scalable à faible complexité. Ainsi, l'algorithme de détection de fréquence image critique est d'abord appliqué pour adapter la résolution temporelle à un niveau image. La vidéo VFR est ensuite encodée par le codeur scalable qui adapte la résolution spatiale à un niveau bloc, dans la couche d'amélioration. Cette méthode améliore les performances des algorithmes précédemment proposés, offrant ainsi une solution qui surpasse les algorithmes d'encodage scalable de l'état de l'art en termes d'efficacité de codage et de réduction de complexité.

3 Conclusion

Ce document propose d'explorer des approches de codage scalable légères basées sur l'adaptation de la résolution spatio-temporelle, soit à travers des algorithmes de pré-traitement ou des outils de codage intégrés dans des codecs. Le but final était de proposer une méthode de compression scalable basse complexité capable d'encoder efficacement des vidéos 4K HFR tout en conservant une compatibilités avec les infrastructures HDTV existantes.

Pour atteindre ce but, trois contributions principales ont été proposées dans le cadre de cette thèse. La première est basée sur des outils de pré-traitement utilisés pour permettre le codage scalable basse complexité avec un seul encodeur HEVC. La seconde contribution se focalise sur un schéma scalable à deux couches utilisant un encodeur très basse complexité basé sur l'adaptation de la résolution spatiale pour la couche d'amélioration, tout en assurant une compatibilité au travers de l'encodeur HEVC utilisé pour la couche de base. Enfin, la troisième contribution s'intéresse à l'adaptation de la résolution temporelle avec le design d'un algorithme de fréquence image variable, placé avant encodage.

Au travers de ces contributions, le déploiement de services vidéo scalables respectant le standard UHDTV pourrait être considérablement facilité, grâce à la faible contrainte, en termes de calculs, des solutions proposées. De plus, ce travail a mené à la publication de plusieurs articles scientifiques dans des conférences et journaux internationaux ainsi qu'au dépôt de deux brevets.

Table of contents

List of figures

List of tables

Acronyms

AI	All Intra
AMVP	Advanced Motion Vector Prediction
ASR	Adaptive Spatial Resolution
AVC	Advanced Video Coding
BBC	British Broadcasting Corporation
BL	Base Layer
BTC	Basic Test Cell
BVI-HFR	Bristol Vision Institute High Frame-Rate
CABAC	Context-Adaptive Binary Arithmetic Coding
CB	Coding Block
CI	Confidence Interval
CNN	Convolutional Neural Network
CRT	Cathode Ray Tube
CSF	Contrast Sensitivity Function
CTB	Coding Tree Block
CTC	Common Test Conditions
CTU	Coding Tree Unit
CU	Coding Unit
DCT	Discrete Cosine Transform
DCTIF	DCT-based Interpolation Filter
DMOS	Differential Mean Opinion Score
DPB	Decoded Picture Buffer
DPD	Decoded Picture Buffer
DSCQS	Double Stimulus Continuous Qualisty Scale

DST Discrete Sine Transform
DVB Digital Video Broadcasting
DWT Discrete Wavelet Transforms

EBCOT Embedded Block Coding with Optimized Truncation
EL Enhancement Layer
EZW Embedded Zero-tree Wavelet

FD Frame Decimation
FN False Negatives
FP False Positives
fps frames per second

GAN Generative Adversarial Network
GLCM Gray Level Co-occurrence Matrix
GOP Group Of Pictures

HD High Definition
HDR High Dynamic Range
HDTV High Definition TV
HEVC High Efficiency Video Coding
HFR High Frame-Rate
HLS High Level Syntax
HM HEVC test Model
HR High Resolution
HVS Human Visual System

IBC International Broadcasting Convention
IEC International Electrotechnical Commission
ILP Inter-Layer Processing
ILR Inter-Layer Reference
ISO International Organization for Standardization
ITU International Telecommunication Union

JCT-VC Joint Collaborative Team on Video Coding
JVET Joint Video Exploration Team

LASTR Locally Adaptive Spatio-Temporal Resolution
LR Low Resolution

MDI Mean Decrease Impurity
ML Machine Learning
MOS Mean Opinion Score
MPEG Motion Picture Expert Group
MSE Mean Squared Error
MV Motion Vector

NAB National Association of Broadcasters

OF Optical Flow
OTT Over-The-Top

PB Prediction Block
POC Picture Order Count
PPS Picture Parameter Set
PSNR Peak Signal-to-Noise Ratio
PU Prediction Unit

QoE Quality of Experience
QP Quantization Parameter

R-D Rate-Distortion
RA Random Access
RC Resolution Component
RDO Rate Distortion Optimization
RDOQ Rate Distortion Optimized Quantization
RF Random Forest
RFE Recursive Feature Elimination
RQT Residual QuadTree

SAD Sum of Absolute Differences
SAO Sample Adaptive Offset
SATD Hadamard transformed SAD
SDR Standard Dynamic Range
SDSCE Simultaneous Double Stimulus for Continuous Evaluation
SDTV Standard Definition TV
SHM SHVC test Model
SHVC Scalable High efficiency Video Coding
SMPTE Society of Moving Pictures and Television Engineers
SR Super Resolution
SSE Sum of Squared Errors
SSIM Structural Similarity Index
SVR Support Vector Regression

TB Transform Block
TL Temporal Layer
TN True Negatives
TP True Positives
TU Transform Unit

UHD Ultra-High Definition
UHDTV Ultra-High Definition TV

VCEG Video Coding Expert Group
VFR Variable Frame-Rate
VLC Variable-Length Coding
VMAF Video Multimethod Assessment Fusion
VQM Video Quality Metric
VVC Versatile Video Coding

WCG Wide Color Gamut

Part I

Introduction

Chapter 1

Introduction

1.1 Context and challenges

The definition of the latest Ultra-High Definition TV (UHDTV) standard [1] aims to increase the user's Quality of Experience (QoE) by improving the currently deployed High Definition TV (HDTV) standard [2] via the introduction of new video signal features such as higher spatial resolution, High Dynamic Range (HDR), wider color gamut and High Frame-Rate (HFR) [3, 4]. Technical definition of the UHDTV signal is available in the BT. 2020 recommendation of the International Telecommunication Union (ITU) [1]. **Fig. 1.1** illustrates the features enabled by the deployment of Digital Video Broadcasting (DVB) Ultra-High Definition (UHD) services in three phases. The UHD-1 Phase 2 specification enables to increase the spatial resolution from High Definition (HD) (1920x1080 pixels) to UHD (3840x2160 pixels) and the video frame-rate from 50/60 frames per second (fps) to 100/120 fps, thus multiplying by a factor 8 the amount of data to be processed before transmission to the end user.

This new video signal standard is expected to soon be deployed by traditional broadcasters and Over-The-Top (OTT) providers, such as Netflix and Youtube, since UHDTV compliant display devices have already been pushed to the consumer market by display manufacturers. The amount of video data transmitted over IP or mobile networks is thus expected to further increase in the upcoming few years due to these new immersive data-heavy video formats. Cisco, through its visual networking index [5], even estimates video traffic to represent 79% of the world total mobile data traffic by the end of 2022.

In addition to this high bandwidth requirement, many users are consuming the same content on a wide variety of devices with heterogeneous video format and network capacities. Content providers thus still have to reach these markets by encoding videos in different formats (spatial resolution, frame-rate etc.) and different bitrates. Therefore, a strong constraint on backward compatibility has to be addressed, which is a challenge for both broadcasters and OTT providers.

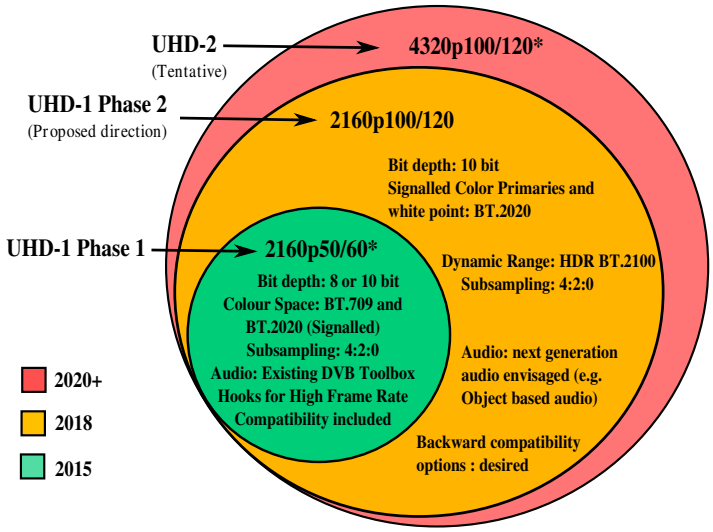


Fig. 1.1 DVB phased UHDTV deployment

High Efficiency Video Coding (HEVC) [6] is the latest video coding standard from the collaboration between the ITU, the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) through their respective Motion Picture Expert Group (MPEG) and Video Coding Expert Group (VCEG) joint effort called Joint Collaborative Team on Video Coding (JCT-VC). A scalable extension of the standard, named SHVC, has also been developed to address scalable encoding use-cases where backward compatibility and several video formats are required.

1.2 Objectives and Motivations

This thesis aims at designing new solutions to decrease the complexity of scalable encoding compared to state-of-the-art codecs. Indeed, although SHVC is a promising solution to address backward compatibility issues encountered by broadcasters and content providers while introducing new video formats to the market, its computationally demanding architecture reaches its limit with the encoding of the data-heavy new immersive video features of the UHDTV video format.

The work of this thesis will thus focus on proposing new efficient ways to achieve a lightweight scalable coding of emerging video formats such as UHDTV while ensuring a backward compatibility with existing HDTV services. The solutions aims at facilitating the deployment of new UHDTV services for broadcasters and OTT providers. Indeed, this thesis is part of a collaborative project between the Vaader team of the IETR academic laboratory and the research institute b<>com, whose objective is to transfer these solutions to the different actors of the audiovisual industry.

1.3 Contributions

The contributions of this thesis are divided into three parts. The first part of this work aims at achieving spatial scalability through only pre/post-processing tools with a single HEVC encoder instance. To this end, a new low-complexity spatially scalable coding scheme is introduced. It is based on a decomposition of the video signal prior to encoding in such a way as to enable spatial scalability using built-in tools of the HEVC standard and a signal reconstruction step performed after decoding. Two different decomposition approaches are investigated in this work:

- **Polyphase-based Decomposition:** This decomposition method is based on the state-of-the-art polyphase decomposition which consists in decomposing each image of the input video into four different sub-resolution images. The newly obtained images are then rearranged sequentially to form a video which is then encoded using a single HEVC encoder instance. Spatial scalability is achieved by only decoding the parts of the obtained layered bitstream containing the desired sub-resolution images. A chroma-phase shift compensation filtering process is proposed to account for a flaw inherent to the polyphase decomposition. A thorough coding performance evaluation of the proposed improvement to the original polyphase decomposition is provided. Several other potential polyphase specific coding tools are investigated and detailed in this work.

- **Wavelet-based Decomposition:** This approach replaces the improved polyphase decomposition by a wavelet-based decomposition. A modification of the conventional wavelet transform scheme is proposed to enable the use of the transformed video within the same low-complexity scalable coding scheme as with the polyphase decomposition. A performance evaluation of the proposed wavelet decomposition with several considered filter banks is detailed in this work. A number of investigated coding tools specific to the format of the wavelet transformed signal are also presented.

The second part of this thesis focuses on the design of a dual-layer scalable encoder composed of a legacy HEVC encoder as Base Layer (BL) encoder and a proposed low-complexity scalable encoder for the coding of the EL. This specific low-complexity EL encoder is based on a proposed algorithm relying on the local adaptation of the spatial resolution. The design of this proposed coding tool can be separated into two complementary contributions:

- **ASR for inter-layer predictions:** This method aims at adapting the spatial resolution of the EL images at a block level to only process the minimal amount of data necessary to correctly portray the spatial details of the EL input images. For each input block, the

proposed ASR algorithm thus determines the critical spatial resolution among a fixed set of possible values based on an analysis of both the block content and its available neighboring blocks. Then, the inter-layer prediction and the rest of the core encoding process are performed at the chosen spatial resolution. This method, together with the re-utilization of BL encoding decisions, allows for a substantial decrease of the encoding complexity compared to state-of-the-art scalable solutions, at the cost of a bitrate overhead due to the additional spatial resolution signaling cost.

- **Extension of the ASR coding scheme to inter-picture predictions:** This method allows for significant improvements of the proposed ASR coding scheme. A first algorithm is proposed, aiming at removing the need to signal the spatial resolution of an EL block when inter-picture predictions are used by the BL encoder for the co-located BL block. To achieve this, the spatial resolution is predicted from the previously coded-frames using the scaled BL Motion Vectors (MVs). A second algorithm is proposed to enable inter-picture predictions, within the EL encoder, to be performed at the resolution chosen by the ASR algorithm. These improvements of the ASR coding scheme lead to a similar coding efficiency for the proposed dual-layer encoder compared to state-of-the-art scalable solutions while dramatically reducing the encoding complexity.

The last part of this thesis investigates the adaptation of the spatio-temporal resolution for HFR content, i.e. videos with a frame-rate of 120 fps. To this end, two novel algorithms are proposed:

- **Variable Frame-Rate:** This method aims at reducing the frame-rate of a video when its inherent motion does not require a 120 fps frame-rate to be portrayed. To achieve this, a Variable Frame-Rate (VFR) model, capable of determining the critical frame-rate, chosen among a fixed number of possible values, for each chunk of 4 consecutive frames of the input video. The model is based on Machine Learning (ML) algorithms trained to perform this specific task using a carefully built dataset of HFR videos annotated with their critical frame-rates. Experimental results shows that the model is capable of generating VFR videos with a perceived quality identical to their corresponding source HFR video but with less frames. In a video coding context, the proposed VFR algorithm achieves significant average bitrate savings and coding complexity reduction compared to state-of-the-art encoding solutions.

- **Locally Adaptive Spatio-Temporal Resolution (LASTR) for low-complexity scalable coding:** This method aims at combining the ASR coding scheme and the VFR solution to propose a lightweight scalable coding scheme based on adaptive spatio-temporal resolution. The VFR model is first used to reduce the frame-rate at a frame-level. The VFR video is then processed by the ASR-based scalable encoder, which

adapts the spatial resolution at a block-level in the EL encoder. This method further increases the performance of the previously proposed algorithms, thus offering an efficient solution that outperforms state-of-the-art scalable encoding algorithms both in terms of coding efficiency and complexity reduction.

1.4 Outline

Chapter 2 includes background information essential to the development of this thesis. First, the different properties of the video signal are presented, in particular the recent improvements brought by the UHDTV signal specifications. Then, the basis of video compression are explained, with a detailed description of the hybrid video coding architecture and a presentation of the common scalability types. Finally, the state-of-the-art standardized video codecs are introduced, with an overview of the coding tools included in both HEVC and its scalable extension SHVC.

Chapter 3 describes the first contribution of this thesis, the improved polyphase-based low-complexity scalable coding scheme. An overview of the related works is first given. Then, the proposed different improvements to the polyphase decomposition are presented, followed by a detailed report of the experimental results.

The second contribution of this thesis, which consists in replacing the previously proposed polyphase approach by a wavelet-based decomposition, is developed in Chapter 4. First, the wavelet transform is briefly introduced together with an overview of related works on wavelet-based coding. Then, the proposed solution is presented together with several additional changes to the HEVC encoding process to account for the wavelet-transformed signal. Finally, the experimental results comparing the performance of the proposed solutions to SHVC.

Chapter 5 introduces the third contribution of this thesis, which proposes a dual-layer low-complexity scalable encoder architecture. First, an overview of related works on downsampling based video coding, upsampling methods and SHVC complexity reduction algorithms is provided. Then, the high-level description of the proposed scalable architecture is presented, followed by a description of the adaptive spatial resolution algorithm, the coding tool on which the low-complexity scalable encoder is based on. A study on resolution mode decisions is also carried out proposing an improved selection process based on content and neighboring blocks. Finally, the experimental results are detailed, in addition to a discussion focusing on the ASR algorithm merits and main identified flaws.

Chapter 6 develops the fourth contribution, extending the ASR algorithm to handle Random Access (RA) configuration and solve the known issues. The algorithm proposed to limit the block-level resolution signaling cost based on a prediction of the resolution mode via motion compensation is first described and evaluated. Then, the addition of inter-picture

predictions within the ASR-based EL encoder is presented and analyzed in terms of coding efficiency and encoding time reduction.

Chapter 7 describes the fifth contribution of this work, which is based on the design of a variable frame-rate algorithm. Related works on HFR video processing are first reviewed followed by a brief introduction to Random Forest (RF) algorithms and a description of the VFR classification problem. Then, the ground truth generation process is described in addition to the development and analysis of the RF training process. Finally, the VFR model output videos perceived quality is subjectively assessed, followed by a description of the performance results.

The last contribution of this thesis, based on the local adaptation of the spatio-temporal resolution, is presented in Chapter 8. The method is first presented, followed by a preliminary performance evaluation.

Chapter 9 concludes this dissertation. The initial goals are first reminded, and the achieved work is then summarized. Finally, future work prospects and potential improvements are discussed.

Finally, Appendix A lists the publications and patents produced during the thesis

Part II

Background

Chapter 2

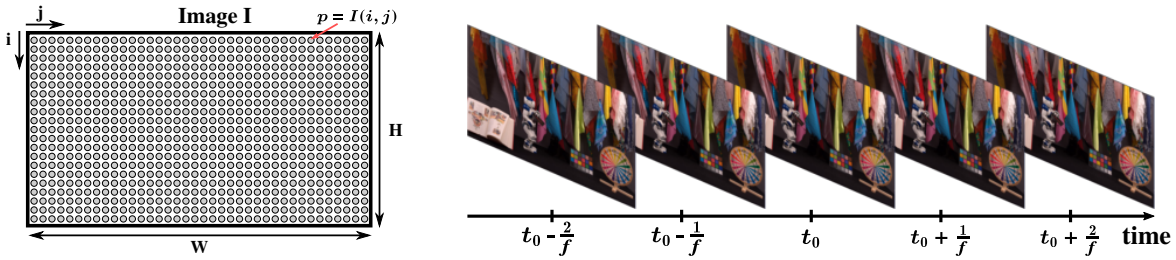
Background on Video Coding

2.1 Video Signal Format

The video signal usually consists on a sequence of frames - also called images or pictures - that are characterized by a series of parameters whose values can be standardized to ensure a correct processing of the video signal. The following sections aim to detail the main characteristics of the video signal and introduce the latest standardized format that will be used in this dissertation.

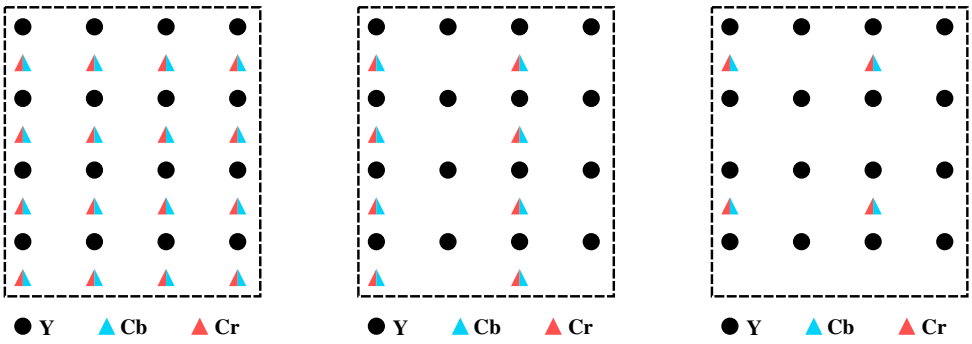
2.1.1 Spatial Resolution

An image is usually represented as a matrix of pixels I whose size $W \times H$ corresponds to the spatial resolution of the image. The image width W is defined as the number of pixels in a row of matrix I while the height H is the number of pixels in each column as depicted in **Fig. 2.1a**. A pixel p, $p = I(i, j)$ with i and j respectively the row and column indexes of the pixel in matrix I, is composed of several components (or channels), depending on the color space under consideration. Typically, either the RGB color space, with one channel



(a) (b)

Fig. 2.1 Representation of (a) an image with a spatial resolution $W \times H$ and (b) a video sequence with a frame-rate f.



(a) 4:4:4 (b) 4:2:2 (c) 4:2:0

Fig. 2.2 Representation of luminance (Y) and chrominance (Cb and Cr) pixel components for common chroma subsampling formats.

per primary color (red, green and blue), or the YUV color space, with one channel for the luminance Y and two for the chrominance U and V, are usually used in digital imaging.

2.1.2 Frame-Rate

In addition to the spatial resolution, the other main characteristic defining a video signal is the frame-rate, usually expressed in fps. A frame-rate f corresponds to the frequency at which the image changes within a video, i.e. an image is displayed for a period of $\frac{1}{f}$ seconds on the display device, as depicted in **Fig. 2.1b**.

2.1.3 Bitdepth and Color Sampling

As previously stated, a pixel is commonly composed of three components or channels. To be stored or transmitted, each of these components is represented, as an integer or in floating point format, using a fixed number of bits, defined as the bitdepth of the video signal. Typical bitdepth values range from 8-bit to 16-bit formats, thus resulting in a total of 24 to 48 bits needed to represent a pixel in RGB color space.

However, in YUV color space, chroma subsampling is used to reduce the number of bits per pixels for a given bitdepth. Indeed, studies on the Human Visual System (HVS) showed a higher sensitivity to luminance compared to chrominance, enabling the possibility of downsampling the chrominance channels without impairing the perceived quality of the video which is one of the main reason of the widespread use of YUV color format in video use-cases.

Typical chroma sampling formats are 4:4:4, the original format with no downsampling, 4:2:2, with a downsampling by 2 in the horizontal direction, and 4:2:0, with a downsampling by 2 in both directions, as depicted in **Fig. 2.2**. Therefore, instead of the 24 bits needed to

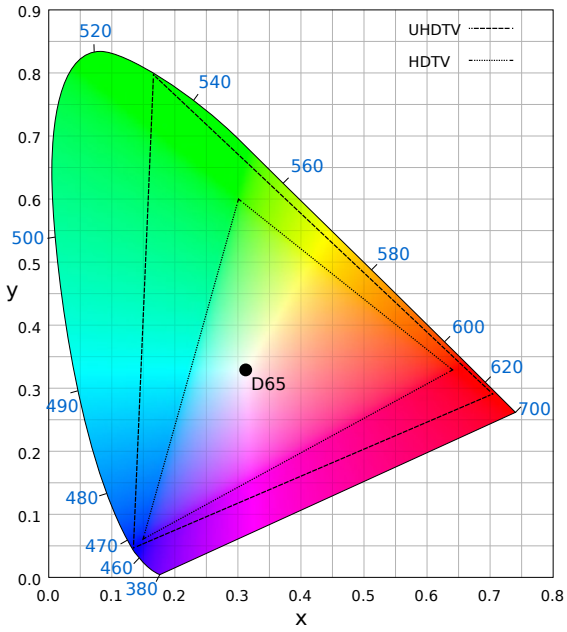


Fig. 2.3 Representation of the Rec. BT.709 (HDTV) and BT.2020 (UHDTV) using the CIE 1931 diagram.

represent a pixel in 8-bit 4:4:4 YUV format, an 8-bit 4:2:0 YUV image will require only 12 bits per pixels to be stored, thus achieving interesting compression performance using only the color space representation.

2.1.4 Color Gamut

Color Gamut is the limitation of the color range that can be represented in an image compared to the range of colors observed in reality. As depicted in **Fig. 2.3**, several color gamuts have been standardized, following the different advances in display technologies and usually associated to a certain television broadcasting standard. For example, ITU Rec. BT.601 [9] defines the color gamut for Standard Definition TV (SDTV) while the color gamut for HDTV has been defined in ITU-R Rec. BT.709 [2]. The wider the color gamut range is, the better the color representation is, thus making Wide Color Gamut (WCG) an important feature to increase the experience of reality in a digital image. To this end, Recommendation BT.2020 [1] defines such a wide color gamut for the UHDTV standard.

2.1.5 Dynamic Range

Natural luminance range from extremely low levels, for example during a moonless night, to extremely high levels for direct sunlight, thus displaying a very high dynamic range, defined as the ratio between the lowest and highest level of luminance. In digital imaging, the dy-

	HDTV	UHDTV
Aspect ratio	16/9	16/9
Spatial resolution	1920x1080	3840x2160, 7680x4320
Frame-rate	60, 60/1.001, 50, 30, 30/1.001, 25, 24, 24/1.001	120, 60, 60/1.001, 50, 30, 30/1.001, 25, 24, 24/1.001
Color gamut	BT.709 color primaries	BT.2020 color primaries
Dynamic range	SDR	SDR, HDR

Table 2.1 Parameter values for HDTV and UHDTV video systems.

namic range is limited both by the capture and display devices, which is an important factor when evaluating the sensation of reality provided by a video system. For a long time, the digital video industry, and more specifically the broadcast industry, has been limited to Standard Dynamic Range (SDR), i.e. luminance levels ranging from 0.1 to 100 cd/m^2, which is very far from the HVS capacities ranging from 10^-6 to 10^8 cd/m^2 [10]. In recent years, the new generations of TV sets brought the support of HDR, with supported luminance levels of up to 1000 or 4000 cd/m^2 with contrast ratios of 1 000 000:1, thus greatly improving the user experience.

2.1.6 UHDTV Standard and Scope of the Study

UHDTV, defined in the ITU-R Rec. BT.2020 [1], is the latest industrial standard for the television broadcast of digital video. Several parameters of the video signal format are designed to improve the visual quality perceived by the end user compared to the previous HDTV standard [2], as summarized in Table 2.1. Particularly, the spatial resolution is increased to reach 2160p or 4320p and 120 fps is added to the list possible frame-rates. In addition, the UHDTV standard includes the support of WCG and HDR.

The DVB project [11], an industrial consortium providing technical standards for broadcast services, separates the introduction of the UHD video format into three different phases UHD-1 Phase 1, UHD-1 Phase 2 and UHD-2, as depicted in **Fig. 1.1**. In this study, only the increase in spatial resolution and frame-rate is considered. Indeed, HDR/WCG format only requires a small number of additional specific coding tools [12–14] and can also be obtained from the SDR format using efficient real-time conversion tools [15, 16]. It has thus been excluded from this study for simplicity. Therefore, the SDR 2160p 120fps video format with BT.709 color space is the reference format for the rest of this dissertation.

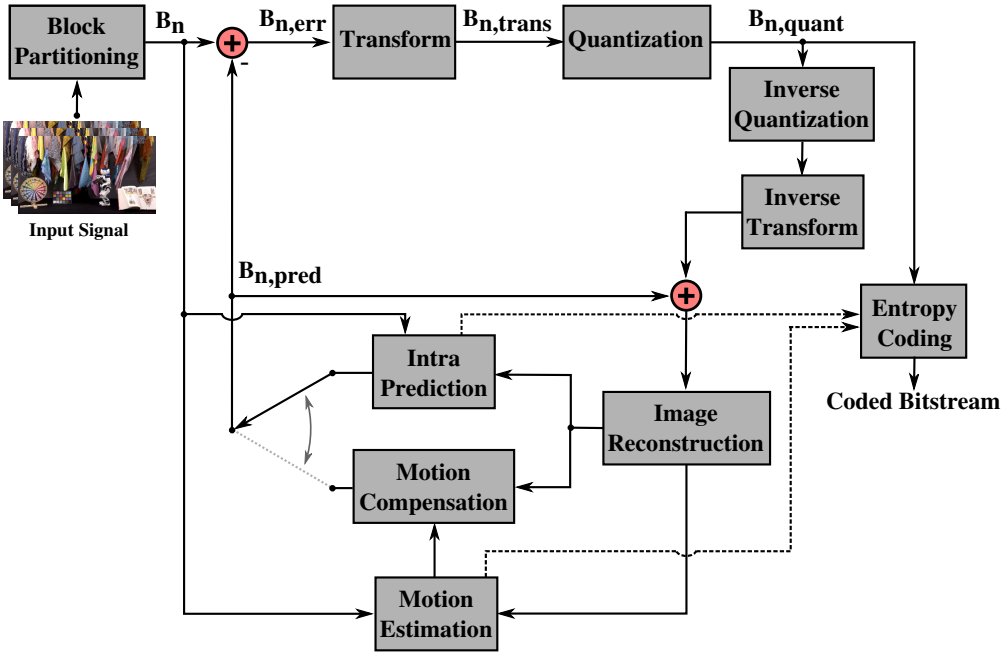


Fig. 2.4 Hybrid encoder architecture.

2.2 Bases of Video Compression

Except for the recently emerging end-to-end deep-learning based video compression algorithms [17, 18], which are out of the scope of this study, and the well-studied wavelet-based video compression theory, introduced in Section 4.1, the most popular video compression solutions rely on block-based hybrid encoding. This section aims at describing this hybrid encoding architecture as well as introducing the concept of scalable encoding. The common performance evaluation method used to compare two coding solutions is also described in this section.

2.2.1 Hybrid Encoder

Hybrid video coding has been used in most of the popular and widely deployed video codecs since the first standardized solution [6, 19–24]. This hybrid architecture is based both on block-based prediction, either using temporal and/or spatial neighboring blocks, and transform of the prediction residuals to achieve the desired compression, as depicted in **Fig. 2.4**. The following sections detail the core coding tools of the hybrid coding scheme, which have been optimized and completed with other algorithms in each new codec generation.

Block Partitioning

Block partitioning is the first step of the block-based hybrid video coding scheme. It consists in dividing the input image I into N blocks B_n of equal size of $W \times H$ pixels, as defined by Equation (2.1), which are then processed sequentially one after the other. In recent codecs, these blocks are further divided in smaller blocks depending on the image content (contours, quantity of details etc.) to achieve a better compression efficiency.

$$\mathbf{B}_n(i,j) = \mathbf{I}(i + \lfloor \tfrac{n}{W} \rfloor,\ j + n\,\%\,W) \qquad \forall (i,j) \in [0,H[\times[0,W[, \tag{2.1}$$

with W and H respectively the width and height of block $\mathbf{B}$, n the index of block $\mathbf{B}$ in image $\mathbf{I}$ in raster scan order, $n = \{1,2,...,N\}$, $\lfloor . \rfloor$ and $.\,\%\,.$ the floor and modulus operators respectively.

Prediction

Each block $\mathbf{B}_n$ is then processed to obtain a predicted block $\mathbf{B}_{n,pred}$ which is used to compute the prediction error residual block $\mathbf{B}_{n,err}$ as follows

$$\mathbf{B}_{n,err}(i,j) = \mathbf{B}_n(i,j) - \mathbf{B}_{n,pred}(i,j) \qquad \forall (i,j) \in [0,H[\times[0,W[. \tag{2.2}$$

The prediction step usually includes a choice between motion estimation and spatial prediction algorithms. Motion estimation consists in finding the block from previously coded images that best matches the current block to be compressed, i.e. the motion compensated prediction block minimizing the error residual energy, and its associated motion vector. Spatial prediction algorithms uses the neighboring pixels from previously coded blocks of the current image being encoded to derive a prediction block that minimizes the error residual energy. Maximizing the prediction accuracy is an important criterion to achieve a good overall compression efficiency, which is the reason behind the highly complex prediction algorithms in recent codecs, as detailed in Section 2.3.2.

Transform

Once the prediction error residual block $\mathbf{B}_{n,err}$ is available, it is expressed in the transform domain before further processing, usually by applying a 2D Discrete Cosine Transform (DCT), to obtain the transformed residual block $\mathbf{B}_{n,trans}$ which decorrelates the residuals and concentrates the residual energy in a smaller number of coefficients. The DCT is used in digital signal processing for its energy compaction property and its highly efficient and optimized

computation algorithms. The transform stage is defined using Equation (2.3)

$$\mathbf{B}_{n,trans}(u,v) = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} \mathbf{C}(u,i)\cdot\mathbf{C}(v,j)\cdot\mathbf{B}_{n,err}(i,j) \qquad \forall (u,v)\in[0,H[\times[0,W[, \tag{2.3}$$

with $\mathbf{C}(x,y)$ the DCT transform matrix elements, whose values are computed as follows for the commonly used orthogonal DCT-II transform of size T

$$\mathbf{C}(x,y) = \frac{P}{\sqrt{T}}\cdot\cos\left[\frac{\pi}{T}(y+\frac{1}{2})x\right] \quad \text{and} \quad P = \begin{cases} 1 & if\ x=0 \\ \sqrt{2} & if\ x>0 \end{cases}.$$

The most recent codecs offer the possibility to choose between different transforms, usually variants of the discrete cosine or sine transforms, to achieve better compression efficiency.

Quantization

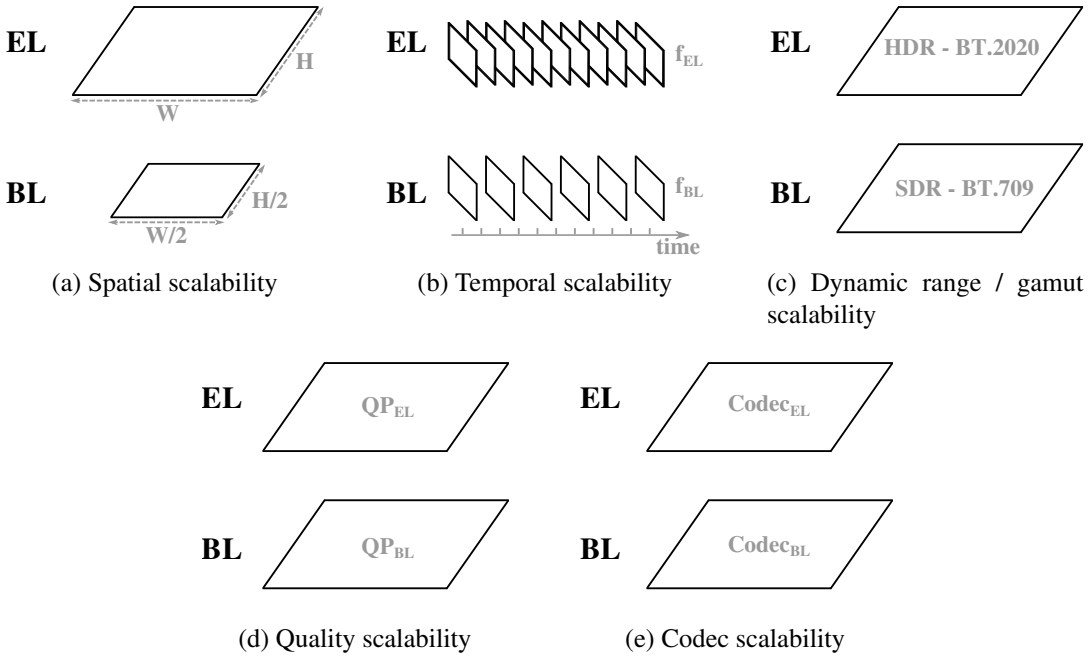
In order to reduce the number of possible real values for the transformed residuals, thus limited the number of bits needed to transmit these coefficients in the bitstream, a quantization step is used. For the uniform quantizer, the most common in video codecs, the quantized residual block $\mathbf{B}_{n,quant}$ is computed as follows

$$\mathbf{B}_{n,quant}(u,v) = \Delta\cdot\left\lfloor\frac{|\mathbf{B}_{n,trans}(u,v)|}{\Delta}+\frac{1}{2}\right\rfloor\cdot sgn(\mathbf{B}_{n,trans}(u,v))\ \forall(u,v)\in[0,H[\times[0,W[, \tag{2.4}$$

with $sgn(x)$ the operator returning the sign of x and Δ the quantization step of the considered uniform quantizer. The quantization step is an encoding parameter that highly influences the resulting bitrate, and thus the reconstruction quality of the decoded pictures.

Entropy Coding

Entropy coding is the last stage of an hybrid video encoder that produces the elements transmitted in the output bitstream. The coding decisions (partitioning, prediction mode) made by the encoder and the quantized transformed coefficients are first binarized to obtain the syntax elements which are then entropy coded using simple variable length coding algorithms or complex context-based arithmetic coding algorithms to achieve better compression efficiency. Indeed, depending on the entropy coding algorithm used, the syntax element will be encoded in the bitstream using more or less bits based on the probability of their value, thus achieving further compression of the data by removing statistical redundancies in the data. The obtained bitrate is highly dependent on the efficiency of the binarization and entropy coding processes, which are detailed in Section 2.3.2 for the recent standardized codecs.



(a) Spatial scalability (b) Temporal scalability (c) Dynamic range / gamut scalability

(d) Quality scalability (e) Codec scalability

Fig. 2.5 Different types of scalability.

2.2.2 Different Types of Scalability

Scalable video coding has first been introduced by the MPEG-2/H.262 video coding standard [21], allowing layered coding with a different configuration and/or video format for each layer. This coding scheme, which can provide significant gains, using inter-layer correlations, over the independent coding of each layer - called simulcast encoding -, has been designed to account for the heterogeneity of the end-users requirements, i.e. from the variable available bandwidth to the different device capabilities. Indeed, depending on the end-user requirement, only the base layer can be transmitted and decoded and the enhancement layer(s) can then be transmitted when the network or the end device become compatible with the higher requirements. In addition, scalable video coding offers an interesting solution for backward compatibility with existing infrastructures when introducing a new video format by using the previously handled format in the BL and the newly introduced one in the EL. In this case, consumers with capable devices can decode both layers while others will only decode the BL, thus avoiding compatibility issues.

To cover these different scalable coding use-cases, several types of scalability, depicted in **Fig. 2.5**, have been considered:

- **Spatial scalability:** the spatial resolution is different in each layer, with usually a factor 1.5 or 2 between the BL and EL resolutions. This is one of the most popular type of scalability due to its adequacy with variable bandwidth and device resolution capability issues.

- **Temporal scalability:** the EL and BL have a different frame-rate, usually the EL frame-rate is a multiple of the BL frame-rate, generally with a factor 2. Most video codecs support temporal scalability due to the Group Of Pictures (GOP) organization and prediction scheme.
- **Dynamic range / gamut scalability:** the dynamic range and/or color gamut of the video signal is different for both layers. This type of scalability is particularly interesting for the deployment of HDR/WCG content with backward compatibility with the common SDR format.
- **Quality scalability:** the EL and BL are encoded with different quality configurations, typically with a different quantization parameter. As for the spatial scalability, this type of scalability is particularly useful to solve varying bandwidth issues while maintaining a minimum base service under critical conditions.
- **Codec scalability:** different codecs are used to encode each layer. The possible inter-layer communications are thus more limited than for other types of scalability, especially if the codecs highly differ from one another. Codec scalability is useful for backward compatibility when deploying a new generation of codecs.

2.2.3 Coding System Performance Evaluation

In order to evaluate the performance of a video coding system P, the video quality of its output compared to the quality of a reference anchor coding system A, either using objective quality metrics or by performing subjective tests. The most basic way to objectively measure the difference between a reference unimpaired image and an image distorted by a coding system is the Mean Squared Error (MSE), computed using Equation (2.5)

$$MSE(\mathbf{I},\mathbf{I}_d) = \frac{1}{W \cdot H} \times \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} (\mathbf{I}(i,j) - \mathbf{I}_d(i,j))^2, \tag{2.5}$$

with W and H respectively the width and height of both the considered reference image $\mathbf{I}$ and distorted image $\mathbf{I}_d$. The widely used Peak Signal-to-Noise Ratio (PSNR) metric, based on MSE and expressed in dB, is usually preferred when evaluating the output quality of a coding system. It is defined as follows

$$PSNR(\mathbf{I},\mathbf{I}_d) = 10 \times \log_{10}\left(\frac{2^b - 1}{MSE(\mathbf{I},\mathbf{I}_d)}\right), \tag{2.6}$$

where b is the bitdepth of both the considered images $\mathbf{I}$ and $\mathbf{I}_d$, i.e. $2^b - 1$ represents the maximum value of each pixel component. For each image of a video, the PSNR is first computed for each pixel component to obtain the weighted PSNR value representing the

overall quality of the image over all color components. The weighted PSNR values are then averaged over all frames to obtain a quality measure for the entire video. For YUV color space, the usual weighting process is defined in Equation (2.7)

$$PSNR_{YUV} = \frac{6 \times PSNR_Y + PSNR_U + PSNR_V}{8}. \tag{2.7}$$

Several more advanced and complex quality metrics, such as Structural Similarity Index (SSIM) [25], Video Quality Metric (VQM) [26] or more recently Video Multimethod Assessment Fusion (VMAF) [27], have been designed to improve the quality evaluation considering certain characteristics of the human visual system. The other evaluation method, through subjective tests, is more reliable to assess the perceptual quality of a video when following the guidelines standardized in the ITU-R BT.500-13 [28] and ITU-T P.910 [29] recommendations. However, these tests require to gather a significant number of participants, which is highly time-consuming, and therefore not always a viable possibility.

Once the evaluated quality measures are available for both coding systems at several configurations (bitrates), a comparison can be performed using the popular Bjøntegaard Delta (BD-Rate) metric [30,31], which can quantify the average bitrate gain for an equivalent measured quality between both coding systems. The BD-Rate metric consists in measuring the average difference between two Rate-Distortion (R-D) curves, one per coding system, each interpolated from at least four bitrate points, as depicted in **Fig. 2.6**.

The first step is thus to interpolate the rate-distortion functions using log-based third-order polynomials, as defined in Equation (2.8)

$$\log(R_X) := \widehat{r}_X(D_X) = a \cdot D_X^3 + b \cdot D_X^2 + c \cdot R_X + d, \tag{2.8}$$

with $\widehat{r}_X$ the interpolated function for coding system X, representing the bitrate R as a function of the distortion D. The distortion is usually the average weighted PSNR but can also be another objective quality metric or MOS obtained through a subjective evaluation.

Then, the BD-Rate value ΔR_{PA} can be computed using Equations (2.9) and (2.10), where the difference between the two interpolated curves is integrated to only take into consideration the overlapping range of measured distortion, shown in gray in **Fig. 2.6**.

$$\Delta\widehat{r}_{PA}(D_L, D_H) = \frac{1}{D_H - D_L} \times \int_{D_L}^{D_H} (\widehat{r}_P(D) - \widehat{r}_A(D))\, dD, \tag{2.9}$$

$$\Delta R_{PA}(D_L, D_H) = 10^{\Delta\widehat{r}_{PA}(D_L, D_H)} - 1, \tag{2.10}$$

with $D_L = \max(\min(D_P), \min(D_A))$ and $D_H = \min(\max(D_P), \max(D_A))$ the boundaries of the overlapping range of measured distortion for both the evaluated coding system P and the considered anchor system A on which the BD-Rate value ΔR_{PA} can be computed.

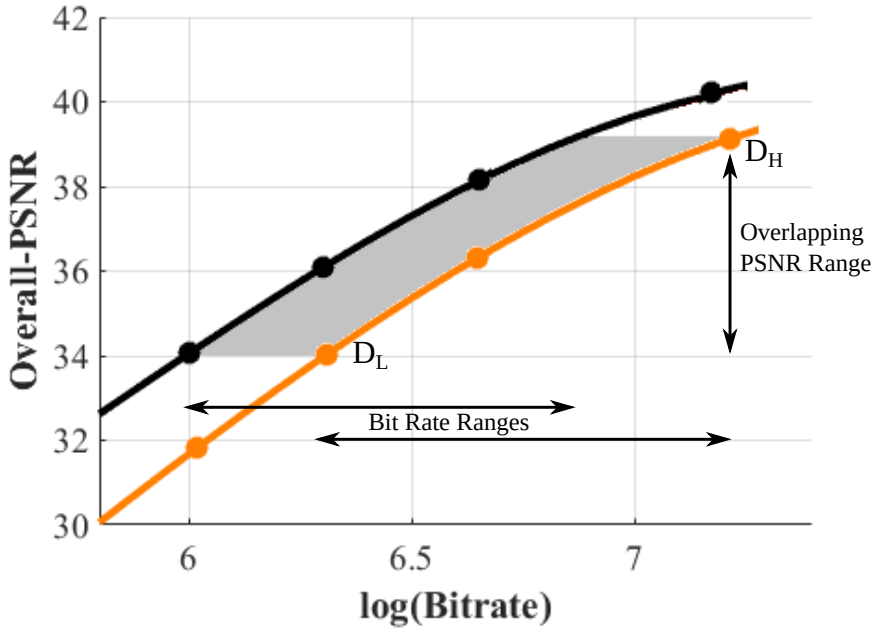


Fig. 2.6 Illustration of Bjøntegaard Delta (BD-Rate) computation.

In addition to the coding efficiency, the complexity of a coding process is can also be measured when evaluating the performance of a coding system. This complexity is usually expressed as a time reduction percentage compared to a reference anchor coding system, as defined in Equation (2.11)

$$TR_{\%}(T_P, T_A) = 100 \times (1 - \frac{T_P}{T_A}) \tag{2.11}$$

with T_X the processing time measured for a coding system X and $TR_{\%}(T_P, T_A)$ the time reduction offered by the coding system P compared to the anchor system A.

Both the BD-rate values and time reduction ($TR_{\%}$) measures are used in this thesis for the evaluation of the proposed algorithm.

2.3 Standardization

In this section, the standardization of video coding systems is addressed. A brief history of video coding standards is first given, followed by a detailed introduction to the different coding standards that fit within the scope of this study, i.e. HEVC and the state-of-the-art scalable video codec SHVC.

2.3.1 Brief History of Video Coding Standards

Video coding aims at reducing the amount of data required to represent a digital video signal, as described in Section 2.1, in either a lossless or lossy manner depending on the desired trade-off between perceived visual degradation and resulting bitrate. A video coding standard defines the different coding tools and standardizes the bitstream format to ensure a straightforward production of encoded videos that can be decoded by any decoder compliant with the worldwide standard.

Over the past few decades, most of the video codecs used in the digital video industry either result from the ITU-T standardization committee, through its VCEG effort, or from the MPEG, a joint effort between the ISO and IEC. Indeed, the ITU-T produced the H.261 (1990) [19] and H.263 (1995) [22] video coding standards while MPEG-1 (1993) [20] and MPEG-4 Visual (1998) [23] have been standardized by the ISO/IEC. However, the most popular and widely used video coding standards have been the product of a collaborative work between VCEG and MPEG, with the H.262/MPEG-2 (1994) [21], H.264/MPEG-4 Advanced Video Coding (AVC) (2003) [24] and H.265/HEVC (2013) [6] standards. Each new codec generation has brought new paradigms and coding tools in addition to reusing and improving the algorithms from the previous codec generation. A new generation video codec, named Versatile Video Coding (VVC), is currently in its final development phase within the Joint Video Exploration Team (JVET), the latest collaboration between VCEG and MPEG. As for the previous video coding standards, VVC is expected to provide a 50% average bitrate gain for an equal perceived quality compared to its predecessor, HEVC.

Since VVC was in its very early development stage at the beginning of this thesis, with a significant encoder complexity increase and without a scalable extension under way, the work presented in this dissertation is solely based on and compared to HEVC and SHVC, its scalable extension, that will both be detailed in the following sections.

2.3.2 HEVC Standard

HEVC has been finalized in 2013 by the JCT-VC, with its first version supporting three different profiles: Main (8-bit video coding), Main10 (10-bit video coding) and Main Still Picture (image coding). Later versions brought support of high bitdepth (12/14/16-bit) encoding and additional chroma subsampling formats (4:2:2 and 4:4:4) as well as two scalable profiles, described in Section 2.3.3, and a multiview and other 3D profiles.

This section focuses on the architecture of the Main profiles, that relies, as depicted in **Fig. 2.7**, on enhanced coding tools from the hybrid coding scheme previously presented, as well as in-loop filters whose roles are to improve the decoded pictures perceived quality before storing them in the Decoded Picture Buffer (DPD). The following sections provides an overview of each of the principal HEVC coding tools, followed by a brief description of

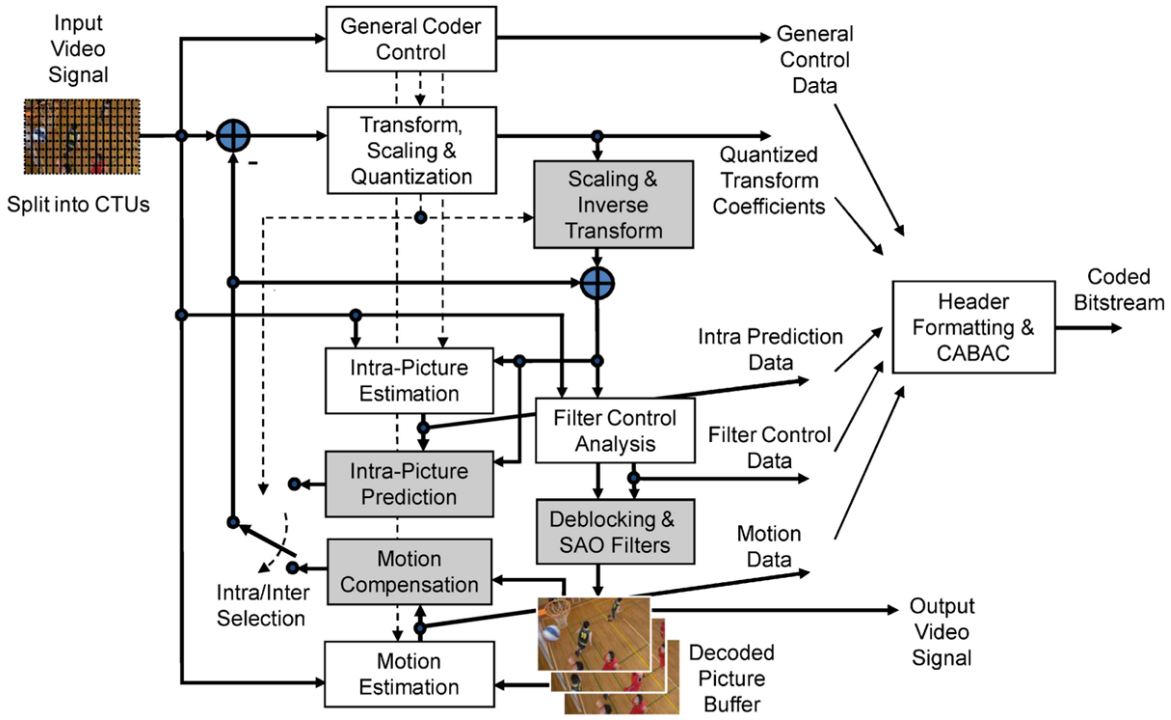


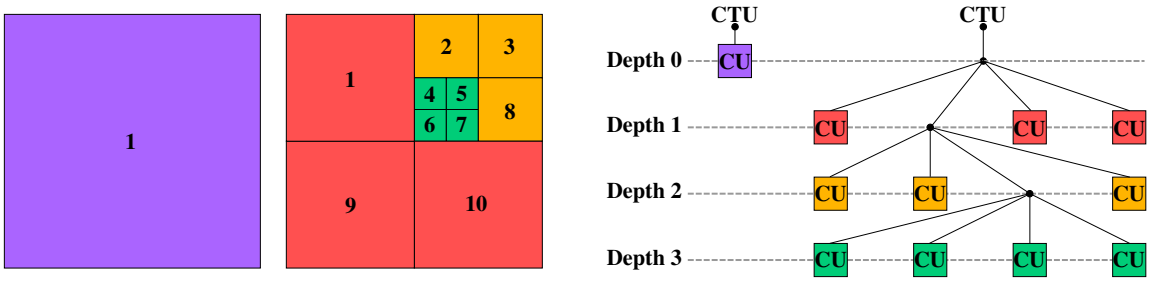
Fig. 2.7 Block diagram of a typical HEVC Encoder [7]

the reference software encoder.

Partitioning

HEVC provides a highly flexible and efficient block partitioning structure to follow both small and large spatial variations of a video signal whose spatial resolution is expected to reach up to 2160p (4K). Indeed, an input video picture is first divided into fixed-length square blocks, called Coding Tree Unit (CTU), which are processed sequentially in raster-scan order. Then, each CTU of size up to 64×64 can be further recursively split into smaller blocks following a flexible quadtree representation [32], as shown in **Fig. 2.8**. The leaf blocks, called Coding Units (CUs), have a minimum size of 8x8 pixels and are processed in descending raster-scan order, as indicated by the ordered CU numbering in **Fig. 2.8a**. Each CU is used as the basis for the prediction and transform/quantization processes, which are respectively performed on subpartitions called Prediction Unit (PU) and Transform Unit (TU). To each of these block units correspond their per-channel (Y, U and V) texture blocks, i.e. Coding Tree Blocks (CTBs), Coding Blocks (CBs), Prediction Blocks (PBs) and Transform Blocks (TBs).

The splitting of a CU into one or multiple PUs can follow the different PU modes depicted in **Fig. 2.9**, depending on the chosen type of prediction for the CU. For intra-predicted blocks, only square PUs are allowed, resulting in a PU of the same size as its corresponding CU, i.e.



(a) CTU Representation (b) Quadtree Representation

Fig. 2.8 Example of HEVC CTU quadtree partitioning.

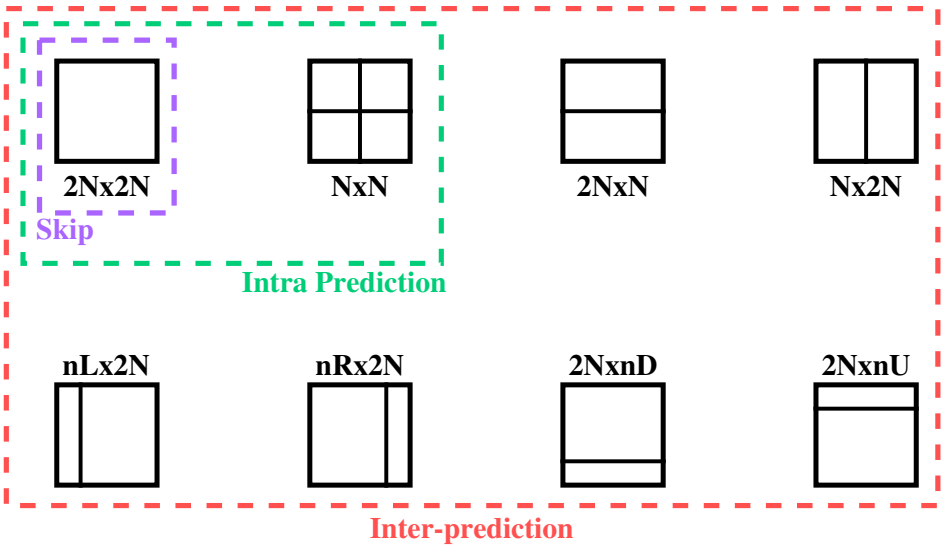
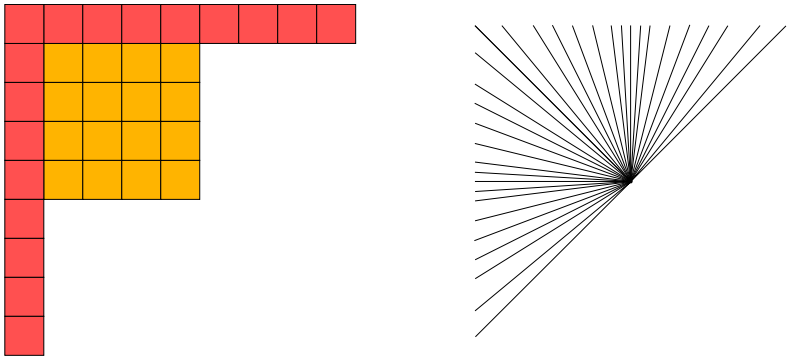


Fig. 2.9 HEVC Prediction Unit (PU) modes.

2Nx2N, or in a splitting of the CU into four PUs of the same size NxN, which is only enabled for the minimum CU size. For inter-predicted blocks, PU modes resulting in two PUs, either by a symmetrical/asymmetrical vertical/horizontal split, are enabled in addition to the NxN and 2Nx2N modes. For the special skipped CUs mode, no splitting is allowed for the prediction stage.

For the transform process, another quadtree representation, called the Residual QuadTree (RQT), is used, with the CU as the root and the TUs as leaves. Several constraints are considered while building the RQT: only square TUs with a minimum TU size of 4x4 pixels are allowed and a TU must not cross PU boundaries.

Each CTU, CU, PU and TU partitioning decision is described through standardized syntax elements to be transmitted in a HEVC bitstream.



(a) (b)

Fig. 2.10 Intra prediction in HEVC with a) the reference and predicted pixels in red and yellow respectively, b) the 33 angular predictors.

Intra Prediction

In HEVC, the current block samples can be predicted from the adjacent reference samples, that have already been processed by the encoder, respectively depicted in yellow and red in **Fig. 2.10a**. Three different types of prediction can be used, namely DC mode, planar mode and angular modes [33]. In DC mode, the average value of available immediate neighboring reference samples are assigned to each pixel of the predicted blocks. Planar mode uses a weighted mean of an horizontal, a vertical and two corner sample predictions, aiming at preserving continuities across block boundaries. For the angular mode, 33 different projection angles can be used in HEVC, as shown in **Fig. 2.10b**, to interpolate the predicted samples from the available reference ones. This complex intra prediction scheme enables significant gains over the previous standards, notably due to the increased number of angular modes which enables a better tracking of edges across adjacent blocks.

Inter Prediction

As most codecs based on an hybrid encoding architecture, HEVC enables the prediction of a block from samples of previously coded reference pictures. The HEVC motion estimation process, which consists in finding the most similar block in the reference pictures, allows for MVs with up to quarter-pixel precision to increase the prediction accuracy [34]. Thus, if a fractional pixel MV is considered, the reference pixels have to be interpolated to the desired sub-pixel position, among the possible positions depicted in **Fig. 2.11**. The HEVC interpolation process uses 8/7-tap and 4-tap 1D filters for the computation of $\frac{1}{4}$ pixel and $\frac{1}{8}$ pixel fractional samples for the luma and chroma channels respectively. The 1D filtering operation, defined in Equation (2.12) is successively performed along the horizontal and vertical directions, in this order, to obtain the interpolated fractional reference block samples. The filter coefficients depend on the interpolation phase used for each filtering operation,

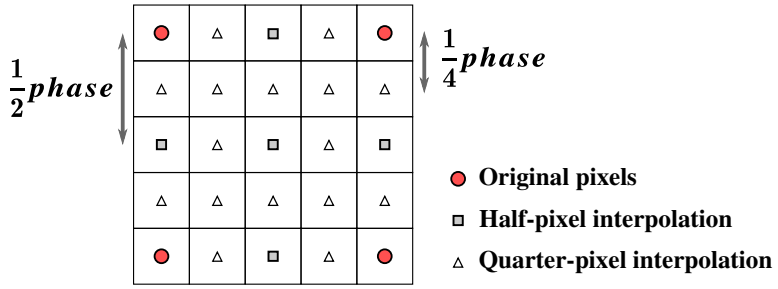


Fig. 2.11 Fractional sample positions for the interpolation of luma pixels.

Table 2.2 Fractional-sample HEVC interpolation filter coefficients.

Channel	phase p	Filter coefficients							
		$\mathbf{f}[p,0]$	$\mathbf{f}[p,1]$	$\mathbf{f}[p,2]$	$\mathbf{f}[p,3]$	$\mathbf{f}[p,4]$	$\mathbf{f}[p,5]$	$\mathbf{f}[p,6]$	$\mathbf{f}[p,7]$
Luma	0	0	0	0	64	0	0	0	0
	1/4	-1	4	-10	58	17	-5	1	0
	2/4	-1	4	-11	40	40	-11	4	-1
	3/4	0	1	-5	17	58	-10	4	-1
Chroma	0	N/A	N/A	0	64	0	0	N/A	N/A
	1/8	N/A	N/A	-2	58	10	-2	N/A	N/A
	2/8	N/A	N/A	-4	54	16	-2	N/A	N/A
	3/8	N/A	N/A	-6	46	28	-4	N/A	N/A
	4/8	N/A	N/A	-4	36	36	-4	N/A	N/A
	5/8	N/A	N/A	-4	28	46	-6	N/A	N/A
	6/8	N/A	N/A	-2	16	54	-4	N/A	N/A
	7/8	N/A	N/A	-2	10	58	-2	N/A	N/A

which can be derived from the desired fractional position, as summarized in Table 2.2.

$$\mathbf{v}_{\mathbf{B}_n,interp}(i) = \frac{1}{\sum_{k=0}^{F-1}\mathbf{f}(p,k)} \cdot \sum_{k=0}^{F-1} \mathbf{f}(p,k) \cdot \mathbf{v}_{\mathbf{B}_n}\left(\frac{i}{2}+k+1-\frac{F}{2}\right), \tag{2.12}$$

with $\mathbf{v}_{\mathbf{B}_n}$ the input vector of block $\mathbf{B}_n$, $\mathbf{v}_{\mathbf{B}_n,interp}(i)$ the output vector interpolated with the filter $\mathbf{f}$ of size F - $F = 8$ for luma and $F = 4$ for chroma -, i the vector sample index and p the phase used for the interpolation, depending on the fractional pixel position.

Once the MVs are selected, several coding tools are included in HEVC to reduce the signaling cost of motion data in the bitstream. On one hand, a complex motion vector prediction algorithm, called Advanced Motion Vector Prediction (AMVP), can be used. It consists in identifying a list of several prediction candidates, then only transmitting the index of the candidate in the list as well as the difference between the MV to be coded and the chosen candidate. The AMVP candidate list is constructed as follows, using the neighboring blocks depicted in **Fig. 2.12**. First, two spatial candidates are selected among the five possible spatial neighboring blocks A_0, A_1, B_0, B_1 and B_2. A temporal candidate is also added to the list,

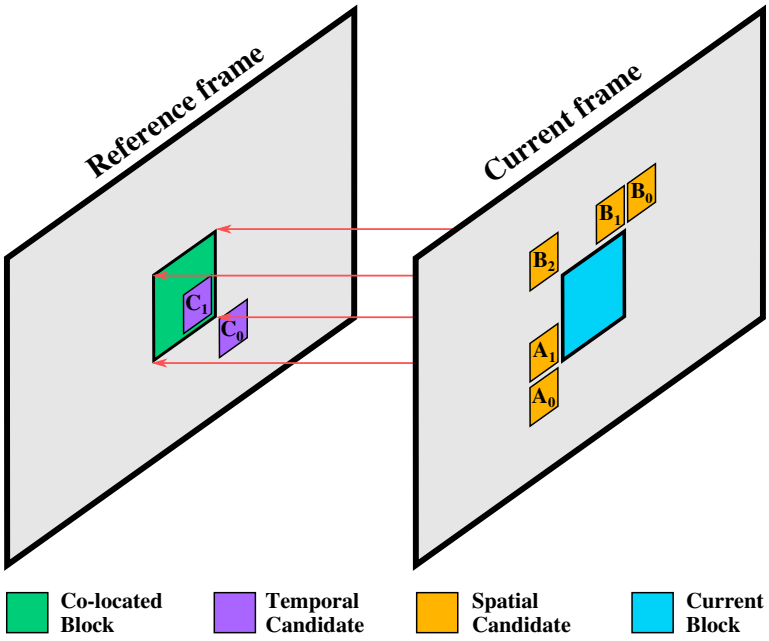


Fig. 2.12 Advanced Motion Vector Prediction (AMVP) candidates

selected between the two possible temporal neighboring blocks C_0 and C_1. If the number of candidates in the list is smaller than two, zero MVs are added to the list.

On the other hand, a new coding tool called Merge-Mode [35] is introduced in HEVC, using the motion data redundancies in a quad-tree structure to reduce the amount of prediction information transmitted in the bitstream. Indeed, merged PUs will completely share their motion data, thus only requiring the signaling of motion data from the first of the merged PUs as well as merge information for each additional PU. The merge information to transmit consists in an index within a merge candidate list. This list has a fixed size of five candidates and is constructed using a similar process as for the AMVP candidate list. The same possible neighboring blocks are considered, with up to four spatial and one temporal candidate selected. However, if the merge candidate list is not complete due to unavailable neighboring blocks, it is first completed with combined motion data from existing candidates or then with zero-motion candidates if the list still contains less than five candidates.

Transform and Quantization

The transform stage of HEVC is similar to its predecessors, with a DCT-II transform kernel used to reduce the number of significant coefficient in the residual signal. The only difference is the use of a Discrete Sine Transform (DST) based transform for 4x4 intra-coded TBs.

The transformed residuals are then quantized and scaled depending on a Quantization Parameter (QP), whose value is in the range 0 to 51. The QP has been designed so that an increase of one represents an increase of the quantization step Δ by ≈ 12 %, i.e. an increase

of 6 in QP leads to a Δ multiplied by a factor of two. Additionally, to completely define the relation between QP and Δ, an absolute step size of one has been associated to a quantization parameter equal to four, leading to the Equation(2.13).

$$\Delta(QP) = \mathbf{g}_{QP\%6} << \frac{QP}{6}, \tag{2.13}$$

where

$$\mathbf{g} = [g_0, g_1, g_2, g_3, g_4, g_5]^T = [2^{-4/6}, 2^{-3/6}, 2^{-2/6}, 2^{-1/6}, 2^0, 2^{1/6}]^T.$$

The quantization operation is defined using Equation (2.14), with $\mathbf{C}(i,j)$, $\mathbf{W}(i,j)$ and $\mathbf{B}_{n,quant}(i,j)$ being respectively the transform matrix coefficient, the quantization weight and the quantizer output at location (i,j) in the TB

$$\mathbf{B}_{n,quant}(i,j) = \left[\left[\left[abs(\mathbf{C}(i,j)) \cdot \frac{16}{\mathbf{W}(i,j)} \cdot \mathbf{f}_{QP\%6} + offset_Q\right] >> \frac{QP}{6}\right] >> s_Q\right] \cdot sgn(\mathbf{C}(i,j)), \tag{2.14}$$

where $\mathbf{f}_{QP\%6} \approx 2^{14}/\mathbf{g}_{QP\%6}$, s_Q is a scaling factor depending on the bitdepth and transform size, and $offset_Q$ is chosen to achieve the desired rounding.

To signal the quantized transformed residuals in the bitstream, HEVC uses new complex but efficient paradigms [36], related to scanning order, significance map, sign coding and coefficient level. A TB is processed by 4x4 sub-blocks, following on of the three available specific scanning orders. Diagonal scan, shown in **Fig. 2.13**, is the most common scanning order in HEVC, available for all TBs. Two additional scanning orders, namely horizontal scan and vertical scan, are available only for 4x4 and 8x8 intra-coded TBs. Based on the 4x4 sub-block separation and the chosen scanning order, the following syntax elements can be used to describe the residuals of a TB:

- ***last_sig_coeff_x***: Horizontal position of the last significant coefficient of a TB in chosen scanning order.
- ***last_sig_coeff_y***: Vertical position of the last significant coefficient of a TB in chosen scanning order.
- ***sig_coeff_flag***: flag indicating if a coefficient is greater than zero. The flag corresponding to the last significant coefficient is inferred to be set to 1.
- ***coeff_abs_level_greater1_flag***: Conditional flag indicating if a significant coefficient is greater than 1. Limited to 8 flags per 4x4 sub-block.
- ***coeff_abs_level_greater2_flag***: Conditional flag indicating if a coefficient greater than 1 is greater than 2. Limited to 1 flag per 4x4 sub-block.

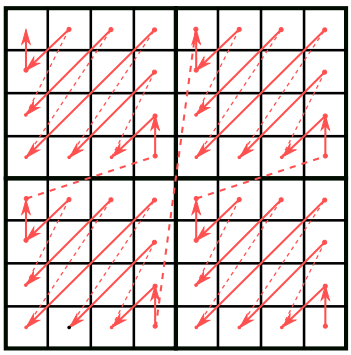


Fig. 2.13 Diagonal scanning pattern used for HEVC residual blocks.

- ***coeff_abs_level_remaining***: Conditional value representing the remaining absolute coefficient level after considering previously signaled flags.

- ***coeff_sign_flag***: Conditional flag indicating the sign of a significant coefficient.

These syntax elements are transmitted for each residual coefficient, depending on the conditional flag dependencies and per sub-block limitations, except for *last_sig_coeff_x* and *last_sig_coeff_y*, signaled at a TB level.

In-loop Filters

Two in-loop filters are integrated in HEVC in order to improve the perceived quality of reconstructed pictures. The HEVC deblocking filter [37] is widely based on the deblocking filter used in AVC. However, it more parallel friendly and has been well optimized in terms of computational cost to limit its impact on the decoding time. The other in-loop filter is a new tool called Sample Adaptive Offset (SAO) [38], which aims at reducing the sample distortion. It consists in first classifying the reconstructed samples into different categories, then obtaining an offset for each category and finally applying the offset to each sample of each category. The offsets are computed at the CTU level only at the encoder side and explicitly transmitted in the bitstream using a set of syntax elements, contrary to the sample classification which is performed both by the encoder and decoder. Both in-loop filters bring 2% average bitrate savings for the same output objective quality.

Syntax Element and CABAC Entropy Coding

Once all syntax elements have been determined by the encoder, HEVC relies on several specific binarization techniques to transpose the syntax elements into bins that can be processed by the entropy coder. Fixed-length or Variable-Length Coding (VLC) entropy coding is used for the signaling of High Level Syntax (HLS) [39] while Context-Adaptive Binary Arith-

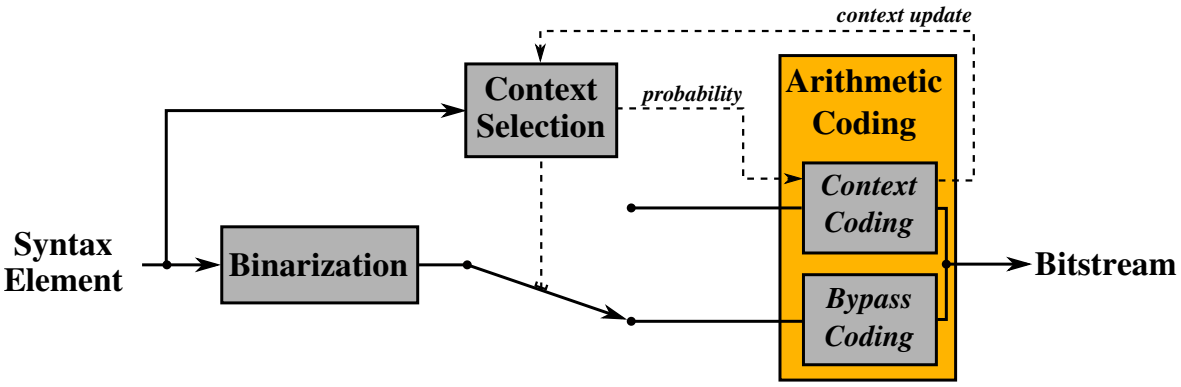


Fig. 2.14 Three key operations in the CABAC engine.

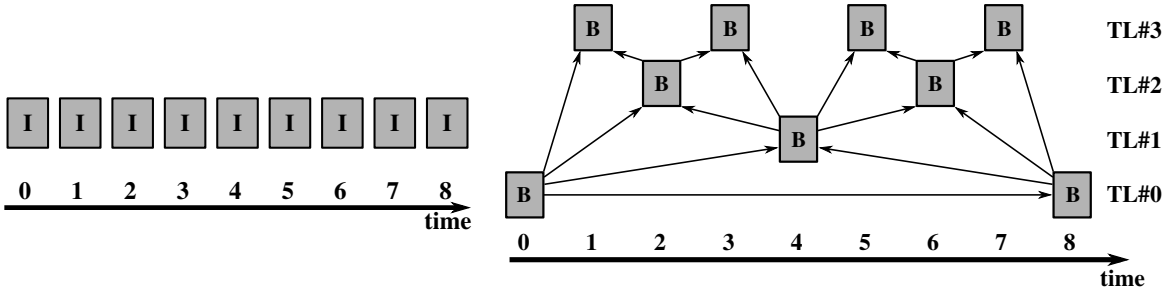
metic Coding (CABAC) [40] is used for all syntax elements related to CUs, PUs, TUs and SAO filtering.

The HEVC CABAC engine architecture is depicted in **Fig. 2.14**. The first step is to binarize the syntax elements to be transmitted in the bitstream. The same binarization techniques as AVC are used, namely unary, truncated unary, k^{th}-order Exp-Golomb and fixed length codes. Almost all syntax elements are binarized using one or several of these techniques. For instance, the residual coding syntax element *coeff_abs_level_remaining* is separated into a prefix and a suffix, respectively binarized with unary coding and fixed-length coding.

Then, a context selection is performed to choose either the context model associated to the considered syntax element or no context modeling. Finally, arithmetic coding is performed on the bins using a probability for each considered bin. This probability is either estimated from the context model, triggering a subsequent context update, or equal to 0.5 if no context has been selected. The latter case is called bypass coding and has been introduced to limit the number of context coded bins, with their complex modeling and feedback update loop, thus increasing the throughput of the HEVC CABAC engine.

Reference Software

The HEVC test Model (HM) reference software [41] is made available by the JCT-VC committee to demonstrate the capacities of the video coding standard. The version used in this thesis is the Test Model 16 (HM16.12) [42], with the two configurations of interest in this dissertation, namely All Intra (AI) and RA, being supported by the reference software. The AI configuration relies only on intra prediction modes for the prediction stage, thus avoiding temporal dependencies between pictures. **Fig. 2.15a** depicts the coding structure of the AI configuration with its independently decodable I pictures. The RA configuration allows for both intra and inter prediction modes, with P (uni-predicted) or B (bi-predicted) pictures organized in GOP of fixed size, as shown in **Fig. 2.15b**. A GOP is composed of hierarchically decodable Temporal Layers (TLs), enabling built-in temporal scalability thanks to the



(a) All Intra (b) Random Access

Fig. 2.15 GOP prediction structure for common HM configurations

possibility to drop one or several of the upper TLs without breaking the temporal prediction dependencies.

Since the purpose of the reference implementation is to demonstrate the capacities of the standard, the HEVC reference software encoder performs an exhaustive search among all possible mode combinations, thus achieving high quality encodings. This efficient exhaustive search is done through a Rate Distortion Optimization (RDO) [43], consisting in evaluating each possible mode combination and then selecting the one minimizing the evaluated R-D cost. The R-D cost J is defined as a linear combination of the distortion D and the bit cost R, weighted by a Lagrangian multiplier λ, as defined in Equation (2.15)

$$J = D + \lambda \cdot R. \tag{2.15}$$

The Lagrange multiplier λ is defined as a function of the QP and its computation has been optimized following the method developed in [44]. The λ computation is also adapted for each possible distortion metric, i.e. either the Sum of Absolute Differences (SAD), the Sum of Squared Errors (SSE) or the Hadamard transformed SAD (SATD) in the reference software, depending on the considered encoding step (partitioning, prediction, etc.).

Starting at the CTU-level with the CU partitioning modes and hierarchically evaluating each subsequent mode combination until in-loop filtering parameters, the RDO-based encoding process provides the bitstream resulting in the best trade-off between distortion and bitrate, driven by the given quantization parameter and derived λ value.

2.3.3 SHVC, the Scalable Extension of HEVC

SHVC [45] has been standardized as annex H of the second version of the HEVC standard. It has been finalized in 2014 by the JCT-VC, enabling the support of spatial scalability, quality scalability, gamut scalability, bit-depth and codec scalability coding schemes in addition to the built-in support of temporal scalability already included in HEVC version 1.

To avoid the mistakes of SVC [46], the scalable extension of AVC, which was not adopted

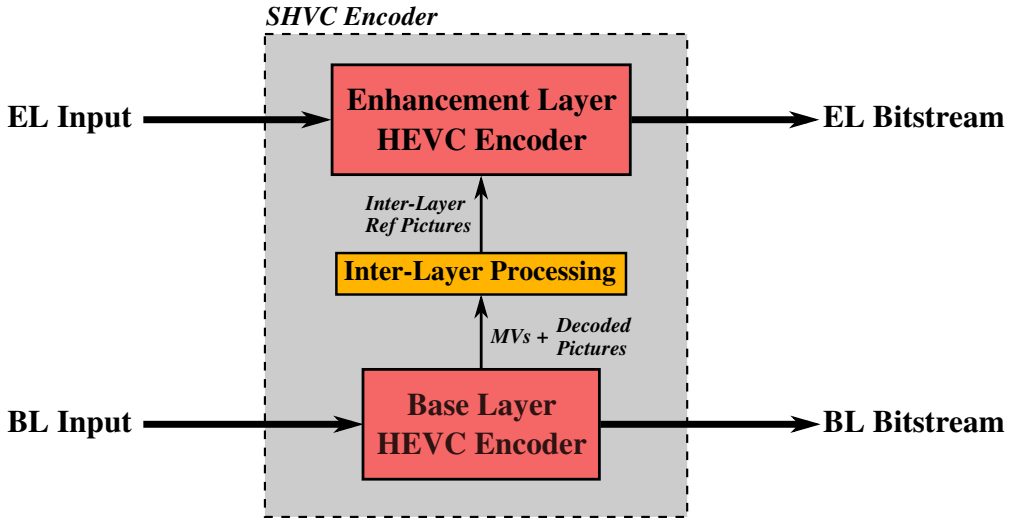


Fig. 2.16 SHVC encoder high lelvel architecture.

by industrials due to the late release and complex architecture of the scalable codec, SHVC has been developed with a focus on simplicity. Indeed, scalability is achieved by coupling one HEVC encoder instance per layer, with HLS changes, and ILP tools, as depicted for a dual-layer scalable coding scheme in **Fig. 2.16**. The high-level syntax changes needed for SHVC are located in the slice header and above, mainly to describe and signal the relevant layer-specific information (number of layer, layer ID, etc.). ILP tools are used to generate the Inter-Layer Reference (ILR) pictures from the BL decoded pictures and adapt the relevant coding information to the format of the EL, depending on the type of scalability used. The ILP tools used for spatial scalability, i.e. the type of scalability of interest in this dissertation, as well as a brief overview of the reference software are presented in the following sections.

Inter-layer Texture Resampling

Due to the difference in spatial resolution between two adjacent layers, a texture resampling [8] is required to upsample the BL decoded pictures to the spatial resolution of the EL in order to generate the ILRs pictures to be used by EL HEVC encoder. SHVC supports a number of different scaling ratios, ranging from 1 to 2, between the BL and EL, the most popular ones being 1.5x and 2x spatial scalability. The resampling is achieved by applying 8-tap and 4-tap interpolation filters to the luma and chroma samples of the BL decoded pictures, respectively. The resampling with a factor 2 reuses the same filter as for the half-pixel interpolation performed in the motion compensation stage, while additional filters have been designed for the 1.5x resampling, as summarized in Table 2.3.

Table 2.3 SHVC ILP upsampling filters [8].

Channel	phase p	Filter coefficients								Note
		$\mathbf{f}[p,0]$	$\mathbf{f}[p,1]$	$\mathbf{f}[p,2]$	$\mathbf{f}[p,3]$	$\mathbf{f}[p,4]$	$\mathbf{f}[p,5]$	$\mathbf{f}[p,6]$	$\mathbf{f}[p,7]$	
Luma	0	0	0	0	64	0	0	0	0	Same as MC filter
	1/3	-1	4	-11	52	26	-8	3	-1	Additional filter
	1/2	-1	4	-11	40	40	-11	4	-1	Same as MC filter
	2/3	-1	3	-8	26	52	-11	4	-1	Additional filter
Chroma	0	N/A	N/A	0	64	0	0	N/A	N/A	Same as MC filter
	1/4	N/A	N/A	-4	54	16	-2	N/A	N/A	Same as MC filter
	1/3	N/A	N/A	-6	52	20	-4	N/A	N/A	Additional filter
	3/8	N/A	N/A	-6	46	28	-4	N/A	N/A	Same as MC filter
	1/2	N/A	N/A	-4	36	36	-4	N/A	N/A	Same as MC filter
	7/12	N/A	N/A	-4	30	42	-4	N/A	N/A	Additional filter
	2/3	N/A	N/A	-2	20	52	-6	N/A	N/A	Additional filter
	7/8	N/A	N/A	-2	10	58	-2	N/A	N/A	Same as MC filter
	11/12	N/A	N/A	0	4	62	-2	N/A	N/A	Additional filter

Inter-Layer Motion Vector Scaling

SHVC allows for the use of MVs from the ILR pictures as temporal candidates in the merge and AMVP candidate list construction algorithms. The temporal motion vector prediction is performed on a compressed motion field, at a 16x16 block level, thus the motion data of the ILR picture is also extracted from the BL compressed motion field.

Therefore , for each 16x16 block in the ILR picture, its colocated 16x16 block is first identified in the BL, taking into account the difference in spatial resolution if spatial scalability is used. Then, the motion information of the ILR block, i.e. the inter-prediction mode, reference pictures and MVs, is copied from its colocated BL block. Finally, the obtained MVs are scaled using the spatial scalability resampling factor to obtain the corresponding correct ILR MVs used for the motion vector prediction process in the EL encoder.

Reference Software

The SHVC test Model (SHM) reference software [47] is largely based on the same implementation as the HM. Indeed, the SHM software instantiates an encoder derived from the HM, for each specific layer, capable of sending, receiving and using inter-layer information from adjacent layers as reference for encoding. The previously described inter-layer processing operations, i.e. texture resampling and MV scaling, are implemented in the SHM within inter-layer processing units communicating with each instance of the specific layer

Table 2.4 SHM downsampling filter coefficients for 1.5x and 2x spatial scalability.

Scaling Factor	phase p	Luma and Chroma channels coefficients											
		$\mathbf{f}[p,0]$	$\mathbf{f}[p,1]$	$\mathbf{f}[p,2]$	$\mathbf{f}[p,3]$	$\mathbf{f}[p,4]$	$\mathbf{f}[p,5]$	$\mathbf{f}[p,6]$	$\mathbf{f}[p,7]$	$\mathbf{f}[p,8]$	$\mathbf{f}[p,9]$	$\mathbf{f}[p,10]$	$\mathbf{f}[p,11]$
1.5x	0	0	5	-6	-10	37	76	37	-10	-6	5	0	0
	1/8	-1	5	-3	-12	29	75	45	-7	-8	5	0	0
	1/2	-1	3	2	-13	8	65	65	8	-13	2	3	-1
	5/8	-1	2	3	-12	2	59	70	14	-13	1	4	-1
2x	0	2	-3	-9	6	39	58	39	6	-9	-3	2	0
	1/4	1	-1	-8	-1	31	57	47	13	-7	-5	1	0

encoders.

In addition, the SHVC reference software implements a downsampling filter bank recommended to generate the BL input video from the EL input original video. Table 2.4 summarizes the downsampling filter coefficients for 1.5x and 2x scalability. They have been designed to minimize the introduced artifacts when coupled with the inter-layer resampling filters. By doing so, the inter-layer prediction error is minimized and the amount of residuals to code is thus reduced.

2.4 Conclusion

In this chapter, the different characteristics of the video signal have first been introduced, in particular the new dimensions of the UHDTV signal. Then, the hybrid encoding architecture, on which most popular video codecs are based on, has been presented followed by a description of the usual scalability types. Finally, a detailed overview of the standardized state-of-the-art encoders, namely HEVC and its scalable extension SHVC, has been given. All this background information on regular and scalable video encoding are essential for understanding the rest of this dissertation.

Part III

Pre/post processing Tools for Low-complexity Spatial Scalability

Chapter 3

Polyphase-based Decomposition

As mentioned in Section 2.3.3, the state-of-the-art scalable solution, SHVC, relies on computationally expensive architecture requiring, in addition to inter-layer processing, one instance of HEVC encoder per layer. This has been a long-time significant impediment to the wide adoption of MPEG scalable video codecs. With the new immersive video formats increasing the amount of data to be processed, the SHVC, and its already computationally demanding architecture [48], becomes even more unsuitable.

With the recent developments proposed within the scope of JVET as a starting point, this chapter proposes an improvement of the a low-complexity scalable coding chain for x2 spatial scalability achieved via purely pre and post processing tools. The contribution of this chapter thus focuses on modifying the polyphase-based decomposition previously presented to JVET within the scope of the future video coding standard development.

This chapter is organized as follows. Section 3.1 provides a detailed description of the polyphase decomposition, the basis of the considered pre-processing tool. Then, the investigated improvements to the low-complexity scalable method are described in Section 3.2, followed by a presentation of the experimental results in Section 3.3. A detailed study on the decisions made by the encoder while processing the polyphase subsampled signal is given in Section 3.4. Finally, Section 3.5 concludes this chapter.

3.1 Related Work

As part of the JVET effort, Thomas *et al.*. proposed a new scalable coding scheme relying on the polyphase subsampling of the input video prior to encoding [49]. This section aims at presenting this new architecture, which enables x2 spatial scalability with a single HEVC encoder instance, instead of the traditional dual-layer scalable approach of SHVC containing on HEVC encoder per layer. The principle of the polyphase decomposition is first detailed, followed by a description of the low-complexity scalable coding scheme, based on this pre-processing tool, is given. Finally, the quality of the polyphase decomposed signal is

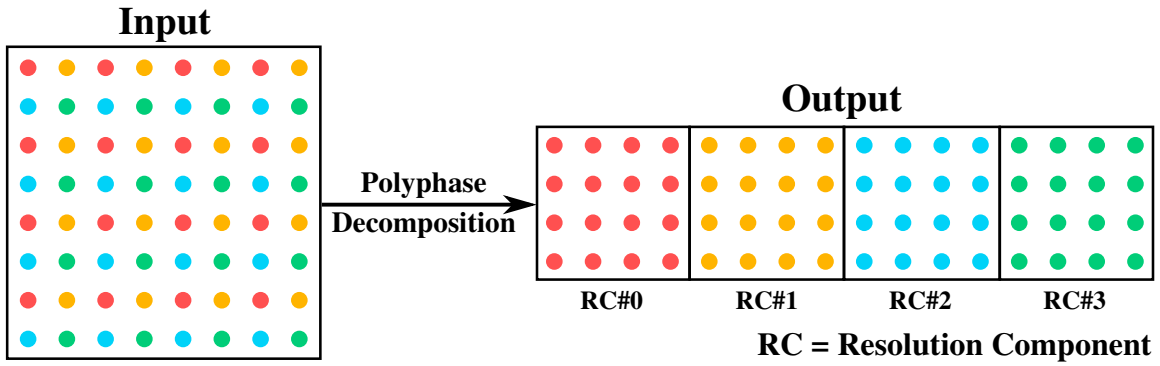


Fig. 3.1 Polyphase subsampling into resolution components (RC).

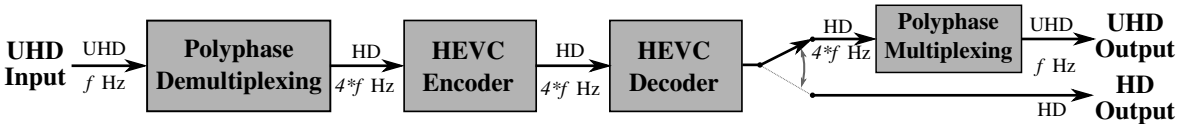


Fig. 3.2 Coding chain for polyphase subsampling scalable scheme.

analyzed.

3.1.1 Principle

The polyphase decomposition is defined as follows. The input video is first decomposed into four sub-resolution components, each Resolution Component (RC) having one over four pixels of each 2x2 block of the original signal. The resulting sub-resolution images are then packed sequentially to form a new video at a quarter of the original spatial resolution but with four times the number of frames, as illustrated in **Fig. 3.1**. The polyphase subsampled video is then fed to an HEVC encoder. Spatial scalability is achieved at the decoder side by choosing to decode only one RC or the entire bitstream. If the four resolution components are decoded, an additional reconstruction step is performed to produce the full resolution decoded video. The coding chain used for polyphase subsampled videos is further detailed in the next section.

3.1.2 Low-complexity Scalable Coding Scheme based on Polyphase Decomposition

Fig. 3.2 illustrates the scalable coding chain used to encode polyphase subsampled videos. The input UHD video is first decomposed into a HD signal following the scheme depicted in **Fig. 3.1**. Then, the HD signal is encoded and decoded using HEVC and finally upsampled to reconstruct the output UHD signal. It is important to note that the signal fed to the encoder

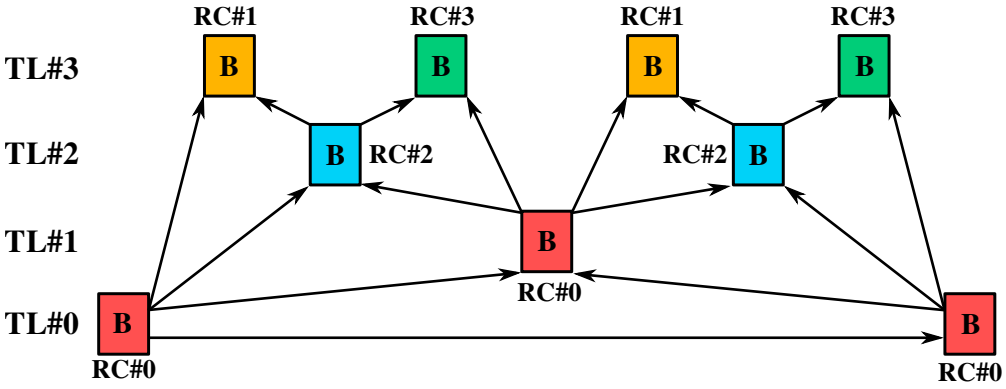


Fig. 3.3 Resolution Components (RC) distribution on temporal layers of HEVC Random Access GOPs.

has an image frequency equal to four times the native one.

The two spatial resolution of the scalable solution can be obtained after decoding, the choice between the HD or UHD output being made during the decoding process. To this end, the temporal layers of the HEVC standard are used to obtain spatial scalability, assuming that the base-layer, i.e. the first RC of the decomposed signal, is occupying the lower temporal layers and that the enhancement layer, i.e. the three other RCs, occupies higher temporal layers. As shown in **Fig. 3.3**, the traditional random access GOP structure of HEVC, in its 8-image or 16-image sizes, guarantees such an organization of the phases in the different temporal layers. Thus, if the desired output is the HD video, the decoder only has to decode the first 2 or 3 temporal layers, for GOP sizes of 8 or 16 images, respectively. For the UHD output, the bitstream has to be decoded entirely, and an additional reconstruction step is necessary to recover the decoded video in its original resolution.

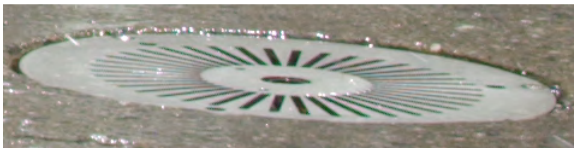
The performance of the polyphase scalable scheme has been assessed in [50, 51], showing promising results compared to simulcast HEVC coding. However, the coding efficiency is greatly varying depending on the type of content, with important losses for highly textured video sequences. This could be due to the different artifacts introduced by the polyphase decomposition, described in the next section, prior to encoding. In addition, it was pointed out that substantial losses can be observed in both chroma channels on most tested sequences, indicating a probable mismanagement of the chroma in the polyphase decomposition scheme.

3.1.3 Quality of the Decomposed Signal

Given the nature of the performed subsampling operation, aliasing artifacts can be introduced, thus possibly impairing the quality of the different RCs. This is especially problematic for the RC serving as BL it can be directly used at the display end.

In [52], Thomas *et al.*. evaluate the artifacts introduced in the base layer of the polyphase subsampled videos of all the different test sequence classes defined by MPEG, from UHD

(a) Original sequence



(b) Patch from original sequence

(c) Patch from polyphase downsampled sequence

Fig. 3.4 Example of aliasing artifacts introduced by polyphase subsampling on ToddlerFountain test sequence

to SD resolutions. It is shown that artifacts introduced are almost negligible for the UHD sequences, whereas the more the resolution decreases, the more disturbing the artifacts are, making the technique unsuitable for resolutions below HD. Focusing on UHD content, the use-case considered in this thesis, several artifacts are still present in the RC output images.

The first source of artifacts is aliasing, as expected with a downsampling operation, which is present in highly textured areas. Overall, on the tested UHD sequences, presented in Section 3.3.1, the downsampled versions are visually good as aliasing only affects small parts of the images, usually in areas with very sharp transitions, such as the book of the *CatRobot* video clip, the power lines of *DaylightRoad* or the drain grill from *ToddlerFountain*, as depicted in **Fig. 3.4**.

A second type of artifacts is also introduced, or amplified, in the processed sequences. Indeed, the base layer obtained by applying the polyphase decomposition on the *CatRobot* sequence shows a very noisy background (the moving scarfs), whereas the original clip does not contain as much visible noise. For other sequences like *DaylightRoad*, or to a lesser extent *Drums*, the existing noise is amplified by the subsampling process, sometimes so significantly that it disturbs the viewer.

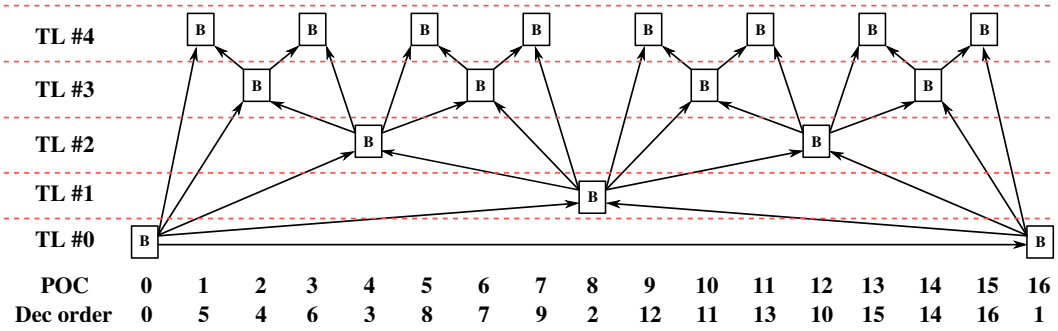


Fig. 3.5 Random Access GOP16 structure.

3.2 Enhanced Polyphase Coding Scheme

This section focuses on exploring possible improvements to the scalable coding scheme based on the polyphase decomposition. First, two modifications to the GOP structure are detailed. They respectively rely on modified TL QP offsets and the addition of intra-layer temporal predictions, thus taking into account the special features of the polyphased decomposed signal. Then, a simple filtering operation is presented to improve the management of chroma pixels during the decomposition stage.

3.2.1 Resolution Component dependent Temporal Layer QP Offsets

In the Common Test Conditions (CTC) for the scalable encoder SHVC [53], JCT-VC defines a GOP structure of sixteen images, as depicted in **Fig. 3.5**, with five different TLs. These TLs play an important role in HEVC because several parameters can be set at a temporal layer level. For instance, in the CTC, a different QP is applied for each temporal layer, the lowest QP being attributed to layer zero. This feature has been defined to favor pictures that are used as reference, i.e. images from the lower temporal layers.

When used for the encoding of a polyphase decomposed signal, RC #0 will always be on the three lower temporal layers, whereas the other three RCs will be either on the fourth or fifth temporal layers, thus having higher QPs. While this feature could be interesting in a scalable scheme where the major part of the bitrate would be attributed to the BL, such a feature could be prejudicial to the quality of the full resolution reconstructed signal. Indeed, with the current CTC, the number of transform blocks, which contain the residuals of temporal predictions between images of different RCs, are hardly present in the encoded fourth and fifth layer images due to their higher QP. Encoder decisions are further detailed in Section 3.4.

In order to improve the coding efficiency for the three RCs of the EL, two alternative configurations, relying on the decrease of the quantization step for these sub-images, are proposed:

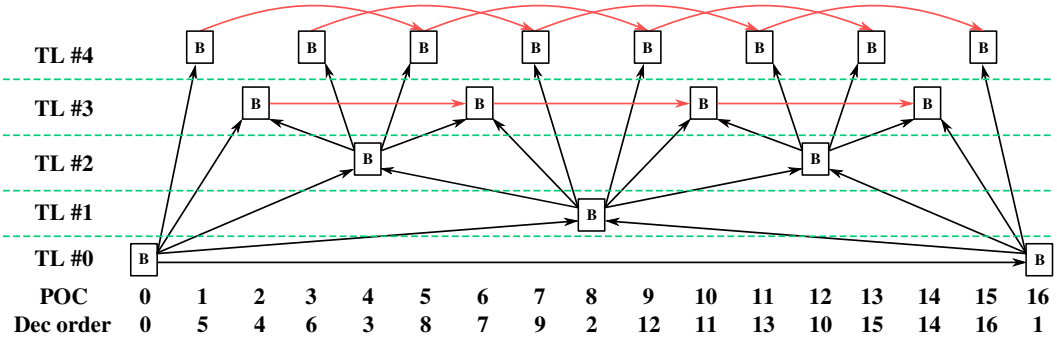


Fig. 3.6 Modified Random Access GOP16 structure.

- Base layer unchanged, QP offsets equal to 3 for TL#3 and TL#4.
- Base layer unchanged, QP offsets equal to 0 for TL#3 and TL#4.

3.2.2 Intra-layer Temporal Predictions

Another idea is to allow for predictions between images of the same RC, i.e. between images of the same temporal layers. This feature adds an alternative choice for the encoder that should be selected when the correlation between two adjacent pixels of the original full resolution image is less interesting, in a RDO sense, than the correlation between pixels of two consecutive full resolution images.

Figure 3.6 depicts a modified GOP structure that allows such intra-layer predictions. It is important to note that the inter-layer predictions between the fourth and fifth temporal layers have been deactivated to limit the additional encoding time induced by the new GOP structure.

3.2.3 Chroma Phase Shift Compensation

The polyphase subsampling technique takes one over four pixels, in each 2x2 block of the input image, to create the four RCs. This decomposition is performed on both the luma and chroma planes of the input images. However, for a video in 4:2:0 format, which is the most commonly used chroma subsampling format, the polyphase decomposition introduces a misalignment between the chroma pixels of the different sub-resolution images, as depicted in **Fig. 3.7**. This misalignment has a significant impact on the motion compensation efficiency in the chroma channels.

Indeed, during the motion compensation step, an encoder usually derives the MVs of the chroma planes from the MVs of the luma plane. For example, in HEVC, the luma MVs are scaled by a factor of two in both directions to compute the chroma MVs. For a polyphase subsampled video, when a frame from a different RC is used as reference for an inter-prediction,

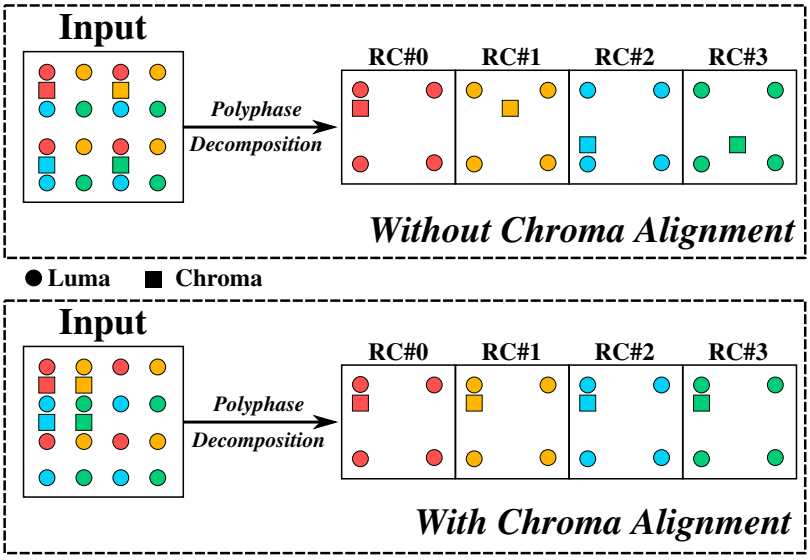


Fig. 3.7 Pixel positions in 4:2:0 format for polyphase subsampling with and without chroma pixel alignment.

the derived chroma motion vector is inherently wrong due to the misalignment of the chroma pixels between RCs. The resulting chroma plane prediction is sub-optimal thus increasing the amount of residuals to encode. To overcome this issue one can either change the chroma motion vector derivation process depending on the RCs or realign the chroma planes before the polyphase subsampling.

In order to avoid changing the core encoding process and to keep the proposed solution entirely as a pre-processing step, a realignment of the chroma planes of the different RCs is proposed, as shown in **Fig. 3.7**, using a simple mean filter before the polyphase subsampling. The mean filter is applied on each 2x2 chroma block of the input full resolution image, as described in Equation (3.1)

$$\begin{cases} \mathbf{I}_{align}(2n,2m) & = \mathbf{I}_{src}(2n,2m), \\ \mathbf{I}_{align}(2n+1,2m) & = \frac{1}{2} \cdot (\mathbf{I}_{src}(2n,2m) + \mathbf{I}_{src}(2n+1,2m)), \\ \mathbf{I}_{align}(2n,2m+1) & = \frac{1}{2} \cdot (\mathbf{I}_{src}(2n,2m) + \mathbf{I}_{src}(2n,2m+1)), \\ \mathbf{I}_{align}(2n+1,2m+1) & = \frac{1}{2} \cdot (\mathbf{I}_{src}(2n,2m) + \mathbf{I}_{src}(2n+1,2m+1)), \end{cases} \tag{3.1}$$

with $\mathbf{I}_{src}$ the input picture, $\mathbf{I}_{align}$ the output chroma aligned picture, n and m the row and column indexes, respectively. The inverse operation is performed after the reconstruction of the output full resolution image from the four decoded RCs.

3.3 Experimental Results

This section provides a study on the performance of the proposed enhanced polyphase scalable coding scheme on UHD sequences. First, the test sequences and experimental setup used for the performance evaluations are detailed. Then, an objective evaluation is performed to compare the polyphase-based coding scheme, with and without the proposed improvements, to SHVC.

3.3.1 Test Sequences

Since the subject of the study is focused on the coding of high resolution content, a dataset comprising only UHD sequences has been defined. The dataset is similar to classes A1 and A2 of the JCT-VC test sequences, with video clips made available by Netflix, SJTU, Huawei and b<>com. Table 3.1 summarizes the characteristics of each test sequence. The sequences have a 5-second duration and a frame-rate of 60 fps, except for *Drums*, which has a duration of 3 seconds and a frame-rate of 100 fps.

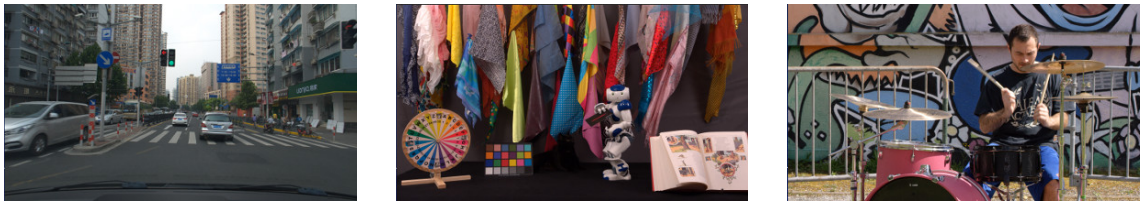
Table 3.1 UHD test set sequences and their characteristics.

Sequence	Resolution (width x height)	Frame rate (fps)	Frames
DaylightRoad	3840x2160	60	300
CatRobot	3840x2160	60	300
Drums	3840x2160	100	300
ToddlerFountain	4096x2160	60	300
Tango	4096x2160	60	294
RollerCoaster	4096x2160	60	300

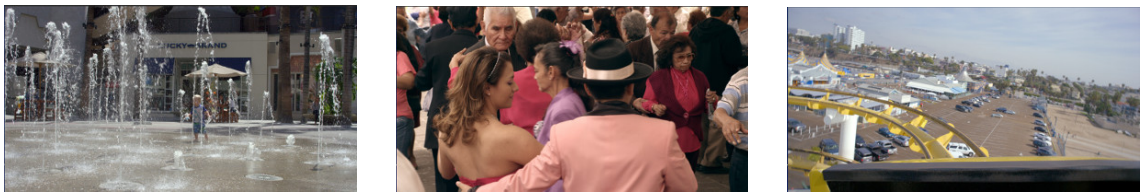
As depicted in **Fig. 3.8**, the database offers a wide range of natural content. The different sequences have been selected from various sources to ensure that they have different spatio-temporal characteristics. Two simple metrics recommended by the ITU-T [29], namely the spatial information SI and temporal information TI, have been computed for each video clip to assess the validity of this statement.

The spatial information measurement (SI), i.e. the level of spatial details in the video sequence, is based on the Sobel filter. Each frame F_n of the test sequence is thus first filtered with the Sobel filter $Sobel(F)$ before computing the standard deviation of the resulting image. Once the entire video clip has been processed, the spatial information SI is obtained by taking the maximum value of the standard deviation of the Sobel-filtered frames over the whole sequence. Equation (3.2) represents this process:

$$SI = \max_{n}\{std_{i,j}(Sobel(\mathbf{F}_n(i,j)))\}, \tag{3.2}$$



(a) DaylightRoad (b) CatRobot (c) Drums



(d) ToddlerFountain (e) Tango (f) RollerCoaster

Fig. 3.8 Screenshots of UHD test sequences.

where $\mathbf{F}_n(i,j)$ is the pixel located at the i^{th} row and j^{th} column of the n^{th} frame.

The temporal information measurement (TI), i.e. the amount of motion between successive frames, is based on the motion difference feature, which is the difference between the pixel values of two consecutive frames. $M_n(i,j)$, the motion difference pictures, is defined as

$$\mathbf{M}_n(i,j) = \mathbf{F}_n(i,j) - \mathbf{F}_{n-1}(i,j). \tag{3.3}$$

The temporal information, TI, is then computed as the maximum, over all frames, of the standard deviation of the motion difference picture, as defined in Equation (3.4).

$$TI = \max_n \{std_{i,j}(\mathbf{M}_n(i,j))\}. \tag{3.4}$$

Fig. 3.9 shows the SI-TI metrics computed for the test sequences comprised in the dataset. As expected, the video clips have different spatio-temporal characteristics, with *Tango* and *RollerCoaster* showing a low amount of details compared to sequences like *Drums*, and with *CatRobot* having a low temporal information measure compared to *ToddlerFountain*.

Despite the simplicity and known flaws of these two metrics for the precise measurement of spatio-temporal information, they are reliable enough to assess the diversity of the spatio-temporal characteristics of video contents. Therefore, the differences observed should lead to various behaviors over the dataset for the compression algorithms that will be tested throughout the different studies carried out.

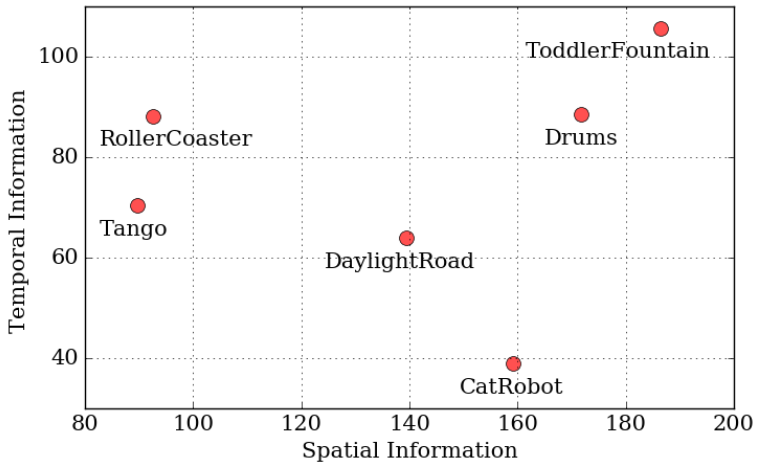


Fig. 3.9 Perceptual spatio-temporal information of the test sequences

3.3.2 Experimental Set-up

Encodings have been performed with the HEVC reference software (HM16.12) [41] and SHVC reference software (SHM9.0) [47]. According to the CTC for RA configuration, a fixed QP scheme has been used for all encodings, with a hierarchical GOP of size 16 and an intra-period of approximately one second for both the reference SHVC coding chain and the proposed one. The performances have been assessed in terms of rate-distortion using the Bjøntegaard delta metric (BD-Rate) [30], representing the average bitrate savings for equal PSNR values. In order to evaluate the proposed scalable coding chain over a wide range of bitrates, QP values ranging from 22 to 40 with a step of two have been used for each test sequence.

3.3.3 Objective Evaluation

Results for Resolution Component dependent Temporal Layer QP Offsets

After encoding the UHD dataset following the coding chain presented in Section 3.1.2, the two specific TL QP offset configurations did not show improved coding performances. The decrease of the QP, i.e. a lower quantization step and a lower λ parameter in the RDO, did increase the amount of transform units signaled in the bitstream, but the resulting higher bitrate was not compensated by a substantial increase in PSNR value for the reconstructed output signal.

Results for Intra-Layer Temporal Predictions

This new GOP structure has been tested with and without the modified temporal layer QP offsets. The coding performances were similar to the original GOP structure, with average

Table 3.2 BD-Rate results for chroma aligned polyphase compared to original scheme.

Sequence	PSNR-Y	PSNR-U	PSNR-V	PSNR-YUV
DaylightRoad	-0.16	-12.5	-11.2	-2.96
CatRobot	-3.55	-28.0	-31.8	-8.98
Drums	-1.94	-15.0	-10.9	-3.51
Tango	-0.10	-7.87	-14.9	-3.66
RollerCoaster	-0.52	-14.5	-7.64	-2.64
ToddlerFountain	-0.03	-4.47	-4.39	-0.11
Average	**-1.1**	**-13.7**	**-13.5**	**-3.6**

BD-Rate values very close to 0%. Indeed, the new reference pictures allowed by the modified GOP are not often chosen by the encoder as temporal predictor, preferring the RC #0 image encoded with a lower QP. With the modified temporal layer QP offsets configurations, the new predictions represent the majority of chosen reference pictures, with a slightly lower resulting reconstructed quality without any decrease in bitrate, compared to the original GOP structure with modified QP offsets.

To conclude this study on modification of the encoding GOP structure, the proposed changes were not retained as they do not offer increased encoding performances, contrary to the chroma phase compensation that improved the prediction quality of both chroma planes, as shown in the next section.

Results for Chroma Phase Shift Compensation

Table 3.2 shows the results of the realignment of chroma pixels of the polyphase subsampled signal presented in Section 3.1. Negative BD-Rate values represent bitrate savings for the proposed decomposition compared to the original polyphase subsampling. It can be observed that the proposed chroma realignment improves the performance for both chroma channels (U and V) with an average gain of 13.6% compared to the original polyphase subsampling. Thus, the losses due to the rounding error introduced by the proposed filtering are more than compensated by the improved prediction step with correct chroma motion vector derivation. Table 3.2 also shows minor gains for the luma channel. This can be explained by the overall decrease in terms of number of bits, due to the improved chroma motion estimation, necessary to achieve the same quality.

The polyphase subsampling scheme with and without the proposed chroma realignment is compared to SHVC in Table 3.3. For equal PSNR-YUV, an average bitrate overhead of 6.6% can be observed for polyphase with chroma realignment. Considering the per-sequence results, it can observed that performance is highly variable, ranging from a 45% overhead to a 13.3% bitrate reduction. Indeed, on one hand, the temporal prediction between

Table 3.3 BD-Rate results (%) for proposed schemes vs SHVC.

Sequence	Original polyphase				Polyphase with Chroma Alignment			
	Y	U	V	YUV	Y	U	V	YUV
DaylightRoad	33.6	128	159	47.8	33.4	104	142	44.9
CatRobot	23.9	152	200	39.9	19.6	80.4	98.1	28.5
Drums	-10.3	74.2	93.1	-1.1	-12.1	46.1	69.4	-4.6
Tango	-17.2	39.9	49.9	-4.1	-17.3	31.0	31.4	-6.5
RollerCoaster	-16.9	32.0	25.5	-11.1	-17.4	11.3	15.0	-13.3
ToddlerFountain	-11.8	59.4	35.1	-9.4	-11.8	50.9	26.9	-9.5
Average	**0.2**	**81.1**	**93.8**	**10.3**	**-0.9**	**53.9**	**63.8**	**6.6**
Encoding $TR_{\%}$	**55**				**55**			
Decoding $TR_{\%}$	**42**				**42**			

the enhancement and base layers allows for the recovery of the spatial details of the full resolution input video for most test sequences, resulting in a bitrate reduction compared to SHVC. On the other hand, for the *DaylightRoad* and *CatRobot* test sequences, the large losses can be explained by a noisy input, especially for *DaylightRoad*, and by the presence of very sharp details, mostly text. The spatial high frequencies of these sharp details are partially or totally lost in the base layer during the decomposition step and are thus too costly to recover due to the high residual energy in the enhancement layer frames. The same behavior is observed in the chroma channels for all sequences, where the smaller resolution of the chroma planes amplifies this effect, which results in high positive chroma BD-Rate values.

Regarding the coding complexity, the polyphase scalable scheme provides a substantial 55% (resp. 42%) reduction in encoding (resp. decoding) time compared to SHVC. The encoding complexity gain could be further increased, by limiting the possible reference frames for the upper temporal layers, without any impact the coding efficiency.

3.4 Study on Encoder Decisions

This section aims at verifying that the encoder decisions follows the expected behavior considering the particularities of the polyphase decomposed signal being encoded. Indeed, the polyphase-based scalable coding scheme relies on the use of temporal predictions to predict spatially adjacent pixels. Therefore, RCs #1, #2 and #3 - respectively the top-right, bottom-left and bottom-right pixels of each 2x2 block of the original input image - are supposed to be temporally predicted from the RC #0 - the top-left pixel. Thus, the majority of blocks should be encoded using a forward inter-prediction with motion vectors corresponding to the phase difference between the sub-images of the base and enhancement layers. For in-

Table 3.4 Prediction mode decisions for polyphase decomposed CatRobot sequence (QP 26).

Prediction mode	**Average number of CUs per frame**					
	RC #1		RC #2		RC #3	
Intra	0	**0 %**	6	**0.3 %**	1	**0 %**
Inter	150	**12.3 %**	459	**24.4 %**	125	**9.6 %**
Skip	1069	**87.7 %**	1419	**75.3 %**	1173	**90.4 %**
Total	1219	**100 %**	1884	**100 %**	1299	**100 %**

stance, the MV (2,0) - half-pixel motion in quarter pixel precision - should be used for RC #1 blocks predicted from their corresponding RC #0 block. It is important to note that RCs #1 and #3 sub-images can also be predicted from RC #2 images due to the particularity of the Random Access GOP structure used in this experiment (see Section 3.2.1 for the detailed structure). To verify these hypothesis, the chosen prediction modes are first studied, followed by an analysis of the selected prediction directions, based on the encoding of the *CatRobot* sequence at QP 26 ($\approx$ 12 Mbps).

3.4.1 Chosen Prediction Modes

First, the prediction modes selected for the encoding of the enhancement layer images have been studied. Table 3.4 shows the average number of CUs for different prediction modes, namely intra, inter and skip. The skip mode corresponds to CUs that contain only one PU of the same size, using block merging (full inheritance of motion parameters from spatially or temporally adjacent blocks) and without any residual data signaled in the bitstream.

As expected, the number of blocks encoded using intra prediction is close to zero for the three RCs of the enhancement layer. Indeed, almost every blocks are coded using temporal prediction - inter or skip mode - thus confirming the interest of the polyphase decomposition scheme. In addition, the majority of these temporally predicted blocks - 95.9%, 92.4% and 96.2% for RCs #1, #2 and #3, respectively - are using the merge mode, suggesting a certain spatial homogeneity among motion vectors of each sub-resolution image.

3.4.2 Chosen Prediction Directions

To confirm the hypothesis on the expected motion vectors and their direction, the PUs have been analyzed. Table 3.5 shows the average number of PUs using each possible temporal prediction direction: L0 usually contains P or B reference images with lower Picture Order Count (POC) numbers (forward prediction), L1 mainly contains B reference images with higher POC numbers (backward prediction) and bi-pred is for bi-directional predictions using both L0 and L1 reference picture lists.

Table 3.5 Chosen temporal prediction directions for polyphase decomposed CatRobot sequence at QP 26.

Prediction direction	**Average number of PUs per frame**					
	RC #1		RC #2		RC #3	
forward	1001	**78.6 %**	1468	**71.9 %**	1209	**88.8 %**
backward	76	**6 %**	99	**4.8 %**	30	**2.2 %**
bi-pred	196	**15.4 %**	475	**23.3 %**	122	**9 %**
Total	1273	**100 %**	2042	**100 %**	1361	**100 %**

As expected, forward prediction is the most chosen mode, mainly corresponding to the use of RC #0 (or #2) images to predict frames of the enhancement layer. Among these forward predictions, more than 90%, in average, of the signaled motion vectors directly match the *phase* difference between the two RCs images. Thus, the experimentally observed encoder decisions correspond to the behavior that was expected while designing the polyphase decomposition scheme.

3.5 Conclusion

In this chapter, improvements to the polyphase decomposition scheme, a pre/post-processing tool to achieve x2 spatial scalability with a single HEVC encoder instance, have been proposed.

First, two modifications of the GOP structure aiming at adapting the coding configuration to the format of the polyphase decomposed signal have been proposed. It has been shown that these modifications, on one hand relying on different TL QP offsets and on the other hand on intra-layer temporal predictions, did not yield any improvements over a legacy HEVC coding scheme, in terms of bitrate savings.

Then, a chroma phase shift filter has been presented to correct a flaw inherent to the polyphase decomposition. It has been shown that this correction leads to 3.6% average bitrate savings compared to the original polyphase decomposition, and an averaged bitrate overhead of 6.6% compared to SHVC with a complexity reduction of 55% and 42% at the encoder and decoder sides, respectively.

However, a high proportion of skipped blocks, and thus the very low amount of transform residuals signaled in the bitstream, is observed. This raises the issue of the capacity to reconstruct the level of details of the original full resolution. Indeed, even with a good prediction stage, a bigger amount of residuals should remain in the bitstream at high bit rates. The lack of transmitted transform residuals induces a level of detail relatively low in the reconstructed output video, hence the BD-Rate values observed, especially on the

chroma planes where the amount of transform blocks is even closer to zero. This limited performance could be the result of the aliasing introduced by the polyphase decomposition performed prior to encoding, which makes the spatial details harder to encode due to the presence of unwanted high spatial frequencies.

Chapter 4

Wavelet-based Decomposition

As stated in the previous chapter, the main flaw of the polyphase subsampling technique is the potential aliasing introduced in the four sub-resolution images obtained after decomposition. Indeed, aliasing creates high frequencies which are costly to encode, for usual video coding algorithms, due to the magnitude and position of the corresponding coefficients in the transformed residual blocks. This is especially a problem for the resolution component which serves as base layer, since it will be encoded with a lower QP than the sub-images of the enhancement layer, if an HEVC encoding with a traditional hierarchical GOP is considered.

To avoid this issue, the use of a wavelet-based decomposition instead of the polyphase subsampling is proposed in this chapter. Indeed, wavelet transforms have the nice property of generating a lowpass version of the input image, thus with much lower aliasing, which could serve as base layer in the considered low-complexity scalable coding scheme. Therefore, the contribution of this chapter focuses on the design of a modified wavelet transform used in lieu of the polyphase decomposition as the pre-processing operation of the proposed scalable coding scheme.

This chapter is organized as follows. Section 4.1 provides a brief introduction to the wavelet transform theory and an overview of standardized wavelet-based coding solutions. Then, the proposed wavelet-based scalable coding scheme is presented in Section 4.2. Several modifications of the encoding process to suit the wavelet transformed signal are investigated in Section 4.3. Finally, experimental results are detailed in Section 4.4 while Section 4.5 concludes this chapter.

4.1 Related Works

Wavelet-based image and video coding has been an important field of research as an alternative to the traditional block-based hybrid encoding. This section aims at introducing the theory of wavelet transforms and its application to video coding. An in-depth introduction

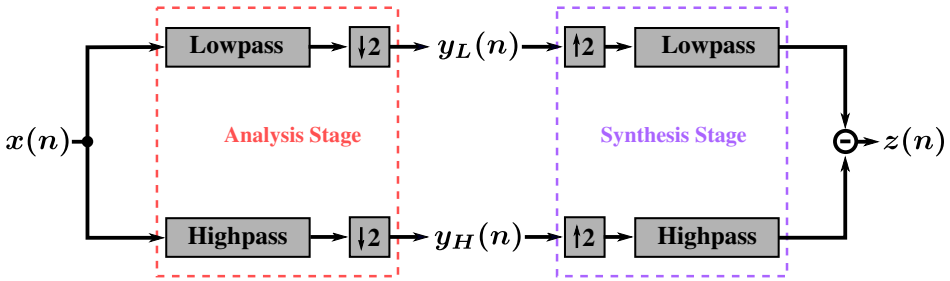


Fig. 4.1 Generic one-level 1-D DWT with analysis and synthesis stages.

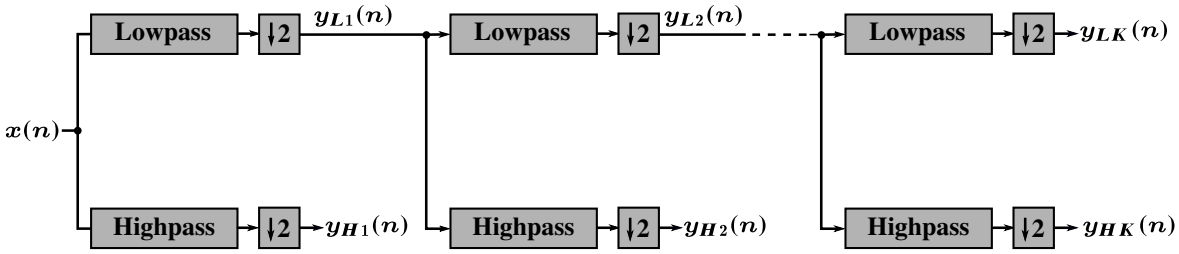


Fig. 4.2 Analysis stage of a K-level 1-D wavelet decomposition. $\mathbf{y}_{f_k}$ represents the wavelet coefficients for frequency sub-band f, $f \in \{L,H\}$, at the k^{th} level of decomposition.

to the wavelet theory is out of the scope of this work and can be found in [54–61]. This section is organized as follows. The key features of the Discrete Wavelet Transforms (DWT) are first introduced, followed by a presentation of the standardized image and video codecs using wavelet transforms.

4.1.1 Background on Discrete Wavelet Transforms

Definition of the DWT

The generic form of a DWT is depicted in **Fig. 4.1**. The analysis stage, i.e. the forward transform, is first performed on the input signal $\mathbf{x}(n)$. It is done by applying the lowpass and highpass wavelet filters. Each filtering operation is followed by a downsampling process which selects the even and odd samples to respectively obtain the lowpass sub-band $\mathbf{y}_L(n)$ and highpass sub-band $\mathbf{y}_H(n)$. The synthesis stage, i.e. the inverse transform, is the performed to reconstruct the output signal $\mathbf{z}(n)$. To this end, an upsampling is first performed on both sub-bands consisting in inserting zeros in-between every two samples. Finally, the output reconstructed signal $\mathbf{z}(n)$ is the sum of the lowpass filtered upsampled lowpass sub-band and highpass filtered highpass sub-band.

The 1-D DWT process can be performed with several levels of decomposition, as depicted in **Fig. 4.2** for the analysis stage. Both a lowpass and highpass filers are applied on the input signal $\mathbf{x}(n)$ followed by downsampling by a factor of 2, thus constituting one level of transform. For each subsequent level, the same process is performed on the lowpass

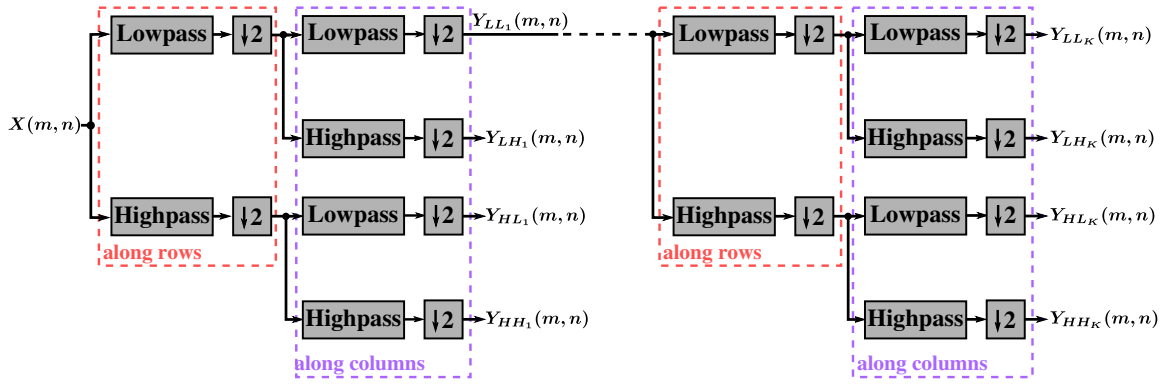


Fig. 4.3 Analysis stage of a K-level 2-D wavelet decomposition. $\mathbf{y}_{fk}$ represents the wavelet coefficients for frequency sub-band f, $f \in \{LL, LH, HL, HH\}$, at the k^{th} level of decomposition.

frequency sub-band of the previous level. The number of levels is a finite number K, with the filter outputs $\mathbf{y}_{H_k}(n)$ and $\mathbf{y}_{L_k}(n)$, $k \in \{1,2,...,K\}$, referred to as k^{th} wavelet coefficients. The wavelet filter coefficients, and subsequent transform properties, depend on the wavelet family, such as the Haar, Daubechies or bi-orthogonal wavelets, to name a few.

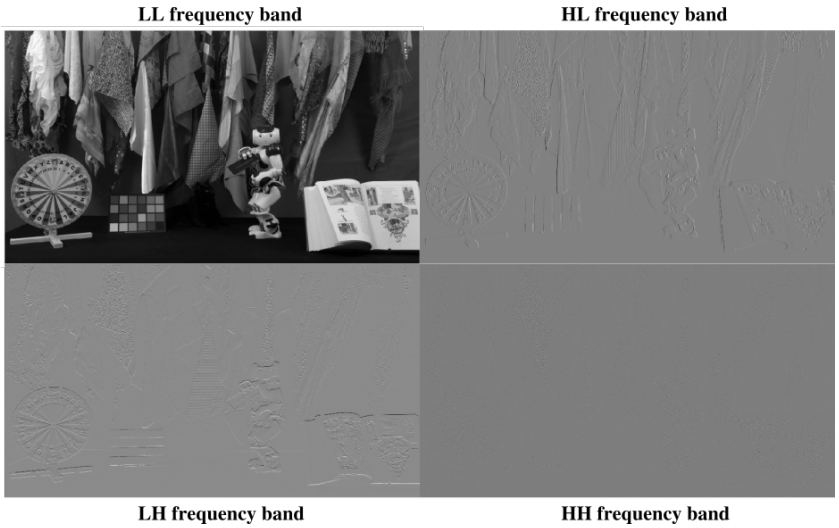
The 1-D DWT process can be extended to 2-D using separable filters [62]. **Fig. 4.3** illustrates the generic form of a K-level 2-D wavelet transform. For each level, the transform process is achieved by successively applying the 1-D lowpass and highpass filters along rows and columns of the input signal $\mathbf{X}(m,n)$, resulting in 4 sub-band signals $\mathbf{Y}_{LL}(n)$, $\mathbf{Y}_{LH}(n)$, $\mathbf{Y}_{HL}(n)$ and $\mathbf{Y}_{HH}(n)$. For each subsequent decomposition level k, the same transform process is performed on the lowpass sub-band of the previous level $y_{LL_{k-1}}$. As for the 1-D transform, $\mathbf{Y}_{LL_k}(n)$, $\mathbf{Y}_{LH_k}(n)$, $\mathbf{Y}_{HL_k}(n)$ and $\mathbf{Y}_{HH_k}(n)$ are referred to as k^{th} wavelet coefficients.

The result of a one-level 2-D Haar transform applied on the input image depicted in **Fig. 4.4a** is shown in **Fig. 4.4b**, with the wavelet sub-bands, namely LL, LH, HL and HH respectively carrying the lowpass filtered image, horizontal, vertical and diagonal contour images.

Lifting Scheme

There are two different conventional filtering schemes to perform compute a wavelet transformed signal: the convolution scheme and the lifting scheme. The convolution scheme consists in traditionally convolving the filter kernel with the input image, which can become computationally expensive for large wavelets, such as the well-known 9/7 Daubechies wavelet [63]. The lifting scheme [64] consists in a sequence of simple filtering operations. The basic idea is to split the input signal in two, the odd samples on one hand and the even samples on the other hand. The highpass sub-band is obtained by applying the highpass filter g on the odd samples whereas the lowpass sub-band is a linear combination of the fil-

(a) Original CatRobot image



(b) One level 2-D decomposition using the Haar wavelet

Fig. 4.4 Example of 2-D DWT with the Haar transform.

tered even sampled, with the lowpass filter h, and highpass sub-band scaled with the scaling operator s. **Fig. 4.5** depicts the block diagram of this lifting scheme for a 2-D DWT.

As for any filtering operation, the case of image boundaries must be handled with care. For wavelet transforms, Usevitch showed in [65] that for the output to be artifact free and without coefficient expansion, the border extension and wavelet must satisfy certain constraints. On one hand, a *whole sample* symmetric extension at image boundaries must be performed before filtering. It consists in symmetrically repeating the signal samples at the exception of the first and last samples of each row or column, which are whole samples and thus only appear once. Therefore, the required number of extended samples is only filter-dependent and can thus be derived from the filter length of the considered DWT. On the other hand, the wavelet filters must be linear phase filters, i.e. either symmetric or anti-

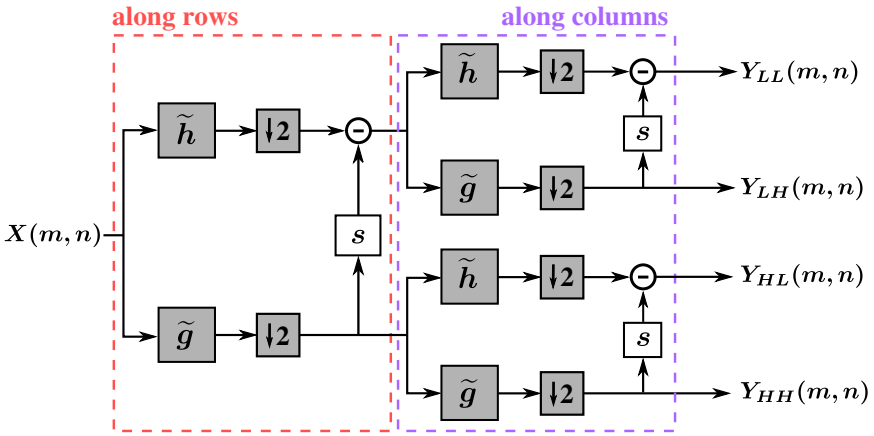


Fig. 4.5 2-D Discrete Wavelet Transform - Lifting scheme.

symmetric. If these two constraints are respected, the wavelet filter pairs (h,g) satisfy the perfect reconstruction property, defined by Equations 4.1 and 4.2. The output image of the wavelet synthesis stage is thus identical to the input image of the analysis stage, even at image boundaries.

$$h(z)\tilde{h}(z^{-1}) + g(z)\tilde{g}(z^{-1}) = 2z^{-l} \tag{4.1}$$

$$h(z)\tilde{h}(-z^{-1}) + g(z)\tilde{g}(-z^{-1}) = 0 \tag{4.2}$$

However, there exists only one linear phase filter set for real-valued and compactly supported orthogonal wavelets, the trivial Haar filter [54]. Thus, a more general form of wavelets have been designed, known as bi-orthogonal wavelets, to allow for symmetric linear phase filters. These wavelets are thus not energy preserving due to their non-orthogonality, which must be taken into account for an optimal coding of wavelet coefficients.

4.1.2 Wavelet-based Standardized Codecs

Based on the breakthrough in wavelet-based coding introduced by Shapiro with the Embedded Zero-tree Wavelet (EZW) algorithm [66], several wavelet coding scheme showed promising state-of-the-art coding performance [67–69], notably the Embedded Block Coding with Optimized Truncation (EBCOT) algorithm [67, 70], and inspired the successor of the well-known DCT-based JPEG still image coding standard [71], JPEG-2000 [72, 73].

Wavelet transform is thus at the core of the JPEG-2000 as depicted in **Fig. 4.6**, with either the Daubechies 9/7 wavelet for the lossy profile or the Le Gall 5/3 wavelet for the lossless profile of the standard. Since then, the standard has been extended to support video coding with Motion JPEG 2000 [74], which introduces the the high level syntax enabling a bitstream containing only frames individually coded with JPEG-2000.

Additionally, in an effort to improve the coding efficiency of wavelet-based video coding,

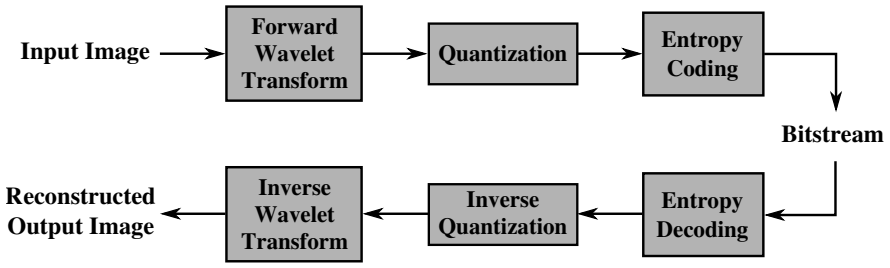


Fig. 4.6 General Block Diagram of the JPEG2000 still image coding standard.

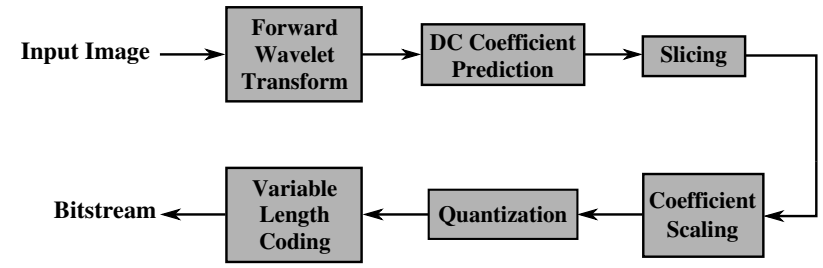


Fig. 4.7 General Block Diagram of the VC-2 video coding standard.

several researchers worked on 3-D wavelet transforms [75–78], the third dimension being a temporal filtering of the images, with optimal motion estimation in the wavelet transform domain [79].

More recently, a second codec, based on the Dirac Pro solution of the British Broadcasting Corporation (BBC) [80], has been standardized through the Society of Moving Pictures and Television Engineers (SMPTE) norm ST 2042-1 [81]. VC-2 is a low-delay solution for for the production industry, and thus focuses on different aspects of video compression than typical broadcast-oriented codecs, namely low latency, low complexity and high fidelity. As depicted in **Fig. 4.7**, the encoder is based on a lifted wavelet transform, chosen among several supported wavelets whose implementation has been optimized to reach a latency equal to a few lines. Due to the hard latency and complexity constraint of the production use-case, a high compression ratio is not required. VC-2 thus allows for compression ratios ranging from 2:1 up to 8:1, depending on the chosen supported wavelet and other coding parameters.

4.2 Proposed Wavelet-based Pre-Processing Tool

4.2.1 Modification of the Decomposition Step

The proposed wavelet-based decomposition process is shown in **Fig. 4.8**, with the modifications depicted in green. First, a DWT is applied on the input signal to produce four different sub-bands. The *LL* sub-band, corresponding to a lowpass filtered version of the input signal, is used as base layer in the scalable coding chain. The *LH*, *HL* and *HH* sub-bands,

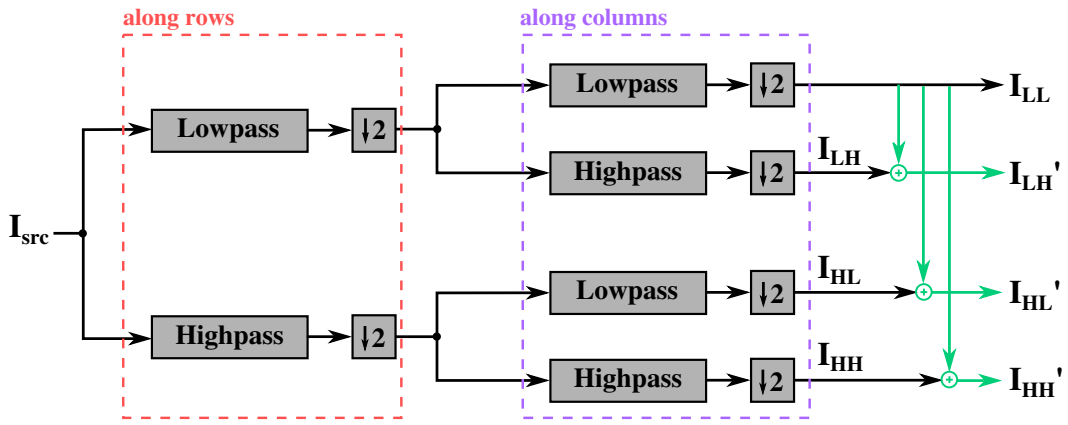


Fig. 4.8 Discrete Wavelet Transform with modifications displayed in green - Analysis stage.

which correspond to horizontal, vertical and diagonal high-frequencies, respectively, form the enhancement layer.

To make the decomposed signal suitable for a standard HEVC encoder, the *LL* sub-band is added to each of the other three high-frequency sub-bands, thus enabling temporal prediction between the base and enhancement layers. Indeed, the residuals of a temporal prediction between a *LL* block and a LH' block with a (near-)zero motion vector would be equivalent to the corresponding original *LH* block. The temporal packing of the sub-bands follows the same process as for the polyphase subsampling, each sub-band corresponding to a different resolution component.

The coding chain integrating the proposed wavelet-based decomposition remains the same as the one described in Section 3.1.2, at the exception of the polyphase decomposition step being replaced by the proposed wavelet-based decomposition. The spatial scalability is thus achieved by only decoding the lowpass filtered sub-band, which composes the lower temporal layers, if only the base layer is needed at the display end. If the original resolution is required, the entire bitstream is decoded and the inverse DWT (synthesis stage) is performed to reconstruct the full resolution output video.

4.2.2 Considered DWTs

Considering the coding chain defined in the previous section, several DWTs can be used, chosen among a vast pool of existing wavelets. Since the wavelet sub-bands are encoded using a standard HEVC encoder, which uses integer format for the input video, the sub-band coefficients must be integers. Thus, only the integer-to-integer DWTs Haar and Le Gall 5/3 are considered in this study.

4.2.2.1 Haar Wavelet

The Haar wavelet is one of the most simple and easy to implement DWT, which has the nice property of being orthogonal with perfect reconstruction even with integer values [54]. For the analysis stage (respectively synthesis stage), the 1-D filtering is performed by applying the lowpass and highpass filters of Equation (4.3), (resp. Equation (4.4)) successively on both rows and columns of the input image, as depicted in **Fig. 4.1**.

$$\textit{Analysis} \begin{cases} \mathbf{a}_{LP}(n) &= \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n) + \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n-1) \\ \mathbf{a}_{HP}(n) &= -\frac{1}{\sqrt{2}} \cdot \mathbf{x}(n) + \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n-1) \end{cases} \tag{4.3}$$

$$\textit{Synthesis} \begin{cases} \mathbf{s}_{LP}(n) &= \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n) - \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n-1) \\ \mathbf{s}_{HP}(n) &= \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n) + \frac{1}{\sqrt{2}} \cdot \mathbf{x}(n-1) \end{cases} \tag{4.4}$$

with n the sample index, $\mathbf{x}$ the input to the considered filter, $\mathbf{a}_{LP}$ and $\mathbf{a}_{HP}$ (resp. $\mathbf{s}_{LP}$ and $\mathbf{s}_{HP}$) the outputs of the considered filters for the analysis (resp. synthesis) stage.

As previously said, the Haar wavelet is very simple. Indeed, the 2-D lowpass filtering basically corresponds to the mean of the pixels of each 2x2 block. Therefore, the *LL* sub-band contains very smooth edges and is lacking sharp details, decreasing the perceptual quality of the base-layer, but also making it easier to encode. Additionally, the high-pass sub-bands mainly contains high coefficients on sharp edges, which can be partially predicted from the *LL* sub-band. This wavelet could thus lead to interesting results, despite its simplicity, when used in the proposed coding scheme.

4.2.2.2 Le Gall 5/3 Wavelet

The Le Gall 5/3 wavelet [82] is more complex than the Haar wavelet previously described. It uses a 5-tap lowpass (respectively highpass) filter and a 3-tap highpass (resp. lowpass) at the analysis (resp. synthesis) stage. The filters are applied successively on both rows and columns and are defined using Equations 4.5 and 4.6.

$$\textit{Analysis} \begin{cases} \mathbf{a}_{LP}(n) &= \frac{3}{4} \cdot \mathbf{x}(n) + \frac{1}{4} \cdot [\mathbf{x}(n-1) + \mathbf{x}(n+1)] - \frac{1}{8} \cdot [\mathbf{x}(n-2) + \mathbf{x}(n+2)] \\ \mathbf{a}_{HP}(n) &= \mathbf{x}(n) - \frac{1}{2} \cdot [\mathbf{x}(n-1) + \mathbf{x}(n+1)] \end{cases} \tag{4.5}$$

$$\textit{Synthesis} \begin{cases} \mathbf{s}_{LP}(n) &= \mathbf{x}(n) + \frac{1}{2} \cdot [\mathbf{x}(n-1) + \mathbf{x}(n+1)] \\ \mathbf{s}_{HP}(n) &= \frac{3}{4} \cdot \mathbf{x}(n) - \frac{1}{4} \cdot [\mathbf{x}(n-1) + \mathbf{x}(n+1)] - \frac{1}{8} \cdot [\mathbf{x}(n-2) + \mathbf{x}(n+2)] \end{cases} \tag{4.6}$$

With its wider lowpass filter, the Le Gall 5/3 DWT should result in a visually better *LL* sub-band than the Haar DWT, making it potentially more suitable for a spatial scalability scheme. However, contrary to the Haar transform, the Le Gall 5/3 DWT is not orthogonal, and thus not energy preserving, which could induce a sub-optimal bit allocation while encoding it with HEVC. Section 4.3.2 explores this issue of non-orthogonality and rate-distortion optimization.

4.2.3 Implementation

The two above mentioned discrete wavelet transforms have been implemented in C++ and added to the software package already used for the modified polyphase decomposition. In order to reduce the computation time, fast algorithms have been considered.

For the Haar DWT, composed of low complexity 2-tap lowpass and highpass filters, the 2x2 kernels of Equation (4.7) have been applied to the $\mathbf{I}_{src}$ input image to generate the transformed image $\mathbf{I}_{DWT}$ and the kernels of Equation (4.8) to reconstruct the output image $\mathbf{I}_{rec}$.

$$Analysis \begin{cases} \mathbf{I}_{LL} = \frac{1}{4} \cdot \mathbf{I}_{src} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} & \mathbf{I}_{LH} = \frac{1}{4} \cdot \mathbf{I}_{src} \cdot \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix} \\ \mathbf{I}_{HL} = \frac{1}{4} \cdot \mathbf{I}_{src} \cdot \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} & \mathbf{I}_{HH} = \frac{1}{4} \cdot \mathbf{I}_{src} \cdot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{cases} \tag{4.7}$$

$$Synthesis \begin{cases} a_{rec} = \mathbf{I}_{DWT} \cdot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} & b_{rec} = \mathbf{I}_{DWT} \cdot \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \\ c_{rec} = \mathbf{I}_{DWT} \cdot \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} & d_{rec} = \mathbf{I}_{DWT} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{cases} \tag{4.8}$$

with $\mathbf{I}_{src}=\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $\mathbf{I}_{DWT}=\begin{bmatrix} I_{LL} & I_{LH} \\ I_{HL} & I_{HH} \end{bmatrix}$ and $\mathbf{I}_{rec}=\begin{bmatrix} a_{rec} & b_{rec} \\ c_{rec} & d_{rec} \end{bmatrix}$ 2x2 blocks of the input, transformed and reconstructed signals, respectively. The division by 2 originally performed at the synthesis stage has been shifted to the analysis stage to keep a 10-bit output compliant with the input format of the HEVC encoder.

For the Le Gall 5/3 DWT, the lifting scheme, introduced in Section 4.1, has been adapted to follow the modified DWT scheme, as depicted in **Fig. 4.9**. The filters composing the lifted Le Gall 5/3 wavelet transform are defined by Equations 4.9 and 4.10. The scaling operator used for prediction in the analysis stage and update in the synthesis stage is defined

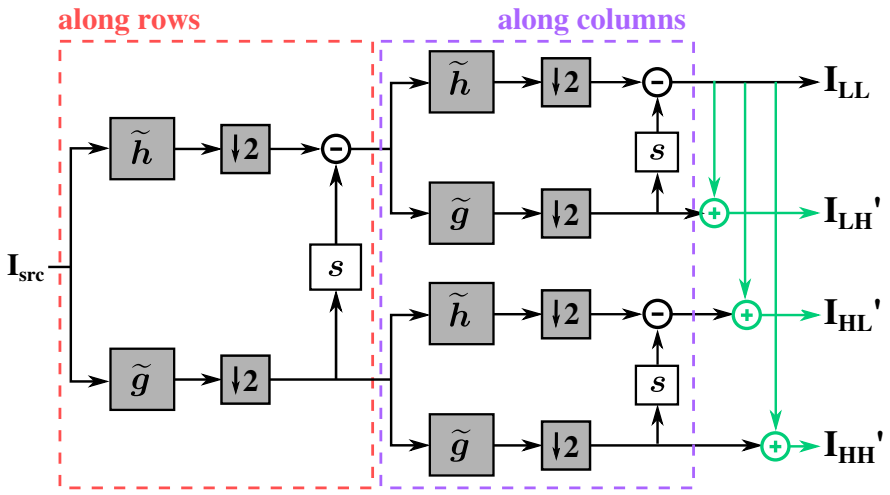


Fig. 4.9 Modified Discrete Wavelet Transform - Lifting scheme.

as $s(z) = -\frac{1}{4} \cdot \left(z^{-1} + 1\right)$.

$$\textit{Analysis} \begin{cases} \tilde{h}(z^{-1}) = 1 \\ \tilde{g}(z^{-1}) = z - \frac{1}{2} \cdot \left(z^2 + 1\right) \end{cases} \tag{4.9}$$

$$\textit{Synthesis} \begin{cases} h(z) = 1 + \frac{1}{2} \cdot \left(z^{-1} + z\right) \\ g(z) = z^{-1} \end{cases} \tag{4.10}$$

This implementation respects the perfect reconstruction property defined in Equations 4.1 and 4.2, with integer values.

4.3 Coding Configuration Modifications

This section presents the investigated experiments on coding configurations specifically designed for the encoding of the proposed wavelet transformed signal. The experimental results will be presented in Section 4.4.

4.3.1 Frequency-dependent Quantization Weighting

One of the main issues observed on the output of the proposed coding chain is a loss of definition due to the lack of information in the decoded high frequency sub-bands, necessary for a correct reconstruction. This lack of information is a consequence of the low amount of transform blocks kept by the encoder in the high frequency sub-bands. Indeed, in these images, the coefficients after prediction and forward transform have particularly low absolute magnitudes, generally smaller than 2 once quantized. Thus, the encoder often finds the

option of not signaling any coefficient a better solution, in a RDO sense, rather than signaling them in a transform unit.

This study aims at increasing the amount of information in the high frequency sub-bands after encoding by decreasing the quantization step for selected high frequencies, depending on the current sub-band, thus favoring the low magnitude transform coefficients during quantization. This can be achieved by applying different frequency-dependent quantization matrices for each sub-band, thus favoring the significant frequencies during quantization.

As stated in Section 2.3.2, the transform step of HEVC uses the two-dimensional DCT on TBs of different sizes, from 4x4 to 32x32 pixels. Therefore, since a wavelet transform splits the signal into four sub-bands, one with the low frequencies, the other with vertical, horizontal or diagonal frequencies, the transform coefficients should be concentrated around one quarter of the transform block, depending on the wavelet sub-band. This is the expected behavior if the chosen DWT has a good frequency splitting property.

For each sub-band obtained with the Haar DWT, **Fig. 4.10** depicts the energy of each coefficient location for the different transform blocks signaled in the bitstream for the coding of *CatRobot* sequence at QP 22. Black and white pixels represent the lowest and highest energies, respectively. As expected, coefficients with the highest energy are located in the quarter corresponding to the frequencies of their sub-band type. The only exception is for the large transform block sizes of the HH sub-band, where the energy of the DC coefficient is higher than the bottom right coefficients, which correspond to diagonal frequencies. With this information, different quantization matrices adapted to each sub-band can be designed with higher weights, i.e. a larger quantization step, for the non-significant frequencies.

As mentioned in Section 2.3.2, HEVC supports quantization weights for each transformed coefficient location (frequency), i.e. the $w(i, j)$ weights in Equation (2.14). These weights can be set for each possible TB size through quantization matrices signaled in the Picture Parameter Set (PPS), which defines several picture-level parameters. The PPS is usually the same for every frame of a bitstream and is thus only sent once for the first frame. However, it can eventually be sent again each time a modification of the PPS parameters is applied, for example to the quantization matrices.

Signaling the quantization matrices in the PPS of each frame - 24 matrices (4 transform block sizes, Y, U and V components, intra and inter prediction modes) - would introduce a non-negligible bitrate overhead. Instead, the scaling matrices are only signaled once for each sub-band type, storing them in a dedicated memory which is then accessed to retrieve the quantization matrices of subsequent frames. The association of a frame to a certain sub-band type from the encoder point of view, and thus to a corresponding set of quantization matrices, is simply done by looking at the POC of the frame. The HEVC reference software has thus been modified to include the signaling of the scaling matrices in the PPS with the necessary memory buffers needed to implement the scheme previously described.

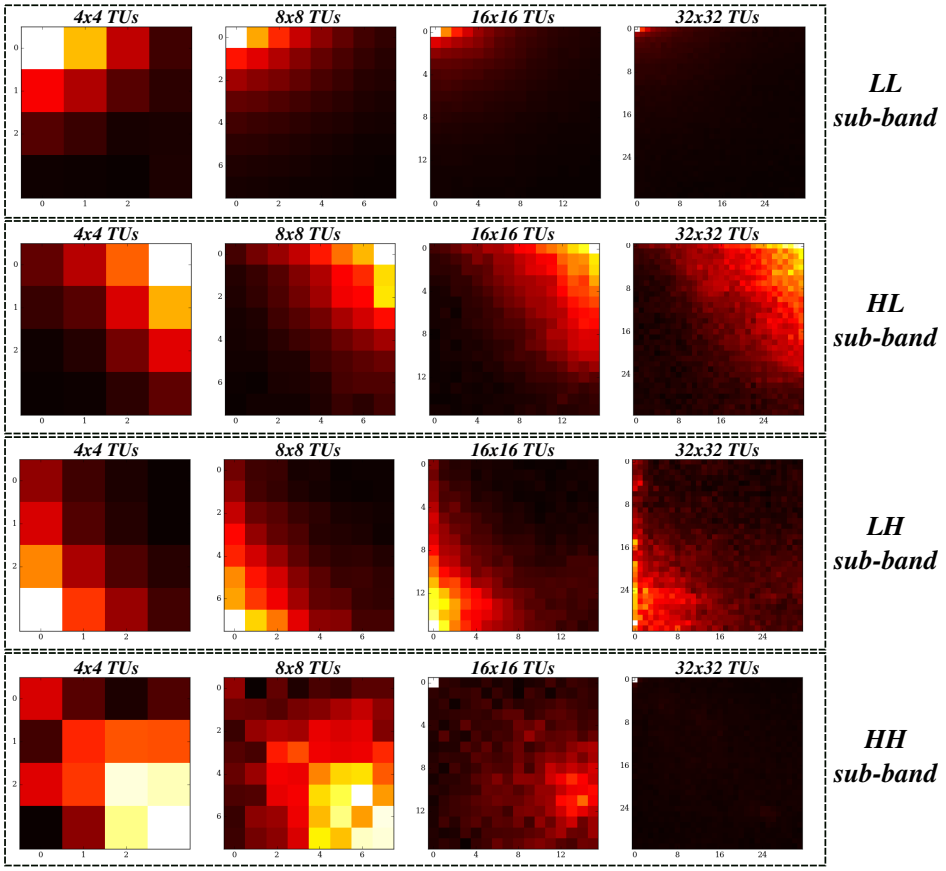


Fig. 4.10 Transform coefficients energy by position in TU and wavelet sub-band.

4.3.2 Optimized Bit Allocation for Non-Orthogonal Wavelets

Since the Le Gall 5/3 wavelet is bi-orthogonal, energy is not preserved throughout the transform process contrary to orthogonal wavelets such as Haar. This non-orthogonality results in a sub-optimal bit-allocation with usual RDO algorithms because each sub-band does not equally contribute to the reconstructed signal. Thus, errors in a particular sub-band lead to a larger output distortion than the same error in a different sub-band.

Usevitch proposed in [83] an optimal bit-allocation for bi-orthogonal wavelets based on a weighting of the quantization error depending on the sub-band filter coefficients. Assuming a two-level wavelet decomposition, the weights applied to the distortion of each sub-band follow Equation (4.11).

$$\omega_{ij} = tr\left((\, \mathbf{g}_j \otimes \mathbf{g}_i \,)^T (\, \mathbf{g}_j \otimes \mathbf{g}_i \,) \right) \tag{4.11}$$

with ω_{ij} the weight applied to the sub-band ij, $i,j \in \{L,H\}$, $\mathbf{g}_x$ the wavelet kernel, here $\mathbf{g}_L = \frac{1}{2} \cdot [1,2,1]$ and $\mathbf{g}_H = \frac{1}{8} \cdot [-1,-2,6,-2,-1]$ for the Le Gall 5/3 wavelet, and $\otimes$ the

Kronecker product. The obtained weights are as follows:

$$\omega_{LL} = 2.25 \qquad \omega_{LH} = 1.078 \qquad \omega_{HL} = 1.078 \qquad \omega_{HH} = 0.517$$

These weights have been transposed into the RDO stage of the HEVC reference software using the following cost function:

$$J = \frac{\omega_{SB}}{4} \cdot D + \lambda \cdot R \tag{4.12}$$

with J the cost function to minimize, D the distortion, R the rate in number of bits required to signal the coding parameters, λ the Lagrangian multiplier derived from the prediction mode and QP, ω_{SB} the weight applied to sub-band SB, $SB \in \{LL, LH, HL, HH\}$.

4.4 Experimental Results

This section provides a study on the performances of the modified DWT coding chain using the previously introduced UHD dataset. As for the polyphase case, an objective evaluation is performed first to compare the encoding scheme, including the proposed imrpovements, to SHVC. Then, the observations of a viewing session aiming at comparing visual artifacts on the outputs of the different coding chains are detailed.

4.4.1 Objective Evaluation

The objective evaluations presented in this section have been carried out using the same coding chain, test sequences and random access encoder configuration as the polyphase scalable coding scheme.

Results for Frequency-dependent Quantization Matrices

Encodings with different sets of quantization matrices have been performed to estimate the impact of the modifications on the coding efficiency. The shape of the tested quantization matrices corresponds to the transform coefficient energy repartition of **Fig. 4.10**, with weights higher than 16 - the normalizing factor - for frequencies with low average energy, in order to set a higher quantization step for the non-significant coefficients compared to the weights of the significant high frequencies.

No improvement in coding performance was observed using the UHD dataset. Indeed, looking at the generated bitstreams, the number of signaled transform coefficients is not higher when using the frequency-dependent quantization matrices, resulting in slightly negative or no impact on the output signal quality. This is mainly due to the fact that allocating a high quantization step to non-significant coefficients tends to result to the same

Table 4.1 BD-Rate results (%) for sub-band weighting vs original Le-Gall decomposition.

Sequence	Y	U	V	YUV
DaylightRoad	-4.61	-14.5	-11.6	-6.85
CatRobot	-5.43	-11.1	-6.9	-6.02
Drums	-3.07	3.74	5.89	-2.01
Tango	0.88	1.38	4.01	1.48
Average	**-3.05**	**-5.12**	**-2.15**	**-3.35**

rate-distortion cost and thus the same encoder decisions as with the original quantization step.

Results for Optimized Bit Allocation for Non-Orthogonal Wavelets

Table 4.1 summarizes the performance comparison for the Le Gall wavelet with and without the presented wavelet sub-band weighting. It can be observed that the proposed sub-band weighting provides average bitrate savings of 3.35 % for equal PSNR-YUV values. More particularly, the weighting shows a substantial increase in coding efficiency for the sequences that present significant losses, i.e. *DaylightRoad* and *CatRobot*.

Indeed, the obtained weights tend to limit the importance of errors in the high-pass sub-bands, thus favoring the low-pass sub-band during bit allocation. Thus, less bits will be allocated to the costly to encode high-pas coefficients due to their smaller contribution to the reconstructed output quality, hence the better results for sequences containing a high amount of spatial details.

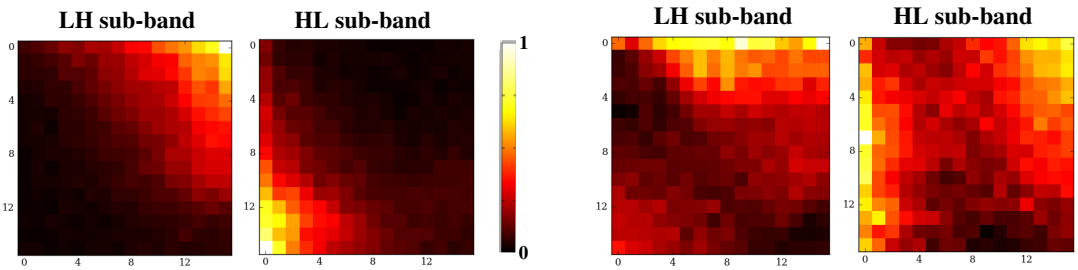
On the contrary, for the sequence *Tango*, which already has a low amount of encoded high-pass coefficients, the weighting does not have a clear impact on the final bitrate by further lowering the amount of transmitted high-pass coefficients, but only lowers the output quality, thus resulting in a positive BD-rate value.

Table 4.2 summarizes BD-Rate values for both Haar and Le Gall wavelet decompositions compared to SHVC. The optimized bit allocation for non-orthogonal wavelets has been activated for the Le Gall wavelet. For equal PSNR-YUV, an average bitrate overhead of 1.9% and 7.1% can be observed for Haar and Le Gall 5/3, respectively.

The results for both wavelet-based decompositions follow the same trend as for the polyphase subsampling but with less bitrate overhead. On average, it can be observed that the scheme based on the Haar wavelet performs better than the Le Gall 5/3 scheme. This is mainly due to the smaller energy present in the transformed coefficients in the high-pass sub-bands after the motion compensation stage. This is illustrated in **Fig. 4.11** which depicts the average energy in the 16x16 TU obtained by encoding the sequence *CatRobot* with a QP of 22. As can be observed, the Le Gall 5/3 scheme transformed residuals are less compact in the

Table 4.2 BD-Rate results (%) and complexity (%) for proposed schemes vs SHVC.

Sequence	**Haar**				**Le Gall 5/3**			
	Y	U	V	YUV	Y	U	V	YUV
DaylightRoad	22.9	45.5	71.2	30.0	36.3	67.3	92.4	44.0
CatRobot	7.5	39.2	55.0	13.4	16.6	53.4	71.6	23.3
Drums	-2.2	45.9	67.8	4.3	2.4	60.9	83.2	9.7
Tango	-24.8	12.1	10.5	-16.0	-23.4	16.1	16.3	-13.7
RollerCoaster	-12.8	13.9	17.8	-8.8	-15.6	10.2	13.6	-11.9
ToddlerFountain	-12.6	12.9	-4.1	-11.5	-11.4	56.4	25.9	-8.7
Average	**-3.7**	**28.3**	**36.4**	**1.9**	**0.8**	**44.1**	**50.5**	**7.1**
Encoding $TR_{\%}$	**54**				**53**			
Decoding $TR_{\%}$	**50**				**40**			



(a) Haar (b) Le Gall 5/3

Fig. 4.11 Normalized average transformed coefficients of 16x16 HEVC TUs for (a) Haar and (b) Le Gall wavelet-based schemes.

DCT domain compared to the Haar wavelet coefficients, especially in the *LH* and *HL* sub-bands. This greatly increasing the number of bits required to signal such transform blocks after the Rate Distortion Optimized Quantization (RDOQ)stage [84]. For the *HH* sub-band, the residual energy is mainly contained in the DC coefficient for both wavelet schemes but is much more diffused in the spatial frequency coefficients for the Le Gall 5/3 scheme. Thus, the number of transmitted TUs is globally higher for the Haar wavelet scheme due to their limited bit cost, resulting in a higher reconstruction quality for the same bitrate.

Encoding and decoding times are also summarized in Table 3.3, showing an average encoding (resp. decoding) complexity reduction of 54% and 53% (resp. 50% and 40%) for Haar and Le Gall 5/3 schemes, respectively, compared to SHVC.

4.4.2 Visualization of the Reconstructed Videos

An informal subjective viewing session with several experts has been organized in order to assess if the perceived quality conforms to the different objective evaluations performed.

Several test points have been selected among the encoded reduced dataset with the same bitrate for each coding chain under test. Overall, the differences in measured objective quality globally correspond to the differences in perceived quality.

For the *CatRobot* sequence, at relatively high bitrate ($\approx 15\ Mbps$), the SHVC encoded sequence shows a better conservation of details on the moving scarfs in the background and on the text in the foreground. For those particular areas, both wavelet-based encoded video clips do perform better than the polyphase scheme, with more readable text on the book pages.

For *DaylightRoad*, the same observations can be made, with more details present on the trees and cable lines with SHVC. With the Le Gall DWT, some very sharp details, such as an antenna at the top of a building, were partly missing from the reconstructed output video, probably due to the high amount of high frequency information in this area, which was filtered out during encoding due to its important coding cost.

For *Drums* and *Tango*, the differences between the reconstructed outputs of the proposed wavelet-based scheme and the SHVC are hardly noticeable at high bitrates, which concurs with the highly similar objective measures at these bitrates.

4.5 Conclusion

In this chapter, a modification to the pre-processing based scalable coding scheme presented in Chapter 3 have been proposed. Indeed, the polyphase decomposition has been replaced by a modified wavelet-based decomposition.

First, the proposed wavelet-based decomposition has been introduced, with a description of the modification of the conventional wavelet transform necessary to make the transformed signal suitable for the considered scalable coding scheme. Then, the chosen implementation of the two considered wavelets filter banks has been presented.

Then, two modifications of the encoder configuration aiming at adapting the encoding process to the format of the wavelet transformed signal have been proposed. It has been shown that the first explored experiment, based on frequency-dependent quantization matrices, did not yield any improvements over a legacy HEVC coding scheme, in terms of bitrate savings. The other proposed improvement consists in optimizing the bit allocation of the RDO process for non-energy preserving non-orthogonal wavelet such as the Le Gall wavelet. It has been shown that this optimization leads to 3.35% average bitrate savings compared to the regular HEVC RDO process. Compared to SHVC, the proposed coding scheme achieved average bitrate overheads of 1.9% and 7.1% for the Haar and Le Gall 5/3 wavelets, respectively, while offering a greater than 50% reduced coding complexity.

However, as for the polyphase based decomposition, despite interesting coding performance for the Haar wavelet, the proposed solution efficiency is highly dependent on the

content being encoded, with notably significant losses for videos with a high amount of spatial details, which is problematic for broadcast and streaming use-cases. The next part of this thesis will thus focus on resolving this issue while achieving further complexity reduction.

Part IV

Dual-layer Low-complexity Scalable Encoder based on Adaptive Spatial Resolution

Chapter 5

Local Adaptation of the Spatial Resolution through Inter-Layer Predictions

Despite the interesting performance of the low-complexity scalable solutions based on pre- and post-processing tools presented in Chapters 3 and 4, they show a limited coding efficiency for content with high and complex spatial details due to the systematic application of the proposed signal decompositions. This is particularly a concern for the UHDTV use-case considered in this thesis.

To avoid this issue, a different way of achieving scalability with little complexity overhead, compared to a single-layer scheme, would be to use a dual-layer scalable encoder with a very lightweight EL encoder and a standard HEVC encoder in the BL to still ensure backward compatibility with HDTV services.

In addition, although the increase in spatial resolution to 3840x2160 pixels, which has been the first dimension of the UHD format to be deployed due to the relatively small changes needed in the production pipeline, offers increased perceptual quality of natural images [4], the visual gain is highly dependent on the content under consideration. This is especially the case if encoding is part of the processing chain [85].

Therefore, in this chapter, a spatially scalable encoder based on a dual-layer architecture involving adaptive spatial resolution is proposed to significantly reduce the complexity of the encoding process. Indeed, the idea is to dynamically reduce the resolution when the UHD resolution is not needed to portray the details of an image while keeping the high resolution when it is justified. By doing this, only useful processing is performed the core encoding process.

This chapter is organized as follows. Section 5.1 provides an overview of the related works on downsampling-based video coding, upsampling and super-resolution algorithms as well as SHVC complexity reduction solutions. The proposed scalable architecture is de-

Fig. 5.1 General architecture of downsampling-based encoding.

scribed in Section 5.2 followed by the detailed presentation of the adaptive spatial resolution algorithm in Section 5.3. A study on the spatial resolution mode distribution is given in Section 5.4. Finally, the experimental results are presented in Section 5.5 and Section 5.6 concludes this chapter.

5.1 Related Work

5.1.1 Downsampling-based Video Coding

Despite the constant effort to increase the capture and display devices spatial resolutions further and further, such resolutions are not necessarily required to represent the critical information of the signal. A well-known strategy consists in downsampling the video prior to encoding and upsampling back to the original resolution after decoding in an effort to reduce the bitrate and/or the coding complexity for a similar reconstructed output quality [86–98]. Adaptive Spatial Resolution (ASR) can be performed at different levels - whole sequence, image or block - in the downsampling-based encoding scheme depicted in **Fig. 5.1**.

5.1.1.1 Sequence-level ASR

Sequence-level downsampling-based encoding is investigated in [87–91]. In [87], authors extend the theory developed by Bruckstein *et al.* in [86] for still image downsampling-based image coding, showing that downsampling a video sequence prior to encoding can improve the coding efficiency compared to conventional AVC full frame coding schemes at low bitrates. In [88], authors consider the downsampling based coding scheme as a joint optimization of rate, distortion and complexity. They design a low-complexity Super Resolution (SR) algorithm which is systematically applied to each frame, achieving better performances than HEVC at low-bitrates, in addition to a significant reduction of the overall complexity. Shen *et al.* [89] use systematic downsampling of inter-coded frames using a specific SR algorithm trained online, while the intra-coded frames are coded at full resolution. This scheme is shown to outperform AVC coding at full resolution for low to medium range bitrates. Dong *et al.* [91] design a model to estimate the downsampling and coding errors, separately, at a given target bitrate to find the optimal downsampling ratio to encode a given video sequence. This optimal ratio $M^\star$ is obtained by minimizing the sum of variances of

the downsampling error σ_D^2 and coding error σ_C^2

$$M^\star = \arg\min_M(\sigma_D^2 + \sigma_C^2). \tag{5.1}$$

The model is shown to improve the coding efficiency compared to conventional coding of HD videos with AVC.

5.1.1.2 Image-level ASR

Despite showing interesting coding efficiency, the sequence-level downsampling-based coding schemes do not take into account the possibly high variance in spatial details between the different frames of a video sequence, thus requiring different downsampling ratios. Wang *et al.* [92] investigate adaptive downsampling at a GOP-level (1-second long). They design spatially scalable R-D models to estimate the downsampling and coding errors and find the best encoding frame-size for a given bitrate, demonstrating a better overall coding efficiency compared to AVC. Further resolution adaptation, at a frame-level, has also been investigated by Afonso *et al.* [93] in which a resolution-quantization optimization module is trained to find a QP threshold QP_{thresh} for which the downsampling of a given image by a factor 2 becomes more efficient than a full-resolution encoding in a R-D sense. For each frame to be encoded, QP_{thresh} is first derived from the downsampling error and then used, together with the input QP, QP_{in}, to select the downsampling factor M and the final QP value QP_{enc} using Equations 5.2 and 5.3

$$M = \begin{cases} 2 & if \quad QP_{in} \geq QP_{thresh} \\ 1 & otherwise \end{cases}, \tag{5.2}$$

$$QP_{enc} = \begin{cases} QP_{in} - K & if \quad QP_{in} \geq QP_{thresh} \\ QP_{in} & otherwise \end{cases}, \tag{5.3}$$

with K a QP adjustment value to account for the lower quantization step needed to achieve the same bitrate at a lower spatial resolution compared to a full resolution encoding. This image-level adaptive downsampling outperforms HEVC in AI configuration at low bitrates. An extension of this work is proposed in [94] where an additional resolution, $M = 1.5$, is considered and the Lanczos3 upsampling is replaced by a Convolutional Neural Network (CNN) based SR algorithm, leading to a better coding efficiency than HEVC in RA configuration, at the cost of a greatly increased complexity due to the use of CNN.

5.1.1.3 Block-level ASR

Since the spatial frequency components can also be highly variable within different areas comprised in a single natural image, several studies have investigated the local adaptation of the spatial resolution, at a block level [95–98]. Lin and Dong [95] have first introduced

the local ASR technique for still image compression using a modified JPEG encoder, allowing the blocks to be coding in four different possible resolutions. For each 16x16 block, an exhaustive search RDO process is used to find the spatial resolution resulting in the best reconstructed quality while not increasing the bit budget compared to the full resolution coding of the block. The bit budget is estimated by counting the number of non-zero quantized DCT coefficients for each possible resolution. This work is extended to video in [96], where the scheme is integrated in a MPEG-2 encoder, demonstrating a better coding efficiency at low bitrates for the local ASR scheme compared to regular MPEG-2 coding. More recently, a CTU-level resolution adaptation scheme [97, 98] has been integrated in the HEVC reference software, allowing for the entire CTU coding process (quadtree partitioning, prediction, transform and quantization) to be performed at a reduced spatial resolution when it leads to R-D gains. The upsampling process to reconstruct the full resolution is chosen between traditional filter bank or a CNN-based SR algorithm. This solution offers significant coding gains compared to regular HEVC coding but also multiplying encoding and decoding times by 3 and 32, respectively.

5.1.2 Upsampling Filter banks and Super-Resolution

In downsampling-based video coding schemes, as well as in the scalable encoder proposed in this chapter, image resizing techniques are usually used to upsample the video to a desired spatial resolution.

Image resizing has been an active field of research for several decades, from the design of the well-known and widely used simple interpolation filter kernels [99–101] to the development of state-of-the-art resizing algorithms based on super-resolution [102–119], which is currently a trending research topic [120–123] through the recent advances in deep learning applied to image processing.

Common Interpolation Filter Banks

Image resampling is generally achieved via traditional sample interpolation using common filter banks. The most simple interpolation algorithms are nearest-neighbor and bilinear interpolation methods, which respectively consist in copying the nearest sample and linearly interpolating from the pixels of the nearest 2x2 neighborhood using Equation (5.4) to obtain an upsampled output sample. These two interpolation methods lead to relatively poor upsampled image quality, with either jagged artifacts due to aliasing for the nearest-neighbor interpolation, or smoother edges for the bilinear interpolation.

$$\begin{aligned}\mathbf{I}_{up}(x,y) &= (1-dx)(1-dy)\cdot\mathbf{I}_{src}(i,j)+dx(1-dy)\cdot\mathbf{I}_{src}(i+1,j)\\ &\quad+(1-dx)dy\cdot\mathbf{I}_{src}(i,j+1)+dxdy\cdot\mathbf{I}_{src}(i+1,j+1),\end{aligned} \tag{5.4}$$

where

$$\begin{aligned} i &= \lfloor x \rfloor, \qquad dx = x - i, \\ j &= \lfloor y \rfloor, \qquad dy = y - j, \end{aligned}$$

with (x,y) the coordinates of the pixel in the upsampled image $\mathbf{I}_{up}$ to be interpolated from the source image $\mathbf{I}_{src}$ and $\lfloor . \rfloor$ the flooring operator.

The most widely-used interpolation techniques are the bicubic [99] and Lanczos [100] filters. Bicubic interpolation is achieved by convolving, successively in both directions, the input image $\mathbf{I}_{src}$ by a filter kernel $\mathbf{f}$ of length 4 whose coefficients are derived from piecewise cubic polynomials, as defined in Equation (5.5)

$$\mathbf{I}_{up}(x,y) = \sum_{m=-1}^{2} \sum_{n=-1}^{2} \mathbf{I}_{src}(i+m, j+n) \cdot \mathbf{f}(m-dx) \cdot \mathbf{f}(n-dy), \tag{5.5}$$

with

$$\begin{matrix} i & = \lfloor x \rfloor, \, dx = x - i, \\ j & = \lfloor y \rfloor, \, dy = y - j, \end{matrix} \text{ and } \mathbf{f}(s) = \begin{cases} \frac{3}{2}|s|^3 - \frac{5}{2}|s|^2 + 1 & 0 < |s| < 1 \\ -\frac{1}{2}|s|^3 + \frac{5}{2}|s|^2 - 4|s| + 2 & 1 < |s| < 2 \\ 0 & 2 < |s| \end{cases},$$

Bicubic interpolation thus uses a 4x4 neighborhood, which results in more visually pleasing upsampled images with sharper edges, at the expanse of a slightly increased computation time compared to bilinear interpolation.

The Lanczos filter bank [100] is arguably the best performing kernel-based interpolation method [124, 125], offering a good trade-off between sharpness, reduction of aliasing and limited ringing artifacts. It is defined as a Lanczos-windowed sinc function. Equation (5.6) defines the decomposed 2D convolution of the input image $\mathbf{I}_{src}$ by a filter kernel $\mathbf{f}_a$ whose length is dependent on the number of lobes a kept by the Lanczos-window

$$\mathbf{I}_{up}(x,y) = \sum_{m=-a+1}^{a} \sum_{n=-a+1}^{a} \mathbf{I}_{src}(i+m, j+n) \cdot \mathbf{f}_a(m-dx) \cdot \mathbf{f}_a(n-dy), \tag{5.6}$$

with

$$\begin{matrix} i & = \lfloor x \rfloor, \, dx = x - i, \\ j & = \lfloor y \rfloor, \, dy = y - j, \end{matrix} \text{ and } \mathbf{f}_a(s) = \begin{cases} \frac{\sin \pi s}{\pi s} \cdot \frac{\sin \pi \frac{s}{a}}{\pi \frac{s}{a}} & 0 < |s| < a \\ 0 & a < |s| \end{cases},$$

The most common implementation of the Lanczos uses a 6-tap kernel, i.e. $a = 3$, since it offers the best reconstruction performance [125] while keeping a reasonable kernel complexity.

More recently, a new DCT-based Interpolation Filter (DCTIF) filter bank, based on the theory developed in [101], has been designed by Alshin *et al.* [34, 126] for the HEVC coding standard. It is used to interpolate pixels to enable half and quarter pixel motion estimation and are described in Section 2.3.2. This filter bank offers a good compromise between visual quality and implementation complexity. To achieve this, longer filters - 7/8 taps for luma and 4 taps for chroma -, with integer implementation and designed to preserve natural high-frequencies present in videos, are used.

SR based on Traditional Signal-processing Tools

The first SR algorithms [102–113] rely on signal processing techniques to produce a High Resolution (HR) image from several Low Resolution (LR) images [127]. Typically, the different LR images represent different "representations" of the same scene, with sub-pixels shifts between one another to enable the HR image generation. They can either be obtained by several cameras capturing the same scene, or from several images of a video sequence where the motion between these images can be controlled or estimated at a sub-pixel precision. Then, from these LR images, different approaches have been designed to generate the HR image.

The non-uniform interpolation approach [102–104], is intuitive and consists in three steps performed successively: registration or motion estimation to obtain the sub-pixel shifts between the different LR images, followed by non-uniform interpolation to generate the HR image and finally a deblurring process to improve the quality of the output image. The frequency domain approach [105–107] uses the aliasing present in the LR images, together with the continuous Fourier transform of an HR image and the discrete Fourier transform of the LR images, to perform the resolution enhancement. The Bayesian Stochastic approach, through a maximum a posteriori method [108, 109], has been designed to tackle the generally ill-posed problem of SR reconstruction, due to the insufficient number of LR images, resulting in robust and flexible SR models. The projection onto convex sets approach [110, 111] utilize prior knowledge of the solution during the reconstruction process, enabling a simultaneous solving of the registration and interpolation stages. The iterative back-projection approach [112, 113] estimates the HR image by iteratively back projecting the error between observed LR images and simulated LR images until the energy of the simulated error is minimized. For an in depth overview of these SR reconstruction techniques, the reader is referred to [128] and [129].

The main issues of these early SR algorithms are the need for several LR images to produce a single HR image and, depending on the considered approach, the non-uniqueness of the solution, the high computational cost or the possible constraints on the LR images. In addition, acceptable reconstruction quality is generally only achieved for scaling factors smaller than 2 with these SR algorithms, which is not optimal for most resizing-based image

processing applications.

Learning-based SR Algorithms

A second category of SR algorithms has been designed using machine learning techniques. The example-based SR approach is first introduced by Freeman *et al.* [114], using a Markov random field to learn the prediction of a HR image from a single LR image. In practice, the Markov network is trained to minimize the error between a HR image a linearly upscaled version of the matching LR image, i.e. to recover the missing high frequencies in the upsampled LR image. This is done by finding the nearest neighbor patch in the LR space thus requiring a very large database of HR image patches and their matching downsampled LR version to correctly generalize the SR problem. The method is extended in [115], with a faster training phase and an interpolation method changed from linear to bicubic, thus reducing the amount of details needed to be recovered by the Markov network. Chang *et al.* [116] have adopted a different approach by using manifold learning, thus considering the similarities between the manifold in the HR patch space and the one in LR patch space to predict the HR output. Since the method uses a combination of several nearest neighbor training patches to generate the HR output, a smaller database is required to achieve good generalization of the SR problem.

Motivated by the advances in the compressive sensing field, Yang *et al.* [117, 118] propose to use a sparse representation of the image patches to recover the HR patch from its LR counterpart. Two compact dictionaries, one to represent the HR patches and the other to represent the LR ones, are jointly trained to ensure coherence between the sparse representations of the patch pairs. The SR process thus consists in first computing the sparse representation of a LR image patch using the relevant patch bases in the trained LR dictionary, and the corresponding sparse representation in the trained HR dictionary is then used to recover the HR output patch. This work is further extended in [119], with an improved coupled dictionary learning method and a faster LR sparse representation computation using a trained neural network. The resulting images are shown to be very realistic with sharp edges, which was one of the issues of the previous SR algorithms.

Neural Networks for SR

More recently, following the current research trend in applied artificial intelligence, CNN-based single image and video SR algorithms have been designed [120–123]. SRCNN, the first end-to-end SR algorithm solely based on a deep convolutional neural network is proposed by Dong *et al.* in [120]. The CNN takes the LR image, previously upsampled to the output resolution using bicubic interpolation, as input and generates the HR output using relatively small network architecture only consisting of three convolutional layers. The first layer extracts feature maps from the LR input image, followed by a second layer that

maps the resulting LR feature maps to the HR space. The final convolutional layer is used to produce the final HR image from the HR feature maps. The number of convolution kernels and their size can be set for each layer and optimized to obtain the desired trade-off between inference time and reconstruction quality. Despite its relative simplicity, SRCNN is shown to outperform the solutions based on sparse representations. Shi *et al.* [121] propose a faster CNN-based SR algorithm, called ESPCN, by keeping all convolutional layers in the LR space, thus not performing the pre-processing interpolation operation, except for the last layer which aggregates the LR feature maps and generates the HR output in a single step. With this architecture, the inference time is reduced by a factor 10 while keeping a reconstruction quality higher than SRCNN.

Both these single image SR algorithms have been extended and optimized for video SR in [122] and [123], respectively. VSRnet [122] takes the same base architecture as SRCNN but extends it to take as input several adjacent frames from a video sequence. These frames are processed independently by the convolution layers until a concatenation, which can happen at different levels of depth in the network, is done to aggregate the adjacent frames before convolution. This network architecture leads to better reconstruction quality compared to single image SR algorithms, in addition to a more temporally consistent output. It is also shown that performing a motion compensation step on the adjacent frame prior to the CNN can lead to better reconstruction quality. Caballero *et al.* [123] also propose an extension of ESPCN for videos, using slow temporal fusion of adjacent frames and spatial transformer motion compensation in addition to the sub-pixel convolution scheme used by ESPCN to achieve a better reconstruction quality than VSRnet while keeping a relatively low computation cost.

A significant number of other CNN-based, and more generally deep learning based, SR algorithms have been recently proposed, including [130–132] to name a few. However, they utilize more complex architectures with deeper networks compared to SRCNN, ESPCN and their video extension, which are already too computationally demanding to be integrated in a low-complexity downsampling-based coding scheme, i.e. the scope of this chapter.

5.1.3 SHVC Complexity Reduction

As previously explained in Section 2.3.3, since the early stages of the HEVC scalable extension development, the main drawback of SHVC is its computational complexity. In an effort to solve this issue, a number of encoding complexity reduction algorithms has been proposed over the last few years [133–146]. These algorithms are usually specific to one of the scalability types supported by SHVC, mainly focusing on either quality/fidelity scalability or spatial scalability, i.e. the most popular scalability use-cases.

Quality/fidelity Scalability

Since the BL and EL are highly correlated in the case of quality scalability - same input video, different QP - the straightforward idea of exploiting the encoding decisions made by the BL encoder to speed up the EL coding has been widely studied [133–140, 147], yielding in a significant EL coding complexity reduction - around $55-65\%$ in average - with a limited impact on the output quality - between 1% and 4% average bitrate overhead.

Tohidypour *et al.* [133] propose a content-dependent complexity reduction algorithm, based on two algorithms. The first one is an adaptive search range of the EL motion vectors, which reduces the motion estimation search range based on the type of motion identified in the co-located BL block. The second algorithm is an early termination scheme that ends the RDO process of the EL encoder once the estimated R-D cost, obtained using the R-D costs of the BL co-located block and the already coded neighboring EL ones, is reached.

In [134], Bailleul *et al.* propose a fast mode decision algorithm based on a statistical analysis of both SHVC EL and BL encoding decisions. The EL RDO process is restricted to use the same CU depth for both layers and certain prediction modes are skipped depending on the decisions made by the BL encoder. In [135], authors also propose to use an offline statistical analysis of the mode decisions of both BL and EL encoders, focusing mainly on the CU depth of the BL to define prediction rules for the depth of the co-located EL CU thus greatly reducing the amount of tested partitioning modes during the RDO stage.

Wang *et al.* [136] design CTU quadtree prediction algorithms for both the BL and EL encoders. The algorithm for the BL relies on the decisions of the previously coded spatial neighboring CUs and the co-located temporal CU which are each tested for the BL quadtree partitioning in a RDO manner. If the R-D cost of the chosen mode does not meet a QP-dependent constraint, the original exhaustive RDO search is performed. For the EL, a quadtree prediction algorithm is proposed, where the co-located BL CU and spatial neighboring CUs are used to predict the final decision, without further pruning to maximize the coding time reduction.

Following the same idea, the algorithm proposed in [137] uses a weighted prediction of the depth of spatio-temporal neighboring CUs to predict the optimal CU depth which is used to reduce the number of tested depth levels in the RDO process. If not enough neighboring blocks are available, the search range is set based on the depth of the co-located BL CU. The EL encoding complexity is thus greatly reduced due to the lighter RDO stage.

Chiang *et al.* [138] extend this idea to reduce the complexity of the RDO determination of the EL PU mode. To achieve this, a block motion complexity index is first computed using the PU modes of spatio-temporal neighboring blocks and co-located BL PU. Based on this index, the algorithm derives the PU modes to exclude from the RDO process, thus achieving an interesting complexity reduction. Recently, a similar fast PU decision method is coupled to fast CU depth decision and early termination in [139], showing significant complexity

gains for a small bitrate overhead.

In [140], Tohidypour *et al.* introduce a machine-learning approach to reduce the complexity of SHVC, capable of predicting the prediction mode of an EL PU. It incorporates a Naive Bayes classifier to find the four most probable modes and their corresponding probability, which are then used to greatly reduce the amount of tested modes in the RDO process. To achieve a better prediction, the ML model is refined online, i.e. its parameters are dynamically updated after processing a sample.

Spatial scalability

Despite the difference in spatial resolution, the base and enhancement layers of a spatial scalability scheme are well correlated, especially if a simple kernel-based upsampling operation on the BL can recover most of the high frequencies of the EL. Thus, researchers also investigated fast SHVC encoding using the BL encoder decisions to speed up the EL processing [141–146].

Zuo *et al.* [141] propose a fast CU depth and intra prediction mode decision algorithm based on the coding information from the BL. A statistical analysis is first carried out to study the correlation between both layers, showing that, depending on the BL CU depth, certain CU depths are rarely used by the EL encoder, thus indicating a possible skip of these modes in the EL RDO process. The study also showed that if the BL encoder uses an angular intra prediction mode, the EL encoder has a high probability to do the same, with a neighboring angular intra predictor. A fast encoding algorithm taking advantage of these specificities is thus designed, leading to an average complexity reduction of 29% for an equal coding efficiency in AI configuration.

In [142], authors reduce the motion estimation RDO process by reducing the motion vector search range in the EL encoder based on the average MV amplitude in the BL encoded bitstream, showing a 28% encoding time reduction with equivalent output R-D performance compared to SHVC reference software. Lu *et al.* [144] propose a complexity reduction of both CU and PU mode decision algorithms offering a 36% complexity reduction over SHVC by using several tools: an early termination of the CU depth determination process based on texture classification and co-located BL CU mode; a reduction of the intra mode prediction search relying on inter-layer, spatial and temporal correlations to remove unlikely intra modes from the RDO process; an inter PU prediction mode decision using the motion complexity and inter-layer dependency to determine the best mode. Shen *et al.* [146] also combine fast CU and PU mode decision algorithms to further reduce the encoding time, leading to average EL encoder complexity reduction of 69% for a coding loss of 2.3%. Wali *et al.* [145] add a skip mode inheritance algorithm from BL to EL to the common fast CU depth and PU size determination algorithms in order to achieve an overall complexity reduction of spatial scalability with an average coding time reduction of 45% for a bitrate

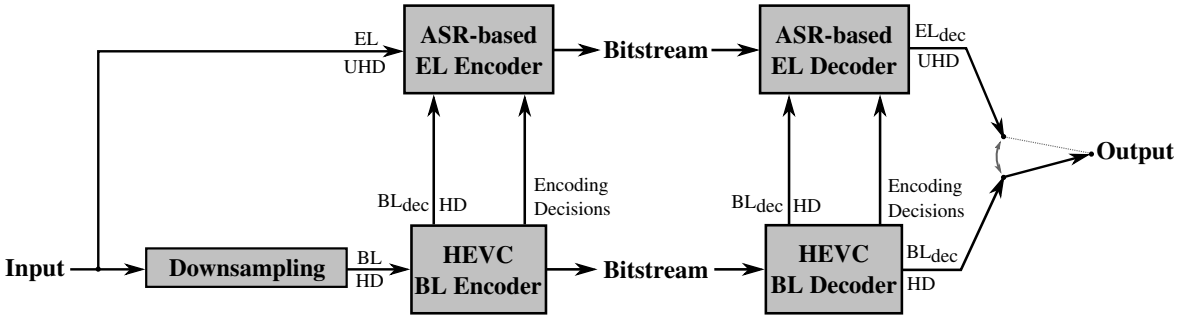


Fig. 5.2 Block diagram of the proposed dual-layer scalable encoder with ASR-based enhancement layer encoder.

overhead of 1.3%.

It is also possible to address both the quality and spatial scalability coding schemes with a single coding complexity reduction algorithm. Tohidypour *et al.* [143] design a probabilistic approach based on machine learning to predict the quadtree partitioning of the enhancement layer. The learned model is a Bayesian classifier using the co-located CTUs in both the corresponding BL frame and the temporally neighboring EL frames as well as spatially neighboring CTUs of the considered EL frame as input. The model offers an average EL encoder complexity reduction of 77% (resp. 79%) with a coding loss of 4.2% (resp. 2.2%) for spatial (resp. quality) scalability scheme compared to the original SHVC reference software in RA configuration.

5.2 Scalable Scheme with Specific Enhancement Layer Encoder

5.2.1 High-level Description

The scalable architecture under consideration, based on the one used in SHVC, is depicted in **Fig. 5.2**. On one hand, the input UHD video is first downsampled to HD resolution and then fed to the base-layer encoder, a standard HEVC encoder. On the other hand, the input UHD video is fed to the enhancement layer encoder, which is based on adaptive spatial resolution and also includes several HEVC features.

Both bitstreams are then processed by the decoder of their respective layer, the EL bitstream only being decoded if the desired output is the higher resolution video, e.g. UHD in our use-case.

Inter-layer communications are enabled in this scalable coding scheme to facilitate the processing of the UHD input video. On one hand, several encoding decisions (block partitioning, prediction mode, motion vectors), made by the BL encoder, are transmitted to the

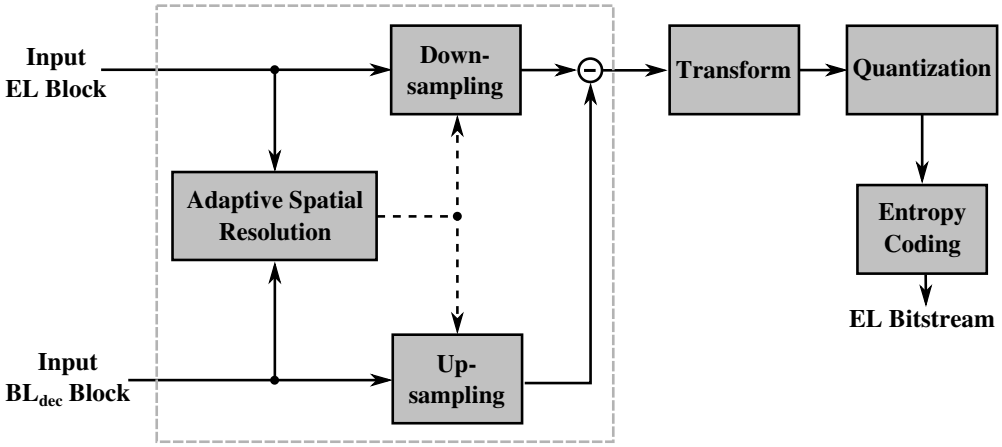


Fig. 5.3 Block diagram of the ASR-based enhancement layer encoder.

EL encoder. On the other hand, the decoded BL frames are also fed to the EL encoder to be used as reference, .

5.2.2 HEVC Base Layer Encoder

As already stated, the base layer encoder is a standard HEVC encoder in order to ensure a backward compatibility with existing infrastructure. The chosen implementation of the HEVC standard is the HM-16.12 encoder [41]. In the different experiments carried out in this study, AI and RA configurations have been used for the BL encoding, both following the JCT-VC common test conditions, with a 1 second intra period and a GOP size of 16 frames for the RA configuration.

AI encoding has been used to assess the performance of the scalable coding scheme with only the decoded BL as reference input to the EL encoder, as will be detailed in Section 5.5.

5.2.3 ASR-based Enhancement Layer Encoder

The enhancement layer encoder relies on several tools to achieve low-complexity scalable coding. The architecture notably uses an inter-layer ASR-based prediction algorithm. It is thus called ASR-IL in the rest of this dissertation. The architecture of the ASR-IL EL encoder is depicted in **Fig. 5.3**. First, the input EL source block and the corresponding decoded BL block are fed to the ASR-IL algorithm, which adapts the spatial resolution based on several factors, as will be detailed in Section 5.3. Once a decision is taken, the decoded BL block (resp. input EL block) is upsampled (resp. downsampled) to the chosen spatial resolution for further processing. The downsampling filters are the same as those used to generate the base layer input, i.e. the same as in the SHVC reference software, which are described in Section 2.3.3. The chosen upsampling filters are identical to the filters used for the fractional pixel interpolation in HEVC, described in Section 2.3.2.

Both signals now having the same resolution, an inter-layer prediction step can be performed on the EL signal with the BL signal as reference. Then, several tools taken from the HM reference software are added to the coding chain, namely the transform step - only the DCT-II implementation - the uniform quantization stage and finally the entropy coding process for transformed coefficients using CABAC. These steps will be further explained in the next section.

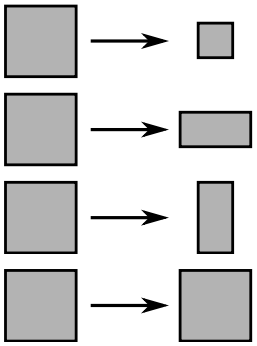
5.3 Adaptive Spatial Resolution

5.3.1 Description of the Technique

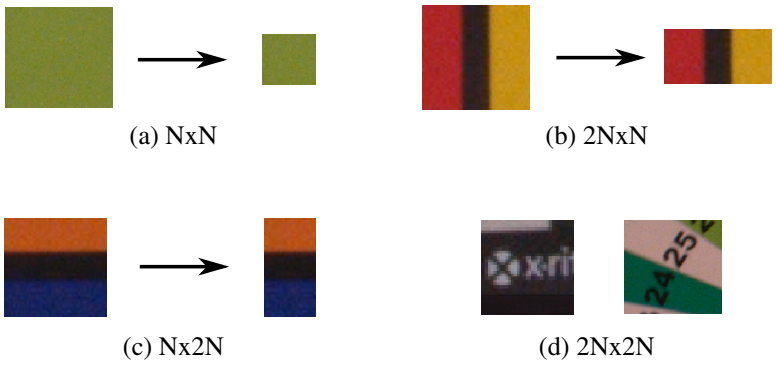
The ASR-IL scheme is at the core of the proposed low-complexity enhancement layer encoder. The idea is that, based on the assumption that improvements on visual quality brought by the UHD format over HD resolution is content dependent, the spatial resolution can be locally reduced with limited to no impact on the perceptual quality. Indeed, for image areas with certain characteristics, such as homogeneous content or smooth edges for example, coding at full resolution might not be beneficial in terms of visual quality compared to a coding step performed at lower resolution followed by an upsampling to the original resolution. On the contrary, for such blocks, the bit cost and complexity could be higher for the full resolution encoding case.

The objective of this study is to assess this hypothesis by designing an adaptive spatial resolution coding scheme that would adapt the resolution, at a block level, depending on the block content and the quantization parameter. The chosen four possible resolution choices, achieved by successive 1D downsampling of the EL input block, are as follows, with N the width of the square block corresponding to the BL resolution, and $W \times H$ designing an EL block chosen resolution of width W and height H:

- $N \times N$: Downsampling in both directions.
- $2N \times N$: Downsampling in vertical direction.
- $N \times 2N$: Downsampling in horizontal directions.
- $2N \times 2N$: No downsampling.



Both rectangular resolutions, $2N \times N$ and $N \times 2N$, are included in the possible spatial resolutions for EL blocks containing strong vertical or horizontal edges respectively, as depicted in Figures 5.4b and 5.4c, which are common in natural content. Figures 5.4a and 5.4d also respectively show intended typical EL block content for $N \times N$ - homogeneous areas or smooth edges - and $2N \times 2N$ - highly detailed areas or sharp edges in several directions - spatial reso-



(a) NxN (b) 2NxN

(c) Nx2N (d) 2Nx2N

Fig. 5.4 Typical EL blocks for each resolution of the adaptive spatial resolution scheme.

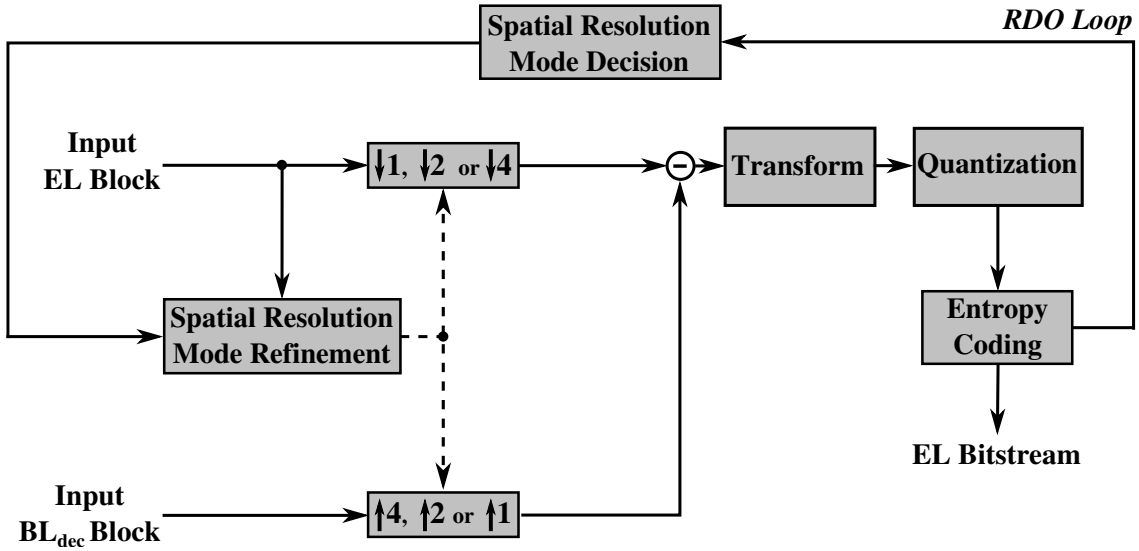


Fig. 5.5 El encoder architecture for RDO based resolution selection.

lutions.

5.3.2 Implementation

To implement the ASR-IL into the proposed low-complexity scalable coding scheme, the chosen resolution mode selection algorithm performs a rate-distortion optimization on the possible choices, as depicted in **Fig. 5.5**. Thus, each possible resolution is tested to select the best one in a RDO sense.

Therefore, the following EL encoding process is performed for each resolution successively and independently. First, the input EL and corresponding decoded BL blocks are respectively downsampled and upsampled to the chosen resolution, according to the scheme depicted in **Fig. 5.6**, to realize a simple prediction step. Then, the resulting prediction residuals are transformed and quantized. At this stage, on one hand the quantized coefficients are

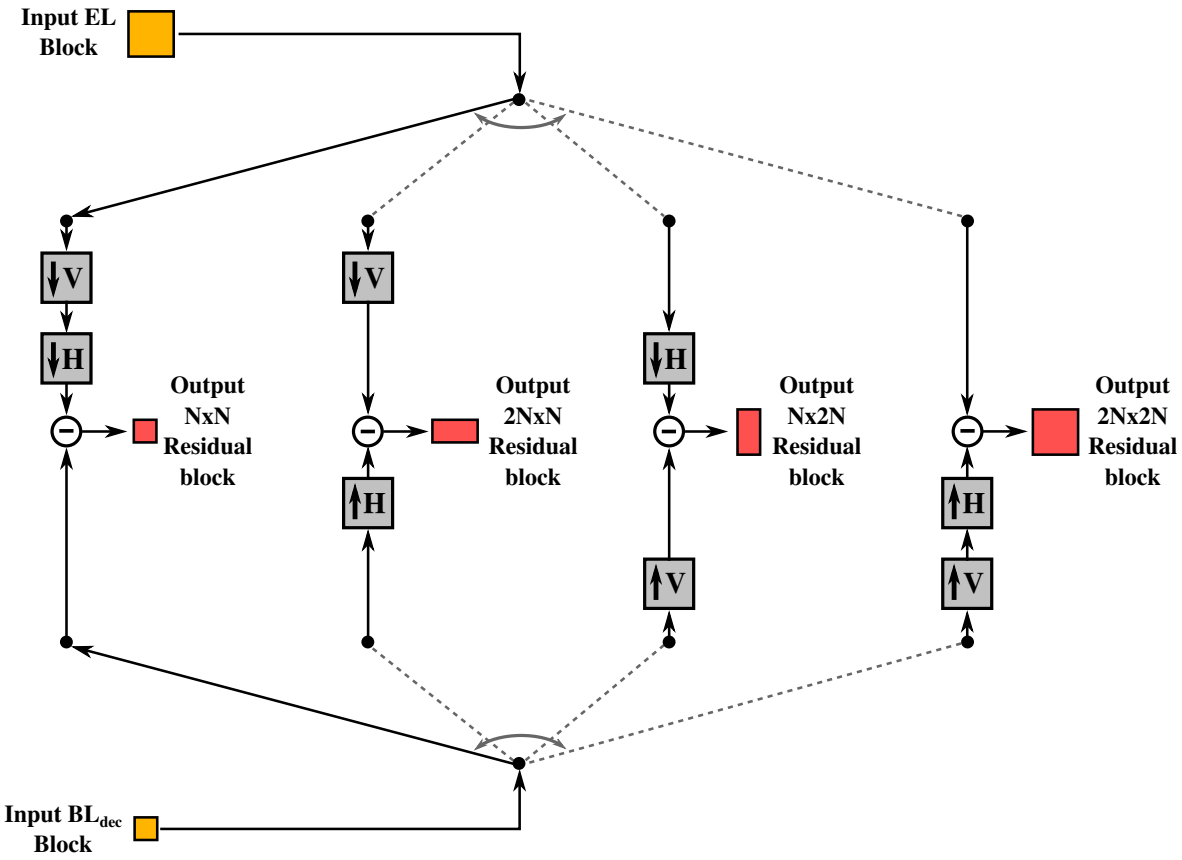


Fig. 5.6 Adaptive resolution selection.

entropy coded to evaluate the bit cost of the chosen resolution. On the other hand the inverse process is performed to reconstruct the decoded EL block in order to assess the distortion with respect to the source signal. The rate-distortion cost is then computed as follows

$$J(s,\lambda) = D(s) + \lambda \cdot R(s), \tag{5.7}$$

with J the join cost of the distortion D and rate R associated to the resolution mode s under test, $s \in \{N\times N, 2N\times N, N\times 2N, 2N\times 2N\}$. λ is a Lagrangian weighting factor depending on the QP. The λ derivation process, given in Equation (5.8) is the same as the one used for intra-coded blocks in the HM reference software with an additional factor, λ_{mod}, to adapt the resulting λ value to the spatial resolution selection case.

$$\lambda(QP) = \lambda_{mod}(QP) \cdot 0.57 \cdot 2^{\frac{QP-12}{3}} \tag{5.8}$$

The λ_{mod} values given in Equation (5.9) have been optimized through extensive testing,

using the same process as in [44].

$$\lambda_{mod}(QP) = \begin{cases} 12 & if \quad QP \leq 24 \\ 13 & if \ 24 \leq QP < 27 \\ 14 & if \ 27 \leq QP < 29 \\ 15 & if \ 29 \leq QP < 31 \\ 16 & otherwise \end{cases} \tag{5.9}$$

The quantization parameter QP, in the range $[0..51]$, is defined as a coding parameter whose value directly depends on the desired encoding quality and bitrate. It is thus set at an encoder level, in the configuration file or command line. However, to compensate for the losses induced by the downsampling and upsampling processes, an offset is applied on the QP value used for $N\times N$, $2N\times N$ and $N\times 2N$ resolutions, as follows:

$$QP_s = QP_{2N\times 2N} - 5 \qquad with\ s \in \{N\times N,\ 2N\times N,\ N\times 2N\}. \tag{5.10}$$

This QP offset value has been obtained after testing a wide range of offset values, from -10 to 10 independently for each resolution, aiming at giving the best rate-distortion performance when averaged over the tested sequences.

Once the RDO algorithm has selected the best resolution in a rate-distortion sense, the chosen mode is signaled in the bitstream, before the transform coefficients. Two new syntax elements, and their corresponding CABAC context, are thus added to the entropy coding process, namely *use_rect_res_flag* and *use_vert_dim_flag* respectively indicating if the EL size corresponds to a rectangular resolution and if the vertical dimension is used.

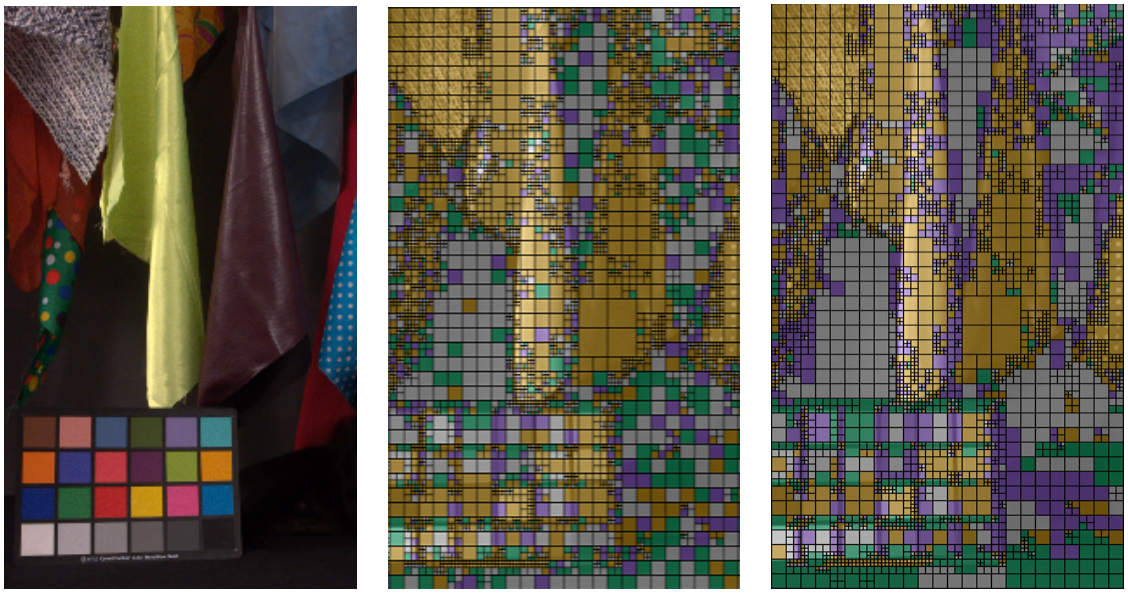
5.4 Study on Resolution Mode Decision

This section aims at investigating the resolution decisions made by the EL encoder to further improve the coding efficiency. The tuning algorithm performed on RDO choices is first presented, followed by a detailed analysis of the resolution choices distribution.

5.4.1 Tuning of RDO choices

After encoding with several configurations, resolution choices have been analyzed to assess if the expected behavior was a correct assumption. For example, the image sample shown in Figure 5.7a, taken from the first frame of the CatRobot sequence, can be used to analyze the resolution mode distribution with its diverse content - homogeneous areas, highly detailed scarfs, vertical and horizontal edges etc.

Figure 5.7b depicts the decisions made by the RDO algorithm for an encoding with a QP



(a) Source image (b) Original RDO algorithm (c) Tuned RDO algorithm

Fig. 5.7 Resolution mode decisions with and without RDO tuning for CatRobot test sequence at QP=22 - yellow=$2N\times2N$, green=$N\times2N$, purple=$2N\times N$, gray=$N\times N$.

value of 22. As expected, for the highly textured scarf at the top left of the displayed area, only the $2N\times2N$ resolution has been chosen. However, as can be seen for most homogeneous areas, such as the black foreground, or the several scarfs, the resolution choices are highly varying, which is not the expected behavior from the ASR-IL algorithm. Highly varying resolution choices induce two problems. On one hand, the overly frequent resolution transitions can be visually disturbing, especially in homogeneous areas. On the other hand, the context adaptation of the CABAC engine cannot achieve a good compression rate if the state of the flags used to signal the chosen resolution are constantly changing.

To explain this resolution variation, an analysis of the rate-distortion cost of each resolution has been carried out at a block level. The study showed that for a significant part of the blocks, several resolutions have a RD cost close to the optimal choice. Thus, the RDO resolution choice could be adapted, depending on characteristics of both the current block and its neighbors, in order to obtain more homogeneous resolution areas. This should lower both the visual impact of frequent resolution changes and the resolution signaling cost while having a limited effect on the objective output distortion.

Therefore, a RDO tuning algorithm has been designed, whose output resolution decisions are depicted in Figure 5.7c. As can be observed, there are much less variations of the chosen resolutions, especially in homogeneous areas such as the black foreground. Overall, the tuned RDO behavior is closer to what was expected when designing the adaptive spatial resolution scheme, notably for spatially homogeneous areas and rectangular resolutions.

Algorithm 1: RDO Tuning algorithm.

Data: Current block **B**, Top neighboring block **T**, Left neighboring block **L**, Current block distortion $D_{B,res}$ and bit cost $R_{B,res}$ for each resolution.
Result: Tuned resolution choice Res_B for current block **B**

$J_{min} \leftarrow$ maximum integer value;
foreach *res* **in** $\{N\times N,\ 2N\times N,\ N\times 2N, 2N\times 2N\}$ **do**
  Compute RD cost $J_{B,res}(D_{B,res}, R_{B,res})$ with equation (5.7);
  if $J_{B,res} < J_{min}$ **then**
    $J_{min} \leftarrow J_{B,res}$;
    $best\_mode \leftarrow res$;
  end
end
$J_{max} \leftarrow J_{min} + 2.56 \cdot W_B$;
foreach *res* **in** $\{N\times N,\ 2N\times N,\ N\times 2N, 2N\times 2N\}$ **do**
  if $J_{B,res} \in [J_{min}, J_{max}]$ **then**
    $best\_modes\_list$.append(*resolution*)
  end
end
if !$checkRectangularResolutions(best\_modes\_list, \mathbf{B}, J_T)$ **then**
  if !$checkneighborsResolution(best\_modes\_list, \mathbf{B}, \mathbf{L}, \mathbf{T}, J_T)$ **then**
    $J_T \leftarrow best\_mode$
end

Algorithm 1 details the RDO tuning process. The first step consists in performing the usual RDO stage, i.e. finding the best resolution mode in terms of rate-distortion trade-off. Then, the list of resolutions with close RD costs compared to the optimal choice, *best_modes_list*, is computed. Finally, the tuning process is carried out by calling two functions, namely *checkRectangularResolutions()* and *checkneighborsResolution()*.

Function *checkRectangularResolutions()* addresses the case of rectangular resolutions by verifying, for both $2N\times N$ or $N\times 2N$ resolutions, if they are present in *best_modes_list*. and if so, the function checks for the presence of strong vertical or horizontal edges, respectively using Equations 5.11 or 5.12 depending on the tested resolution. If both conditions are met for one of the two rectangular resolutions, the tuned RDO decision is set to that resolution.

$$\mathbf{S}_{B,x} > T_{grad} \cdot \mathbf{S}_{B,y}, \tag{5.11}$$

$$\mathbf{S}_{B,y} > T_{grad} \cdot \mathbf{S}_{B,x}, \tag{5.12}$$

with $\mathbf{S}_{B,x}$ and $\mathbf{S}_{B,y}$ the average Sobel-based gradient, along horizontal and vertical axes respectively, for the image block B. T_{grad} is a threshold used to configure the strength of the RDO tuning algorithm. $T_{grad} = 2$ proved to be the value giving the best results in terms

Function checkRectangularResolutions

Input: The list of resolution modes to test *best_modes_list*, current block **B**, the tuned resolution to be updated Res_B.
Output: *True* if the tuned resolution has been updated, *False* otherwise.

Compute $\mathbf{S}_{B,x}$ and $\mathbf{S}_{B,y}$ with Equation (5.13);
if $2N\times N \in$ *best_modes_list* **and B** contains strong vertical edges (Eq. (5.11)) **then**
    $J_T \leftarrow 2N\times N$;
    return *True*;
else if $N\times 2N \in$ *best_modes_list* **and B** contains strong horizontal edges (Eq. (5.12)) **then**
    $J_T \leftarrow N\times 2N$;
    return *True*;
else
    return *False*;
end

Function checkneighborsResolution

Input: The list of resolution modes to test *best_modes_list*, current block **B**, left block **L**, above block **T**, the tuned resolution to be updated Res_B.
Output: *True* if the tuned resolution has been updated, *False* otherwise.

Compute $\mathbf{S}_{B,x}$, $\mathbf{S}_{B,y}$, $\mathbf{S}_{L,x}$, $\mathbf{S}_{L,y}$, $\mathbf{S}_{T,x}$ and $\mathbf{S}_{T,y}$ with Equation (5.13);
if $Res_L \in$ *best_modes_list* **and** *B* similar to **L** (Eq. (5.14)) **then**
    $J_T \leftarrow Res_L$;
    return *True*;
else if $Res_T \in$ *best_modes_list* **and** *B* similar to **T** (Eq. (5.14)) **then**
    $J_T \leftarrow Res_T$;
    return *True*;
else
    return *False*;
end

of tuning potential and rate-distortion performance. The Sobel-based gradients are obtained using the Equation (5.13)

$$\mathbf{S}_{B,d} = \frac{1}{(W-1)(H-1)} \cdot \sum_{i=1}^{H-1} \sum_{j=1}^{W-1} k_{\mathbf{S}_d} * B(i,j) \tag{5.13}$$

with $\mathbf{S}_{B,d}$ the average gradients for the $W \times H$ input block **B** along direction d and $*$ the convolution product. The Sobel convolution kernels $\mathbf{K}_{S_x}$ and $\mathbf{K}_{S_y}$ used to respectively compute

horizontal and vertical gradients are defined as

$$\mathbf{K}_{S_x} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad \mathbf{K}_{S_y} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

Function *checkneighborsResolution()* addresses the neighboring resolution derivation case. Similarly to the rectangular resolution case, two conditions have to be met for a neighboring block resolution to be chosen. First, this resolution has to be in *best_modes_list*, and secondly the neighboring block must share similar characteristics with the current block. The similarity is based on several gradient and size ratios being comprised in certain value ranges, as detailed in Equation (5.14). T_{width} and T_{corr} are thresholds used to configure the tuning strength, with optimal values of $T_{width} = 2$ and $T_{corr} = 1.5$.

$$\mathbf{B}\ similar\ to\ \mathbf{V}\ if \begin{cases} \frac{1}{T_{width}} \leq & \frac{W_B}{W_H} & \leq T_{width} \\ \frac{1}{T_{corr}} \leq & \frac{\mathbf{S}_{B,h}}{\mathbf{S}_{V,h}} & \leq T_{corr} \\ \frac{1}{T_{corr}} \leq & \frac{\mathbf{S}_{B,v}}{\mathbf{S}_{V,v}} & \leq T_{corr} \\ \frac{1}{T_{corr}} \leq & \frac{\mathbf{S}_{B,h} \cdot \mathbf{S}_{V,v}}{\mathbf{S}_{B,v} \cdot \mathbf{S}_{V,h}} & \leq T_{corr} \end{cases} \tag{5.14}$$

with **B** the current block, **V** the neighboring block under test, W_i the width of block i and $\mathbf{S}_{i,d}$ the average Sobel-based gradient of block i along direction d.

5.4.2 Resolution Mode Distribution

After the design of the RDO tuning algorithm, the next step of the study on resolution mode decisions is to assess the distribution of the chosen resolutions depending on the QP. The whole dataset has thus been encoded with the proposed ASR-IL encoder. **Fig. 5.8** shows the ratio of blocks in each resolution over a wide range of QP values.

Overall, the same trend can be observed on every sequences: the ratios of N×N and $2N$×$2N$ are respectively gradually increasing and decreasing when the QP becomes larger, while the ratios of rectangular blocks, i.e. in $2N$×N or N×$2N$ resolutions, remain roughly constant over the entire range of QP values.

The main difference between the different sequences comprised in the dataset is the QP value at which the majority choice switches from $2N$×$2N$ to N×N. Indeed, for sequences

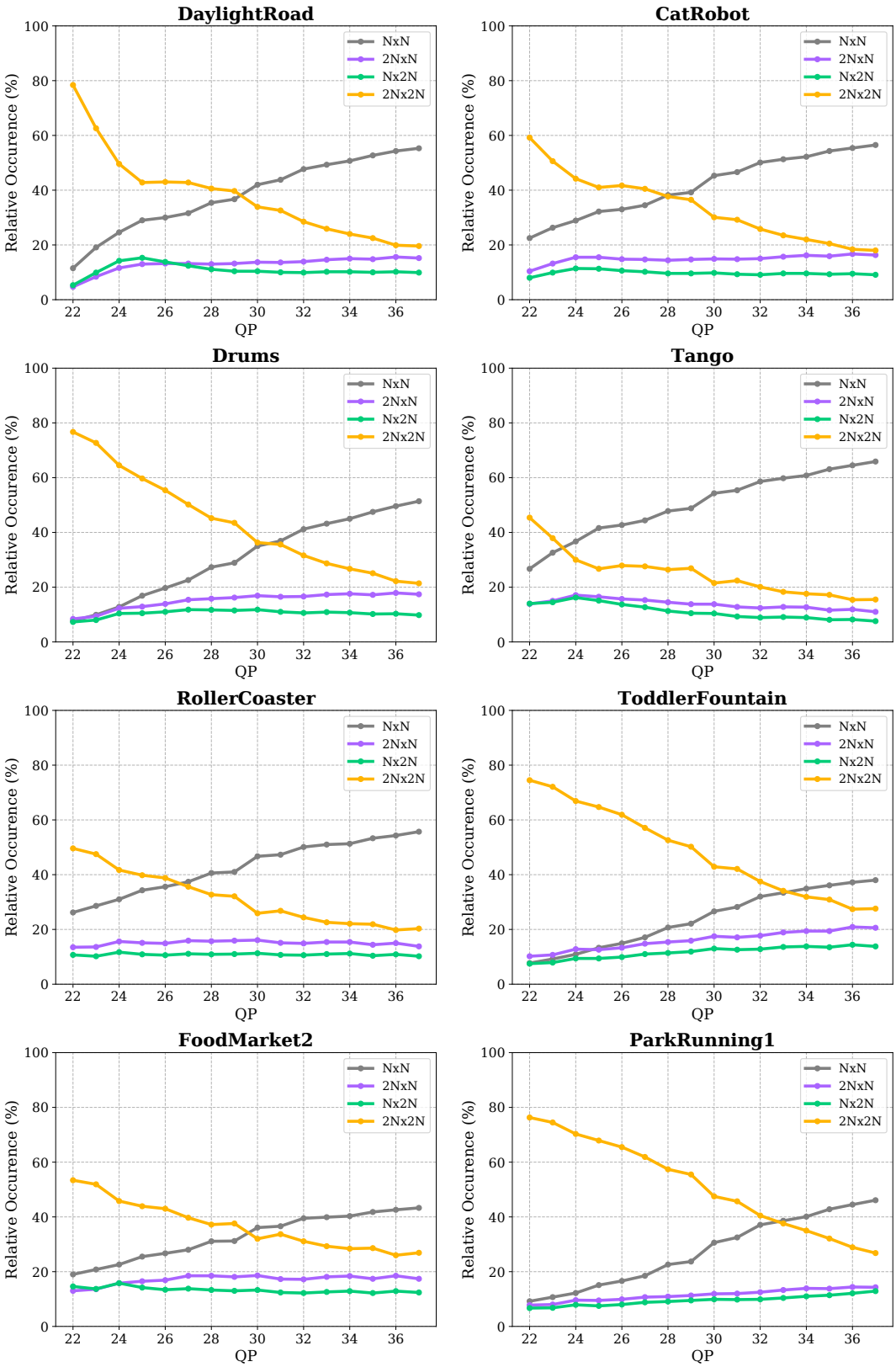


Fig. 5.8 Distribution of resolution modes after RDO with tuning enabled.

with a relatively low amount of spatial details, such as *Tango* or *RollerCoaster*, the $N \times N$ resolution becomes the majority choice at low QP values, respectively 24 and 27. This can be explained by the good quality of the BL encoding and the efficient upsampling stage which achieves a good detail restoration from the decoded base layer for these specific sequences, thus making the coding of prediction residuals of the $2N \times 2N$ mode less effective in terms of rate-distortion.

On the contrary, the *Drums*, *ParkRunning1* and *ToddlerFountain* sequences, which comprise a large proportion of areas with high spatial frequencies, respectively the water fountains, the different people and trees and the textured background, the majority resolution switch occurs at high QP values of 32-33. In this case, the prediction residual coding is mandatory for low QP values, in a RDO sense, to recover the spatial frequencies that have been lost during the base layer processing.

For the case of rectangular resolutions, since they are intended for blocks with strong vertical or horizontal edges which are common in natural content but generally do not cover wide areas, the observed lower selection ratios correspond to the expected behavior. In addition, the limited variation of these ratios over the range of QP values can be explained by the fact that the dedicated tuning function, *checkneighborsResolution()*, does not take the QP into account.

5.5 Experimental Results

This section focuses on evaluating the performance of the proposed low-complexity enhancement layer in terms of both rate-distortion and complexity gains. First, the proposed ASR-IL scalable encoder is compared to SHVC. Then, the impact of the adaptive spatial resolution tool is analyzed by comparing different versions of the algorithm.

5.5.1 Objective Evaluation

The dataset previously used for the performance evaluations carried out in Chapters 3 and 4, comprising six UHD sequences, has been extended to eight test sequences and used to evaluate the proposed scalable encoder. The reference encodings have been performed using the SHVC reference model, SHM9.0 [148], with an AI configuration for the base layer and an enhancement layer with P images having only their corresponding BL image as reference, as depicted in **Fig. 5.9**. The QP values used for the encodings range from 22 to 37 with a step of one.

BD-rate results are summarized in Table 5.1, positive values corresponding to a bitrate overhead, considering the whole bitstream i.e. both the BL and EL information, for the proposed ASR-IL encoder compared to SHVC for the same EL output PSNR value. It is important to note that the same downsampling filters are used in both codecs for the BL

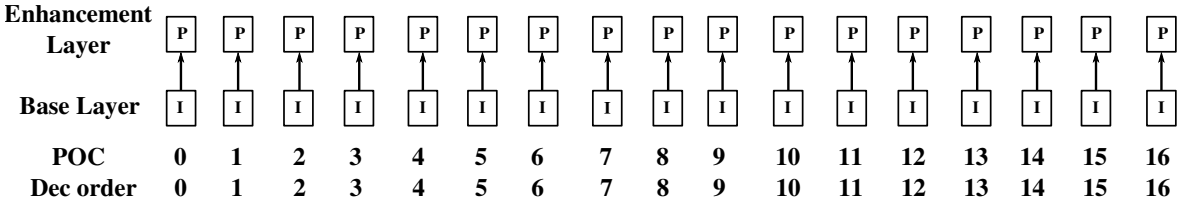


Fig. 5.9 Scalable prediction scheme with AI mode in the BL and P_{BL}-only in the EL.

Table 5.1 BD-Rate results (%) for proposed ASR-IL scalable encoder with and without resolution signaling compared to SHVC (BL in AI configuration, EL in P_{BL}-only).

Sequence	With resolution signaling cost				Without resolution signaling cost			
	Y	U	V	YUV	Y	U	V	YUV
DaylightRoad	4.43	2.11	-1.49	3.67	1.08	-1.33	-4.81	0.35
CatRobot	3.16	9.71	12.98	5.4	-0.17	6.53	9.72	2.08
Drums	5.15	9.33	7.67	6.03	1.84	6.15	4.6	2.75
Tango	3.63	3.67	4.55	4.22	-0.49	-0.25	0.69	0.19
RollerCoaster	4.0	10.65	7.81	5.38	0.05	6.84	4.15	1.46
ToddlerFountain	5.41	6.44	2.14	5.4	2.78	4.06	-0.09	2.81
FoodMarket2	5.08	7.92	8.26	6.06	1.73	4.85	5.18	2.77
ParkRunning1	0.54	7.39	4.23	2.09	-1.07	5.98	2.83	0.53
Average	**3.92**	**7.15**	**5.77**	**4.78**	**0.72**	**4.1**	**2.78**	**1.62**

input generation, thus inducing the same decoded BL samples in both coding chains. Two separate cases have been reported in the results, one with the resolution mode signaling taken into account in the bitrate computation and one without. Indeed, this allows to assess separately the coding potential of the ASR-IL scheme and the signaling cost of the resolution mode.

With the resolution mode signaling cost taken into account, the average bitrate overhead of the proposed ASR-IL encoder is 4.8 % compared to SHVC, for an equal PSNR-YUV value. The per-sequence results are fairly homogeneous, around $4 \sim 6$ %, except for the *ParkRunning1* which shows a reduced average overhead of 2.1 % that can be explained by the particularly better performance at low-bitrates for the ASR-IL scalable encoder. In addition, for every sequences, the losses are higher in the chroma planes compared to the luma one. It can be explained by the fact that, due to the higher sensitivity to the luma channel, a greater weight is applied to the luma distortion compared to the errors present in the chroma channels during RD cost computation in the RDO stage. This weighting process could be changed to obtain a better coding performance in the chroma channels, but this would be to the detriment of luma performance due to the higher overall bitrate for the same luma PSNR.

Figure 5.10 shows the rate-distortion curves for the proposed ASR-IL scalable encoder

Table 5.2 Complexity reduction for the proposed scalable encoder compared to SHVC (BL in AI configuration, EL in P_{BL}-only).

	Enhancement Layer	**Overall (EL + BL)**
Encoding $TR_{\%}$	88 %	72 %

and SHVC over a wide range of QP values. As can be observed, the performance of the ASR-IL encoder is equal or higher than SHVC at low bitrates while SHVC performs better when the bitrate increases. Indeed, at low bitrates, the majority of the blocks of the ASR-IL encoder are either skipped or encoded in one of the rectangular resolutions, as detailed in Section 5.4.2, thus reducing the residuals to encode and focusing mainly on the strong edges. On the contrary, at high bitrates, the EL resolution, i.e. $2N\times2N$, is selected for the majority of blocks in the ASR-IL algorithm. In this case, the encoder does not profit much from the reduced resolution, and is thus close to what is done in SHVC, without the independent EL partitioning and in-loop filtering present in the reference software, which explains the difference in performance at high bitrates.

If the resolution mode signaling is not taken into account in the bit cost computation, the average bitrate overhead of the proposed ASR-IL scalable encoder is, in average, 1.62 % for equal PSNR-YUV compared to SHVC. The per-sequence BD-Rate values follow the same trend as with the signaling cost taken into account, but with an average loss reduction of 3.1 %. The sequences benefiting the most from the adaptive resolution have a higher signaling cost than those for which the majority resolution remains $2N\times2N$ on a wide range of QP values. Indeed, when the chosen resolution is varying frequently, the signaling, which simply consists in two context coded bins, costs more bits after entropy coding than a nearly invariant resolution.

Complexity-wise, the proposed ASR-IL scalable coding scheme achieves a 72 % encoding time reduction compared to the SHVC reference software, as reported in Table 5.2. If only the EL layer is taken into account, the proposed scalable architecture is 8.5 times faster (88% EL encoding $TR_{\%}$) than SHM at the encoder side. This complexity gain is mostly due to the low computational demand of the RDO stage of the ASR-IL encoder, with the few modes to test and the derived EL partitioning.

5.5.2 Impact of the Spatial Resolution Adaptation

In order to assess the coding potential of the ASR-IL scheme, two different experiments have been carried out without taking into account the resolution mode signaling cost: the first is a comparison between the proposed scalable encoder and the same architecture without the adaptive resolution, i.e. with all blocks coded at full EL resolution ($2N\times2N$); the second experiment compares the proposed ASR-IL encoder to a constrained version, without the rectangular resolutions, i.e. only with $2N\times2N$ and $N\times N$ enabled. Table 5.3 summarizes the

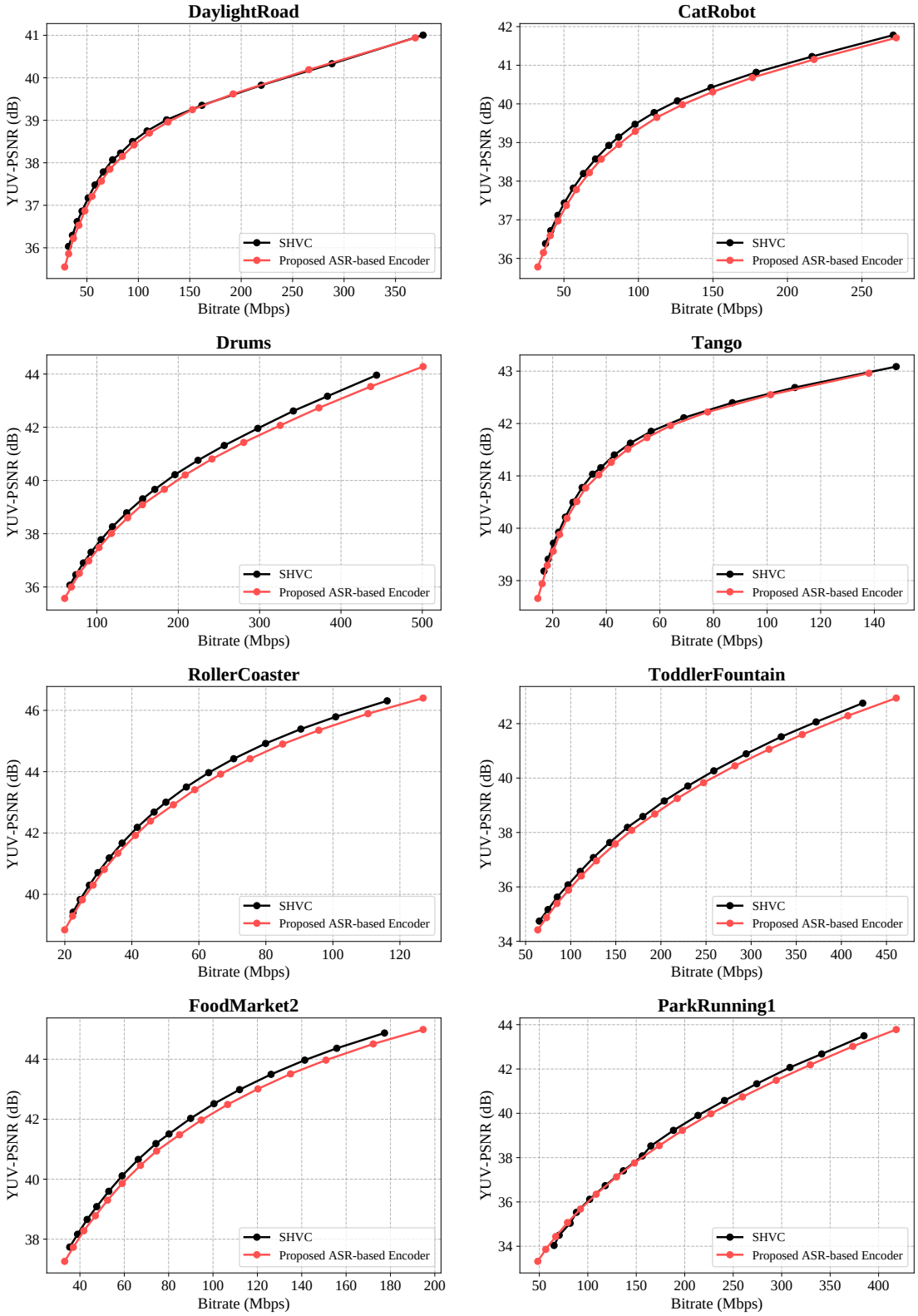


Fig. 5.10 Rate-Distortion curves for the proposed ASR-IL scalable encoder with resolution mode signaling cost compared to SHVC (BL in AI configuration, EL in P_{BL}-only).

Table 5.3 BD-Rate results (%) for proposed ASR-IL scalable encoder compared to versions either with only the full resolution or without rectangular resolutions.

Sequence	**ASR-IL vs ASR-IL w/ only *2Nx2N***				**ASR-IL vs ASR-IL w/o *2NxN* and *Nx2N***			
	Y	U	V	YUV	Y	U	V	YUV
DaylightRoad	-3.75	-8.82	-8.83	-4.92	-1.01	-4.7	-5.03	-2.05
CatRobot	-4.76	-2.13	-1.56	-4.11	-0.69	-2.74	-2.58	-1.14
Drums	-2.5	-3.42	-2.47	-2.57	-0.77	-0.95	-1.09	-0.8
Tango	-3.92	-7.77	-5.04	-4.61	-1.57	-2.1	-2.29	-1.9
RollerCoaster	-4.24	-1.85	-2.56	-3.77	-1.12	-0.15	0.13	-0.89
ToddlerFountain	-1.75	-2.42	-2.38	-1.87	-0.67	-1.95	-1.76	-0.88
FoodMarket2	-3.09	-2.75	-2.65	-2.99	-1.19	-0.88	-0.98	-1.16
ParkRunning1	-1.6	1.15	-1.92	-1.3	-1.99	1.58	0.52	-1.34
Average	**-3.2**	**-3.5**	**-3.43**	**-3.27**	**-1.13**	**-1.49**	**-1.64**	**-1.28**

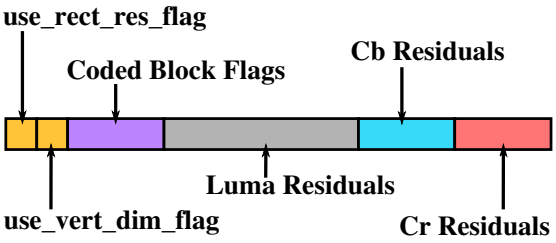
results of these experiments.

The first experiment has been designed to assess the raw gain provided by the proposed scheme. Results show an average bitrate reduction of 3.27 %, for the same PSNR-YUV value, for the ASR-IL scheme compared to a version of the encoder with only $2N\times2N$ blocks. As can be expected from the resolution mode distributions of each sequence, detailed in Section 5.4.2, sequences such as *Drums*, *ParkRunning1* and *ToddlerFountain*, for which the $2N\times2N$ resolution remains the majority class for a wide range of QP values, the gains are below average with respectively 1.87 %, 1.3 % and 2.57 % bitrate reductions. On the contrary, for sequences that profit from the reduction in resolution even at low QP values, around 4 % gains are achieved by the ASR-IL encoder.

The second experiment focuses on the efficiency of the rectangular resolutions $2N\times N$ and $N\times2N$. The BD-rate values show an average gain of 1.28 % for the proposed ASR-IL algorithm compared to a constrained version only allowing the BL and EL resolutions to be chosen. It can be observed that the obtained results do not directly correlate with the proportion of blocks coded at one of the rectangular resolutions. Indeed, due to the RDO tuning part where the resolution is derived from the neighboring block, the blocks can be signaled as a rectangular resolution even if there are no residuals to encode, thus resulting in the same reconstructed signal as if it would have been processed in the BL resolution.

Resolution	use_rect_res	use_vert_dim
$N \times N$	0	0
$2N \times N$	1	0
$N \times 2N$	1	1
$2N \times 2N$	0	1

(a) Resolution Mode flag values



(b) Syntax elements

Fig. 5.11 Resolution Mode signaling scheme.

5.5.3 Impact of Resolution Signalization

Syntax Elements and Entropy Coding

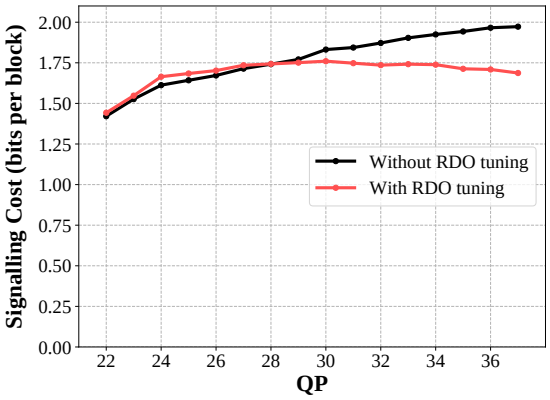
The resolution mode chosen by the ASR-IL algorithm is signaled, at a block level, using two different flags. The first, *use_rect_res_flag*, indicates if the current block is processed in a rectangular resolution or not. The second, *use_vert_dim_flag*, indicates if the vertical dimension is used in the chosen resolution. Both these flags are entropy coded, each with a dedicated CABAC context, to reduce the bit cost induced by the block level signaling.

Figure 5.11a summarizes the flag values for each resolution. Different combinations have been tested in an effort to reduce the signaling cost, including decision trees inspired by the intra mode prediction used in HEVC [33], but they did not provide any significant improvement due to the small flexibility of the scheme (only 4 modes for 2 flags).

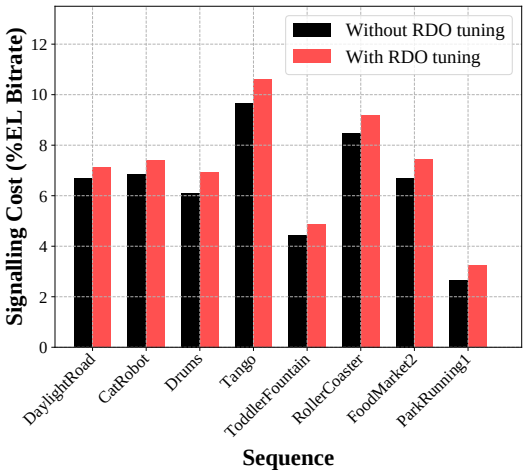
Since the syntax elements used to signal the transform coefficients, identical to those described in Section 2.3.2, require the dimension of the block to be entropy coded (and decoded), the flags indicating the resolution mode are signaled at the beginning of each EL block, as depicted in Figure 5.11b.

Cost Evaluation

In Section 5.5.1, it was shown that the resolution mode syntax elements were responsible for an overhead of around 3 to 4 %. In this section, the signaling cost is further analyzed, especially the impact of RDO tuning on entropy coding. Figure 5.12a depicts the average bits per block necessary to signal the chosen resolution with and without RDO tuning enabled. The



(a) Per-QP average signaling cost per block, averaged over all sequences



(b) Per-sequence signaling cost (%EL_bitrate) averaged over all QP values

Fig. 5.12 Average resolution signaling costs for ASR-IL encoder with and without RDO tuning.

results are presented per QP value, averaged over all sequences. It can be seen that, without RDO tuning, the cost per block is gradually increasing over the entire QP range, reaching the 2 bits/pixels threshold at high QP values, meaning that the entropy coding does not provide any gain at this point. Indeed, in addition to the overall more evenly distributed resolutions, there is a high variability in chosen resolutions, making it impossible for the CABAC engine to predict the flag values thus resulting in a high signaling cost. With the RDO tuning enabled, the variability is greatly reduced, which results in a much lower signaling cost at high QP values.

However, despite the decrease in average bits per blocks, the signaling still represents a significant part of the bitrate dedicated to the enhancement layer. Indeed, as depicted in Figure 5.12b, with or without RDO tuning, the resolution mode syntax elements represent, averaged over the QP range, from 3 to 10 % of the EL bitrate, depending on the sequence. For instance, for the *Tango* sequence, which is the one that profits the most from the ASR-IL scheme, the signaling reaches an average value of 10.5 % of the EL bitrate, with a peak values over 15 % for high QP values. It is important to note that these cost values are computed by only taking into account the EL bitrate, which is different from the signaling costs presented in Section 5.5.1 where the overall ($BL+EL$) bitrate was considered.

In addition, it can be observed on Figure 5.12b that, even with a lower signaling cost in number of bits, the contribution of the syntax elements signaling to the EL bitrate is higher with RDO tuning enabled. Indeed, the tuning algorithm does not only impact the resolution signaling but also the amount of residuals to encode due to the different mode decisions. In general, the tuning algorithm has a tendency to lower the overall bitrate due to the increase

in the number of blocks coded at lower resolutions, which explains why the resolution mode syntax elements still represent a similar or slightly higher portion of the transmitted bits compared to the case without decision tuning.

5.6 Conclusion

In this chapter, a low-complexity scalable encoder has been proposed to tackle the issue limiting the deployment of state-of-the art scalable solutions such as SHVC. The proposed encoder, called ASR-IL, relies on a dual-layer encoder to achieve x2 spatial scalability, using a standard HEVC as the base layer and a low-complexity encoder based on the local adaptation of the spatial resolution to process the enhancement layer.

The proposed EL encoder locally selects the spatial resolution, at a block level, from four possible resolutions depending on both the content of the block and the decisions taken to encode its spatially neighboring EL blocks and co-located BL block. Experimental results revealed that the proposed ASR-IL architecture can achieve average encoding complexity reductions of 88% and 72% for the EL encoder and overall architecture, respectively, with an average bitrate overhead of 4.8% compared to SHVC in the same configuration - AI for the BL and inter-layer predictions only for the EL (P_{BL} only).

The local adaptation of the spatial resolution appears to provide significant gains over a regular coding at a fixed resolution. However, these gains are not high enough to compensate the necessary signalization of the block-level resolution in the bitstream. If the signaling cost of the chosen resolution could be considerably reduced, the ASR-IL scalable coding scheme would reach the performance of SHVC, while offering a dramatic reduction of the encoding complexity.

Chapter 6

Extension of the Adaptive Spatial Resolution Tool to Inter-predicted Frames

As identified in Section 5.5, the main issue of the ASR coding tool is the signaling cost of the resolution mode chosen by the ASR algorithm at a block level. This chapter aims at solving this issue by extending the ASR algorithm in order to handle a BL encoded in RA configuration, i.e. using inter-picture predictions to encode the video frames.

Indeed, with inter-predicted BL frames, the transmitted MVs can be used by the EL encoder to predict the optimal spatial resolution, in a R-D sense, from previously encoded EL frames. In this case, the resolution signaling would not be required, leading to an overall greatly limited signaling cost overhead for the ASR coding tool.

In addition, the BL MVs can also be used to perform inter-picture predictions within the EL encoder. In this case, the popular RA coding configuration is enabled, and broadcast and streaming use-cases can thus be addressed.

This chapter is organized as follows. First, a solution to limit the signaling overhead of the ASR-IL scheme, based on resolution mode derivation via motion compensation is proposed in Section 6.1. Then, a second extension of the ASR scheme is presented in Section 6.2, consisting in using inter-picture predictions in the EL while considering adaptive spatial resolution. Finally, Section 6.3 concludes this chapter.

6.1 Adaptive Spatial Resolution with Mode Derivation

This section focuses on proposing on extending the previously proposed ASR-IL solution with an algorithm based on inter-layer ASR with mode derivation, thus called ASR-IL-MD in the rest of this dissertation. The mode derivation scheme uses motion compensation to

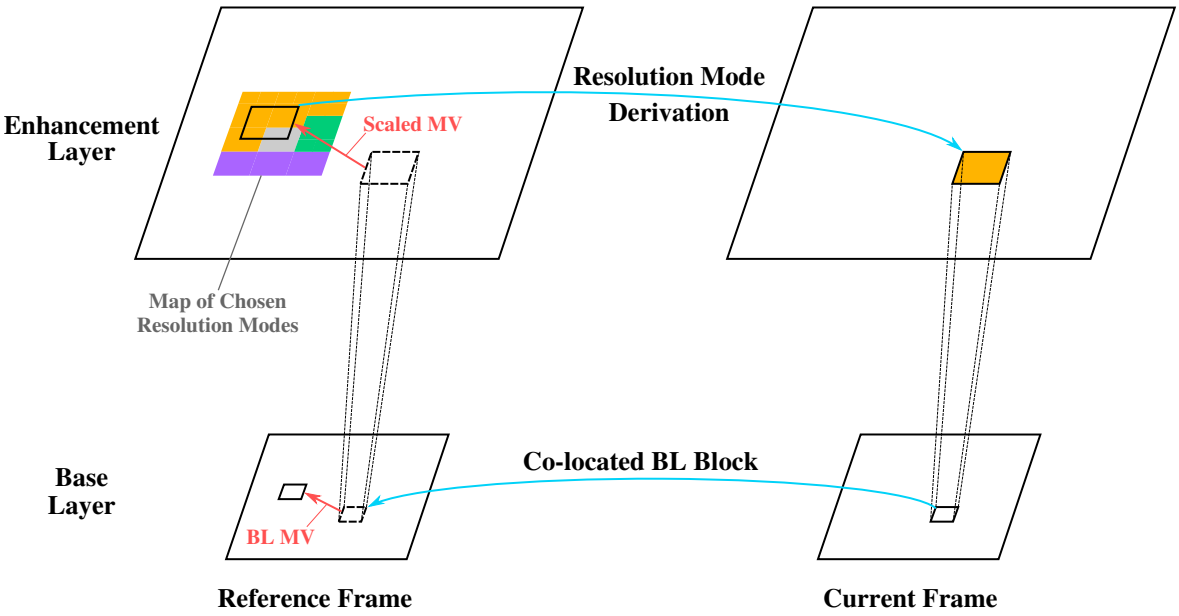


Fig. 6.1 Example of resolution mode derivation via motion compensation.

predict the optimal spatial resolution in order to avoid transmitting the resolution mode syntax elements for the majority of blocks, in an effort to greatly reduce the overall signaling cost. The ASR-IL-MD method is first presented, followed by a description of its implementation into the existing architecture. Finally, the coding efficiency is analyzed in terms of bit rate gains over SHVC and impact on signaling cost.

6.1.1 Description of the Technique

The idea behind mode derivation via motion compensation is straightforward: select the resolution chosen for the corresponding motion compensated block in the reference frame. To avoid performing a highly computationally demanding motion estimation step in the EL encoder, the motion vectors of the base-layer are directly reused, after an appropriate scaling, to derive the resolution mode of the EL blocks, as depicted in **Fig. 6.1**. Therefore, if the BL encoder is configured in Random-Access mode to allow for inter-picture predictions, this method, whose architecture is shown in **Fig. 6.2**, can be used for every EL block whose corresponding BL PU has been encoded using inter-picture predictions.

As for the scheme presented in Section 5.3, all EL blocks are encoded only using their co-located block in the base layer as reference for pixel value predictions, while motion compensation is only used to derive the spatial resolution, as depicted in **Fig. 6.3**. EL blocks with a corresponding intra-picture coded BL PU will follow the same process as the previous architecture (RDO, tuning, resolution mode signaling etc.).

With this scheme, the impact of resolution signaling on the resulting bitrate is expected to drop significantly compared to the previous scheme. Indeed, in RA configuration, especially

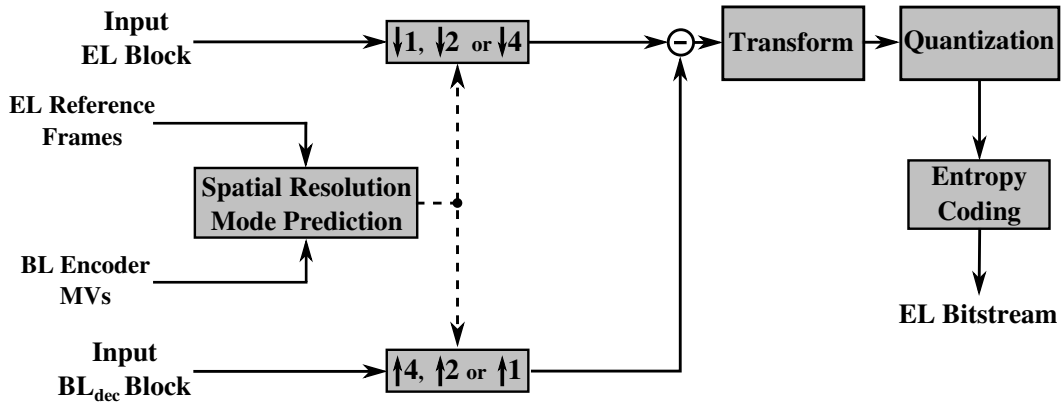


Fig. 6.2 EL encoder architecture for resolution mode derivation via motion compensation.

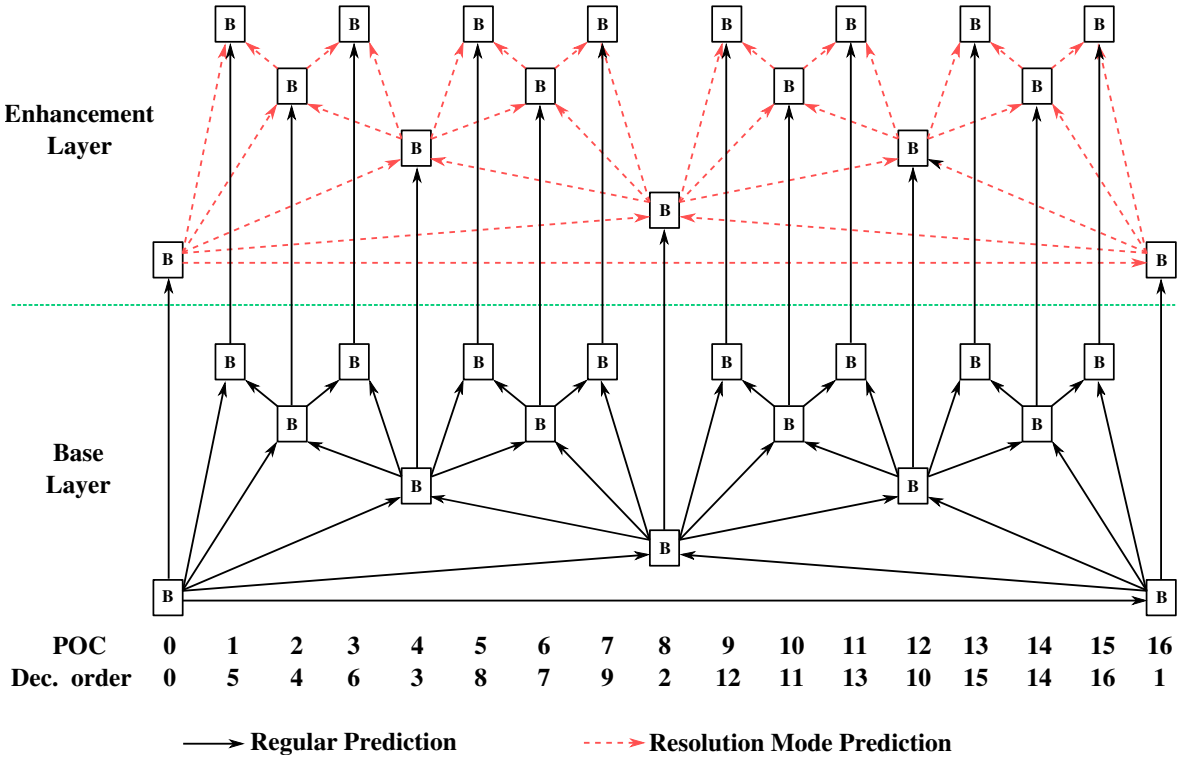


Fig. 6.3 Prediction scheme for the proposed scalable encoder with RA configuration in the BL and resolution mode prediction enabled in the EL.

with the 16-image GOP used in the different studies presented in this report, the majority of blocks are inter-coded, notably in the upper temporal layers, which should limit the number of blocks requiring the resolution to be signaled.

The second advantage of this mode derivation via motion compensation scheme is the temporal coherence of the resolution decisions. With the previous scheme, due to the independent decisions from one frame to another, an intra-refresh effect was clearly visible, with objects often being coded in different resolutions in adjacent frames. The resolution deriva-

tion, and inter-coded base layer, should bring more temporal stability to the ASR decisions, resulting in a more visually pleasing reconstructed output video.

The main drawback would be the additional parsing dependency required at the encoder side, between both layers, due to the BL motion vectors being required before processing the EL block.

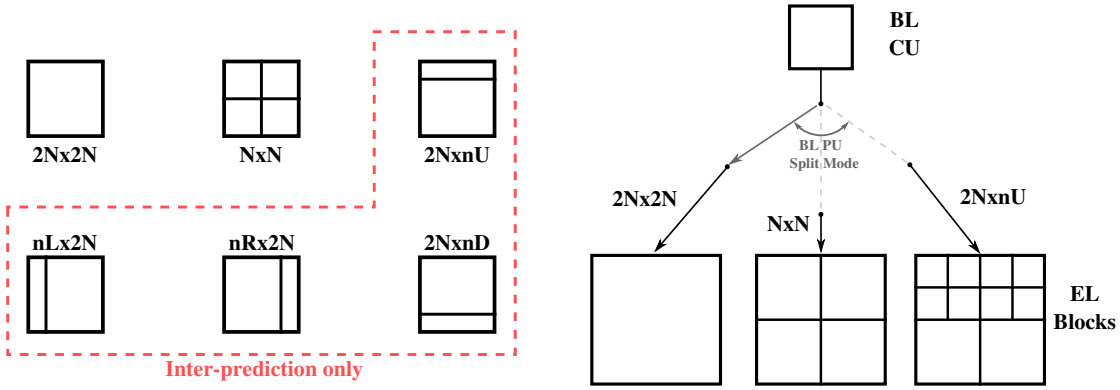
6.1.2 Implementation

The first step to implement the mode derivation technique based on motion compensation is to add the management of reordered pictures coding and HEVC GOP structure to the EL encoder architecture. Indeed, the base layer being coded in RA configuration, the EL encoder must follow the same coding order and keep identical reference frame buffers to allow for a correct utilization of the motion estimation performed in the base layer. Then, assuming a correct reference frame buffer management, the BL motion vectors, previously retrieved from the base layer bitstream, can be directly used as it is in the enhancement layer encoder. The only processing needed is a scaling of the MVs to match the EL resolution, in this case a multiplication by a factor two in each direction.

The second step is to compute, for each reference frame, its resolution map containing the resolution mode used to encode each pixel of the frame. Then, from the motion vector and the associated reference frame, the block resolution is directly derived from the resolution indicated in the corresponding motion compensated area in the resolution map. The derivation process is defined in Equation (6.1)

$$res = \begin{cases} 2N\times 2N & if\ \mathbb{P}(x = 2N\times 2N) > p_{Th} \\ y^* = \underset{y\in L}{max}\,\mathbb{P}(x = y) & otherwise \end{cases} \tag{6.1}$$

with *res* the derived resolution, $L = \{2N\times 2N, N\times 2N, 2N\times N, N\times N\}$ is the set of possible resolutions; $\mathbb{P}(x)$ the probability of having a pixel of the EL reference block coded in resolution x, with $x \in L$. The probability threshold for the selection of the $2N\times 2N$ ($p_{Th} = 0.3$) resolution has been introduced to avoid losing important details when the majority of the reference pixels has not been coded in the EL input resolution. Indeed, contrary to the intra-picture partitioning, the inter-picture partitioning does not always depend on the contours and texture of the image but rather on the similarity with the reference frame in a given area. Thus, it is often the case to obtain different kinds of textures and details within a single inter-predicted block, which requires special handling if the corresponding blocks in the reference frame have been encoded using different resolutions. The value of 0.3 has been selected after extensive experiments. In the case of bi-predictions, the derivation process remains the same, with the probabilities computed across both reference blocks.



(a) BL PU splitting modes (b) EL size derivation scheme

Fig. 6.4 EL partitioning derivation scheme from BL prediction unit size.

Table 6.1 EL block partitioning derivation depending on BL CU size and PU split mode.

BL CU size \ PU Split Mode	$2N\times2N$	$N\times N$	$2N\times nX$ *or* $nX\times2N$
(8×8)	(16×16)	$4\times(8\times8)$	$8\times(4\times4)+2\times(8\times8)$
(16×16)	(32×32)	$4\times(16\times16)$	$8\times(8\times8)+2\times(16\times16)$
(32×32)	$4\times(32\times32)$	$4\times(32\times32)$	$8\times(16\times16)+2\times(32\times32)$
(64×64)	$4\times(64\times64)$	$16\times(32\times32)$	$32\times(16\times16)+8\times(32\times32)$

The last modifications made for the implementation of the proposed ASR-IL-MD scheme, apart from the tuning algorithm optimization presented in the next section, concern the EL partitioning derivation process from the base layer encoder decisions. In addition to the generally higher block size, the inter-prediction in HEVC also induces additional PU modes compared to the two splitting modes ($2N\times2N$ and $N\times N$) allowed for intra-predicted frames, as depicted in Figure 6.4a. Thus, the EL block size derivation process optimized for the previous architecture is not necessarily the best for a base layer including inter-picture predictions.

Therefore, several derivation rules, based on multiples of the PU base layer sizes, have been tested to find the optimal EL partitioning. Figure 6.4b depicts the derivation process used in the proposed scalable encoder with the BL encoded in RA configuration. The only exception is a limitation to 64x64 size for the largest EL blocks, thus mainly affecting the BL CUs of size 64x64. Table 6.1 summarizes all possible resulting partitioning depending on the BL block size and the PU split mode.

6.1.3 Adaptation of the RDO Tuning Optimization

Due to the introduction of the new resolution mode derivation tool, which increases the importance of decisions taken after the RDO tuning stage, each threshold value, optimized for the previous coding structure, must be re-evaluated to obtain the best coding efficiency with the mode derivation enabled. Thus, for both main parameters of the RDO tuning algorithm, i.e. $TGrad$ and $TCorr$, an optimization study has been carried out.

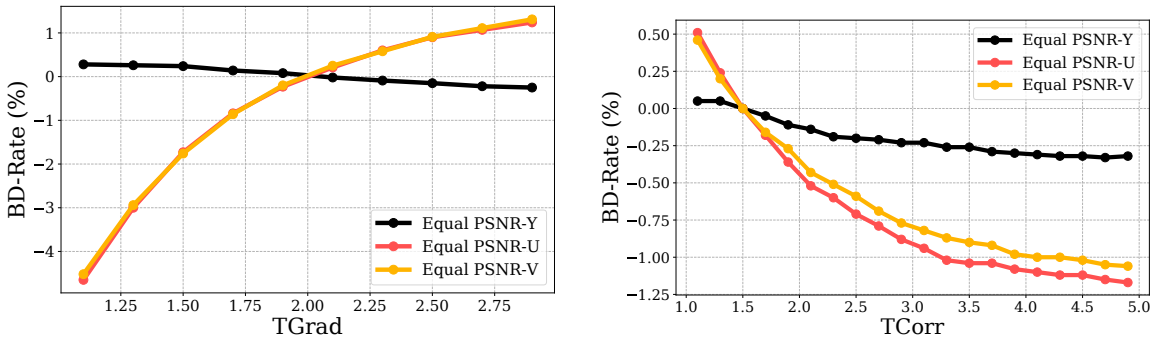
Figure 6.5a shows the results achieved using different values for $TGrad$, the threshold controlling the strength of the *checkRectangularResolutions()* function, the first step of the tuning algorithm in charge of detecting vertical or horizontal contours and selecting the appropriate rectangular resolution if necessary. As depicted on the graph, low values of $TGrad$, i.e. below the previously used value of $TGrad = 2.0$ thus increasing the amount of rectangular blocks, result in losses in the luma channel and gains in both chroma channels, while the exact opposite is observed when the value of $TGrad$ increases.

This behavior can be explained by the fact that spatial frequencies are generally higher in the luma channel, hence the increased losses induced by downsampling the block in one direction. This is further confirmed looking at the per-sequence performance, where low values of $TGrad$ result in high losses for sequences with a high amount of details, due to the high coding cost of the residuals obtained after upsampling, while smaller gains are achieved for sequences like *Tango*, for which the low amount of details is easily recovered by the upsampling process. Therefore, in order to not further decrease the performance on the sequences for which the encoder is already struggling to match state-of-the-art coding efficiency, the optimal static value for $TGrad$ remains 2.0.

However, this statement is only true if the QP is not taken into account. Indeed, for low QP values, since the decoded base-layer has nearly been deprived of all high spatial frequencies, even for sequences like *ToddlerFountain*, it often becomes more interesting, in a RDO sense, to prioritize either only vertical or horizontal details instead of the costly recovery of all frequencies. Thus, the overall performance can be increased by making the $TGrad$ value as a function of the quantization parameter. The final implementation uses the linear relationship of Equation (6.2) to derive $TGrad$ from the QP value

$$TGrad(QP) = \begin{cases} 1.2 & if\ QP < 22 \\ 3.12 - 0.053 \times QP & if\ 22 \leq QP \leq 37 \\ 2.0 & if\ QP > 37 \end{cases} \tag{6.2}$$

Figure 6.5b shows the results achieved using different values for $TCorr$, the correlation threshold used to assess the similarity between the current block and its above or left neighbor in the second stage of the RDO tuning algorithm (function *checkneighborsResolution()*). It can be seen that a better performance in both luma and chroma channels is obtained



(a) TGrad optimization

(b) TCorr optimization

Fig. 6.5 RDO tuning threshold optimization.

when $TCorr$ increases. In fact, the best results are obtained with an infinite value of $TCorr$, meaning that the neighbor resolution would be automatically selected if it is present in the *best_modes_list*, which contains the different modes having a RD cost close to the optimal resolution. This case, for which bitrate savings are obtained for all sequences compared to the original value of $TCorr = 1.5$, has thus been chosen in the final implementation of the RDO tuning algorithm.

6.1.4 Objective Evaluation

In order to perform a fair comparison between the proposed architecture and the state-of-the-art, SHVC has been limited to only use the decoded BL frames as reference to encode the corresponding EL picture, the base layer being encoded with inter-picture predictions enabled (RA configuration) for both SHVC and the proposed ASR-IL-MD scalable encoder. Apart from this modification, the same encoding parameters as the study presented in Section 5.5.1 have been used.

Table 6.2 summarizes the BD-rate results for the proposed encoder compared to SHVC. Overall, the proposed ASR-IL-MD achieves an average bitrate overhead of 0.5 % compared to SHVC, for equal PSNR-YUV. The results obtained for both the U and V chroma channels are better than the performance of the luma channel, with respectively -1.7 %, -2.6 % and 0.9 % average BD-rate values.

Due to the high variability of the per-sequence and per-channel performance, it is difficult to perform a general analysis of the results, except from the fact that the ASR-IL-MD encoder performs better than the ASR-IL algorithm on sequences that benefit from the changes in spatial resolution, and shows equivalent losses on the other sequences with high spatial details. In particular, the chroma performance is highly variable depending on the sequence, with BD-rate values ranging from 15.3 % bitrate savings to 14.1 % bitrate overhead for *DaylightRoad* and *ParkRunning1* sequences, respectively.

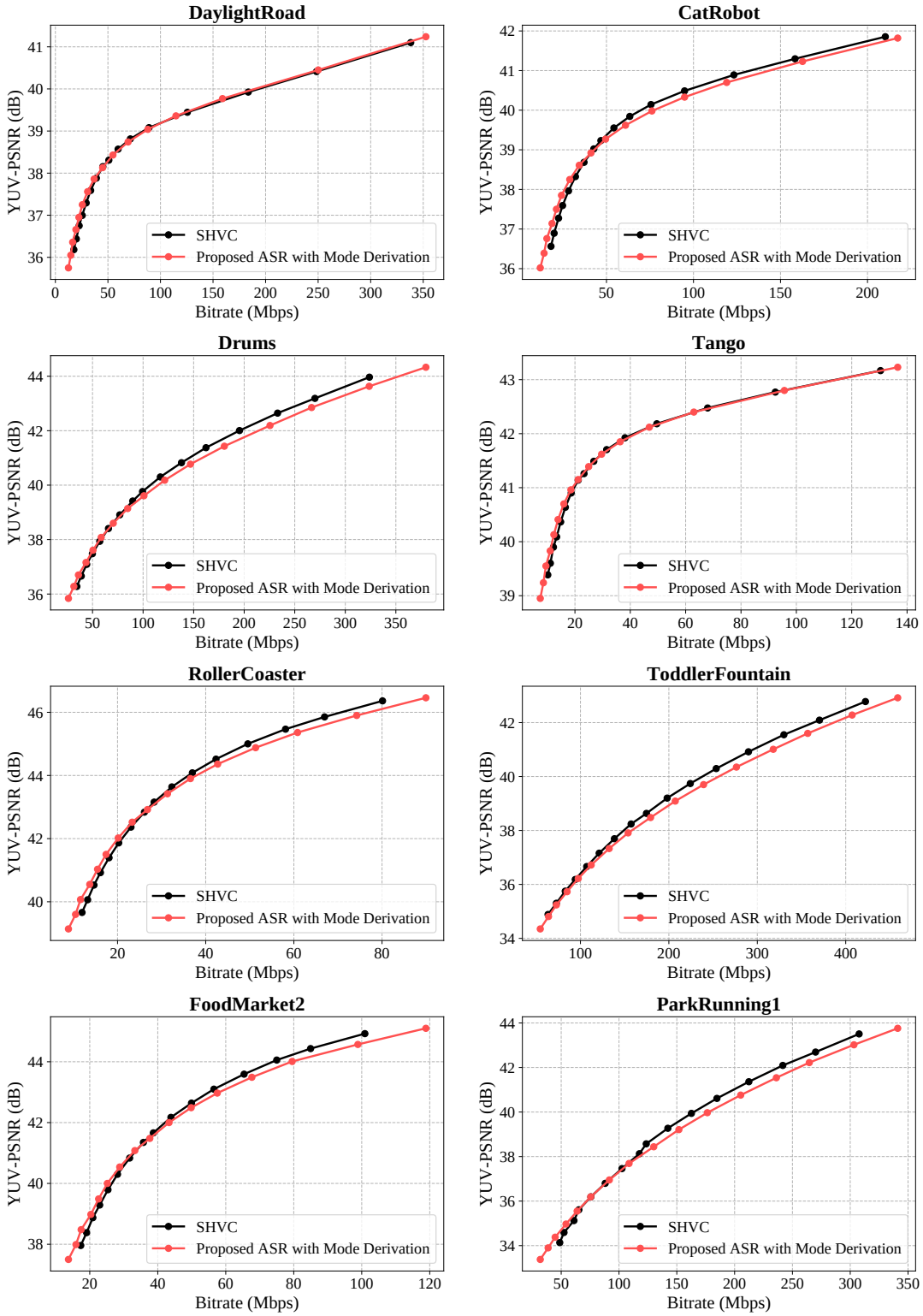


Fig. 6.6 Rate-Distortion curves for the proposed ASR-based scalable encoder with mode derivation compared to SHVC (BL in RA configuration, EL in P_{BL}-only).

Table 6.2 BD-Rate results (%) for proposed ASR-based scalable encoder with mode derivation compared to SHVC (BL in RA configuration, EL in P_{BL}-only).

Sequence	**Y**	**U**	**V**	**YUV**
DaylightRoad	-0.31	-15.3	-16.41	-3.98
CatRobot	-1.48	-4.42	3.71	-0.85
Drums	4.16	0.86	0.49	3.81
Tango	-2.58	-5.96	-7.63	-3.7
RollerCoaster	-0.36	-0.03	-1.26	-0.2
ToddlerFountain	7.65	-3.17	-9.23	5.69
FoodMarket2	0.22	0.29	0.12	0.39
ParkRunning1	-0.12	14.12	9.69	2.92
Average	**0.9**	**-1.7**	**-2.57**	**0.51**

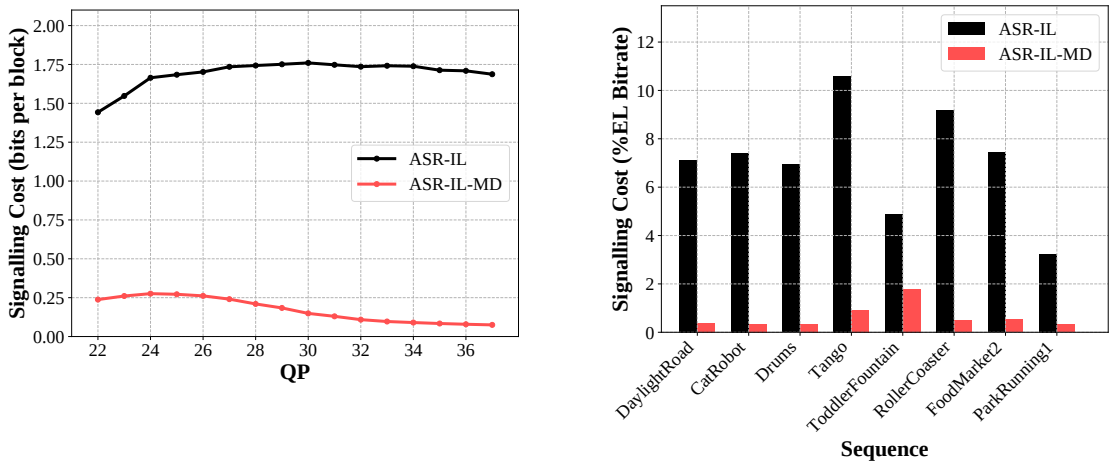
Table 6.3 Complexity reduction for the proposed scalable encoder compared to SHVC (BL in RA configuration, EL in P_{BL}-only).

	Enhancement Layer	**Overall (EL + BL)**
Encoding $TR_{\%}$	96 %	47 %

The rate-distortion curves depicted in Figure 6.6 show that the proposed ASR-IL-MD encoder performs better than SHVC at low bitrates, where the algorithm can often affect lower resolutions. On the contrary, this tendency is reversed for high QP values, especially for sequences where the $2N\times2N$ resolution is selected for a large majority of blocks. The main gain over SHVC is obtained for high QP values, the proposed encoder performing either similarly, for sequences with a limited amount of spatial details, or slightly worse than SHVC at high bitrates. This trend can be explained by the fact that, at low QP values, the behavior of the ASR-IL-MD encoder is very close to SHVC, especially for highly detailed content, but without key coding tools such as in-loop filters, reference frame filtering or EL-specific partitioning, that have not been implemented for complexity reasons. At lower bitrates, the absence of these coding tools is compensated by the gain brought by the adaptive spatial resolution, which is increased by the lower signaling cost, as will be shown in the next section.

A significant change compared to the previous ASR-IL proposition, due to both the BL coded in RA configuration and the resolution mode derivation, is the added temporal coherency which results in a greatly improved subjective quality, as observed by expert viewers. Indeed, the intra-refresh effect is no longer present thanks to the inter-predicted base layer pictures and the resolution affected to a given area or object remains stable from one frame to another thanks to the motion compensated resolution mode derivation.

Concerning the complexity of the ASR-IL-MD architecture, the average encoding times



(a) Per-QP average signaling cost per block, averaged over all sequences

(b) Per-sequence signaling cost (%EL_bitrate) averaged over all QP values

Fig. 6.7 Average costs for resolution mode signaling with and without mode derivation via motion compensation.

are summarized in Table 6.3. If only the enhancement layer is taken into account, the proposed EL encoder represents only 4 % of the processing time of the EL encoder of the SHVC reference software. The lower computation time of the ASR-IL-MD solution compared to the ASR-IL encoder is due to the RDO stage skipping when the block resolution can be derived from both the scaled MVs and reference frame. Considering the overall complexity, i.e. both the base and enhancement layers, the encoding complexity reduction reaches 47 % compared to the SHM. This high difference between the EL-only and overall results can be explained by the increased complexity brought by the inter-picture predictions in the BL.

6.1.5 Impact on Signaling Cost

The last part of this section focuses on the impact of the resolution mode derivation via motion compensation technique on the resolution mode signaling cost. Table 6.4 summarizes the BD-rate results for the proposed ASR-IL-MD encoder with and without the signaling cost taken into account. It can be seen that, on average, the resolution syntax elements signaling represents a bitrate overhead of 0.45 %, which is a clear improvement compared to the previously evaluated cost results of more than 3 % for the ASR-IL encoder. As expected, the resolution mode derivation greatly reduces the number of blocks for which the syntax elements have to be signaled in the bitstream, the only ones remaining being the blocks coded using intra-predictions in the base layer.

The per-sequence signaling costs are stable in general, around a 0.3 % overhead, except for two sequences, namely *Tango* and *ToddlerFountain* with respectively a 0.62 % and

Table 6.4 Signaling cost results (BD-Rate %) for the ASR-IL-MD scalable encoder.

Sequence	**Y**	**U**	**V**	**YUV**
DaylightRoad	0.3	0.32	0.31	0.3
CatRobot	0.29	0.28	0.28	0.29
Drums	0.28	0.28	0.27	0.28
Tango	0.62	0.63	0.62	0.62
RollerCoaster	0.35	0.35	0.34	0.35
ToddlerFountain	1.14	1.15	1.17	1.14
FoodMarket2	0.4	0.39	0.39	0.4
ParkRunning1	0.25	0.25	0.25	0.25
Average	**-0.45**	**-0.46**	**-0.45**	**-0.45**

1.14 % bitrate overhead. For the *Tango* sequence, it can be explained by the high variability in chosen resolutions compared to other sequences. For *ToddlerFountain*, the above average signaling cost is mainly due to the high amount of intra-coded blocks in the base layer. Indeed, this sequence presents important unpredictable movements, especially with the numerous water fountains. Therefore, inter-predictions are not effective, in a RDO sense, compared to intra predictions, which results in a high amount of intra-coded blocks in the BL, thus limiting the use of the resolution mode prediction in the EL.

Figure 6.7a shows the mean signaling cost, in bits per block, averaged over all sequences, for each tested QP value for the ASR-IL and ASR-IL-MD encoders. As can be expected from the BD-rate results, the signaling cost is at least divided by six for every QP when the mode derivation is enabled. The cost is gradually decreasing when the quantization parameter increases due to the overall lower amount of intra coded blocks in the base layer, at lower bitrates.

Figure 6.7b depicts the per-sequence portion of the EL bitrate taken by the resolution mode signaling, averaged over all QP values, for both the ASR-IL and ASR-IL-MD encoders. The cost reduction factor is nearly constant for every sequences, except with *ToddlerFountain* for which the mode derivation cannot lead to great signaling cost reductions, as previously explained.

6.2 ASR-based EL Inter-Predictions

This section focuses on extended the ASR-IL-MD scheme to fully support the RA configuration, i.e. with inter-picture predictions used in both the BL and EL. The proposed algorithm, called ASR-IP-MD, enabling motion compensation with adaptive spatial resolution is thus presented in order to offer a lightweight scalable coding solution compatible with both broadcast and streaming use-cases. The ASR-IP-MD method is first described, followed

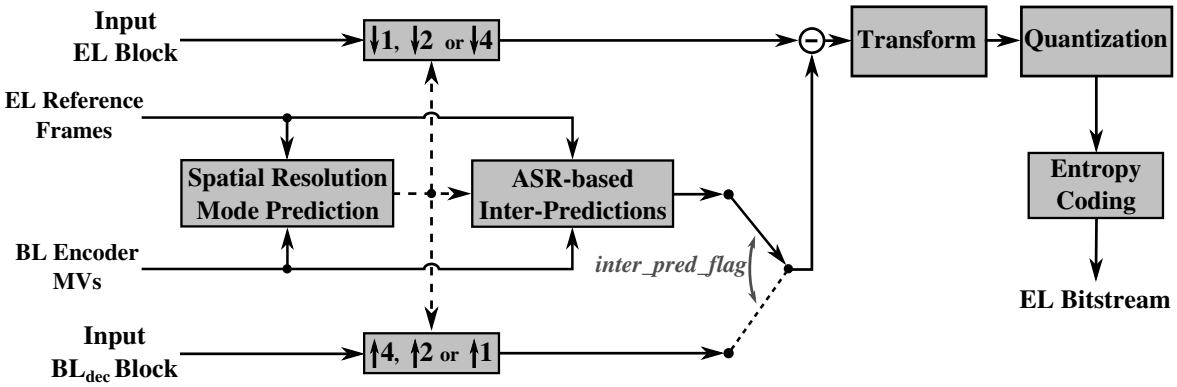


Fig. 6.8 El encoder architecture for ASR-based inter-predictions.

by a detailed presentation of the ASR-based inter-predictions implementation. Finally, the performance of the proposed encoder is evaluated and compared to SHVC.

6.2.1 Description of the Technique

The idea behind the addition of ASR-based inter-picture predictions in the EL encoder is straightforward: enable the same motion-compensation process as in HEVC but at a potentially different spatial resolution than the resolution of the input and reconstructed reference EL frames. The proposed corresponding ASR-IP-MD encoder architecture is depicted in **Fig. 6.8.**

Once again, to avoid a time consuming motion estimation algorithm, the MVs used for the EL inter-picture predictions are the scaled BL MVs. Therefore, EL inter-picture predictions are only enabled for EL blocks with an inter-coded corresponding PU in the base layer.

As for the previous proposed solutions, the resolution in which the prediction step, and subsequent core encoding operations, are performed depends on the resolution chosen by the mode derivation scheme. The algorithm proposed in Section 6.1 for the ASR-IL-MD encoder is thus reused to derive the spatial resolution based on the decisions made in the reference frames. **Fig. 6.9** shows the prediction structure and coding dependencies of the proposed ASR-IP-MD encoder.

EL inter-picture predictions are not systematically used when MVs are available. Indeed, a choice between inter-picture predictions and inter-layer predictions, i.e. using the upsampled BL corresponding block as reference, has been integrated in the EL encoder architecture. The prediction mode decision is taken in a RDO manner. Both prediction modes are thus tested sequentially and the one showing the best performance, in a R-D sense, is selected. The chosen mode is transmitted in the output bitstream using a specific context-coded syntax element, *inter_pred_flag*.

To summarize, three predictions schemes are available in the proposed ASR-IP-MD en-

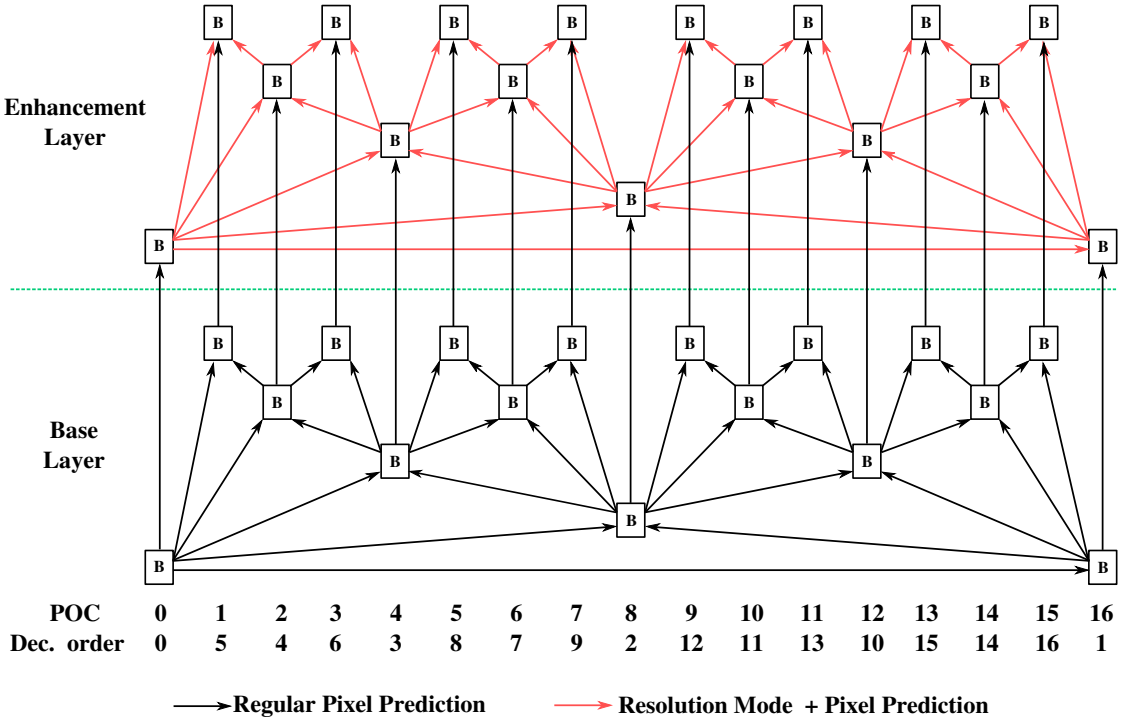


Fig. 6.9 Prediction scheme for the proposed ASR-IP-MD scalable encoder in RA configuration in both layers and resolution mode prediction enabled in the EL.

coder to process the EL blocks:

- **Inter-layer prediction with RDO-based resolution selection:** used systematically for EL blocks with a corresponding intra-coded BL PU (cf. Chapter 5 presenting the ASR-IL encoder);
- **Inter-layer prediction with resolution mode derivation:** can be used for EL blocks with a corresponding inter-coded BL PU (cf. Section 6.1 presenting the ASR-IL-MD encoder);
- **Inter-picture prediction with resolution mode derivation:** can be used for EL blocks with a corresponding inter-coded BL PU (presented in this section)).

These three predictions schemes are enabled in the implementation and performance evaluation of the ASR-IP-MD encoder presented in the following sections.

6.2.2 Implementation

A first solution to implement inter-picture predictions with an adaptive spatial resolution is intuitive and straightforward, as depicted in **Fig. 6.10**. Indeed, if the chosen resolution is $2N \times 2N$, the reference EL block, obtained using the corresponding reference frame in the

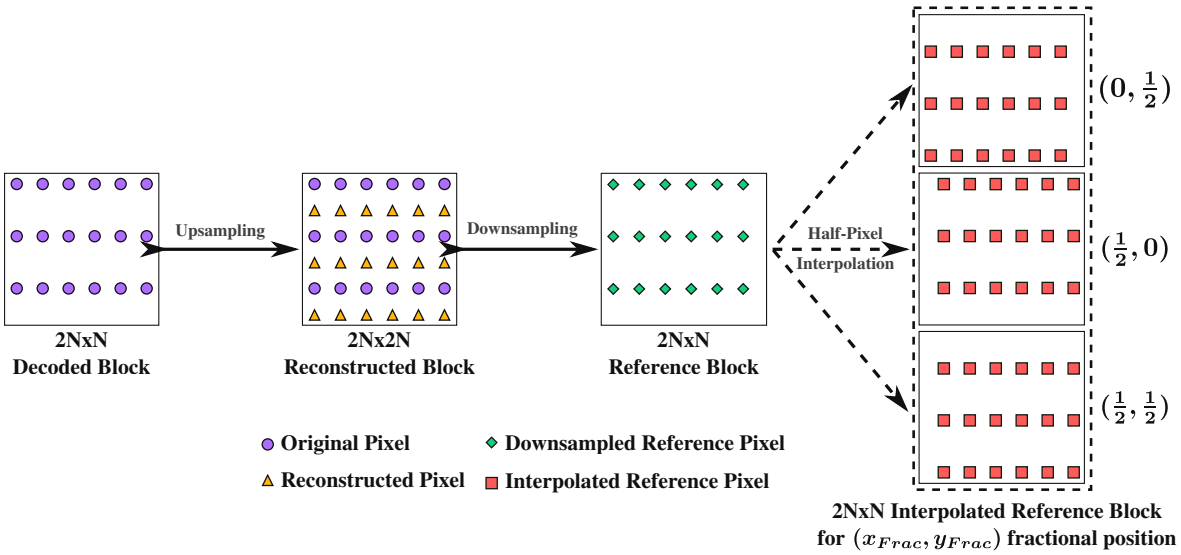


Fig. 6.10 Example of straightforward implementation of ASR-based inter-predictions with sub-optimal prediction of a 2*N*×*N* EL block.

Decoded Picture Buffer (DPB), is used as it is for the prediction. Otherwise, the reference EL block can simply be downsampled to the chosen resolution and be used as it is for prediction. However, in this case, due to the resolution derivation process, an upsampling to the output EL resolution has been performed on most reference pixels for the subsequent storing of the reconstructed reference frame in the DPB. Therefore, the predicted pixels of the EL block being encoded are obtained using an upsampling operation followed by a downsampling step. This first solution is thus greatly sub-optimal and produces a high prediction residual energy. This phenomenon is further amplified if fractional pixel interpolation is directly used as in HEVC. Indeed, since BL MVs are in quarter-pixel precision, the scaled MVs are in half-pixel precision, which requires an additional interpolation step to be performed after the upsampling and downsampling operations.

Instead, due to the inherent properties of the upsampling process used in the ASR scheme, a more efficient implementation, depicted in **Fig. 6.11**, is proposed. Indeed, since the considered upsampling filter bank copies even samples, the reference EL block can be obtained in the chosen resolution, smaller than 2*N*×2*N*, by only taking the even pixels. In this case, for the 2*N*×*N* resolution, the pixels used for predictions are identical to the decoded pixels before upsampling to the output EL resolution. The only sub-optimal case is when the EL reference block overlaps with blocks coded in different spatial resolutions in the reference frame. Indeed, a downsampling operation is needed for pixels not corresponding to the chosen majority resolution. However, this happens for a small amount of pixels and is thus not considered as an issue compared to the previously described implementation.

For the fractional interpolation process, the implementation can also be optimized, as depicted for the 2*N*×*N* resolution in **Fig. 6.11**. Since the filter bank for the interpolation

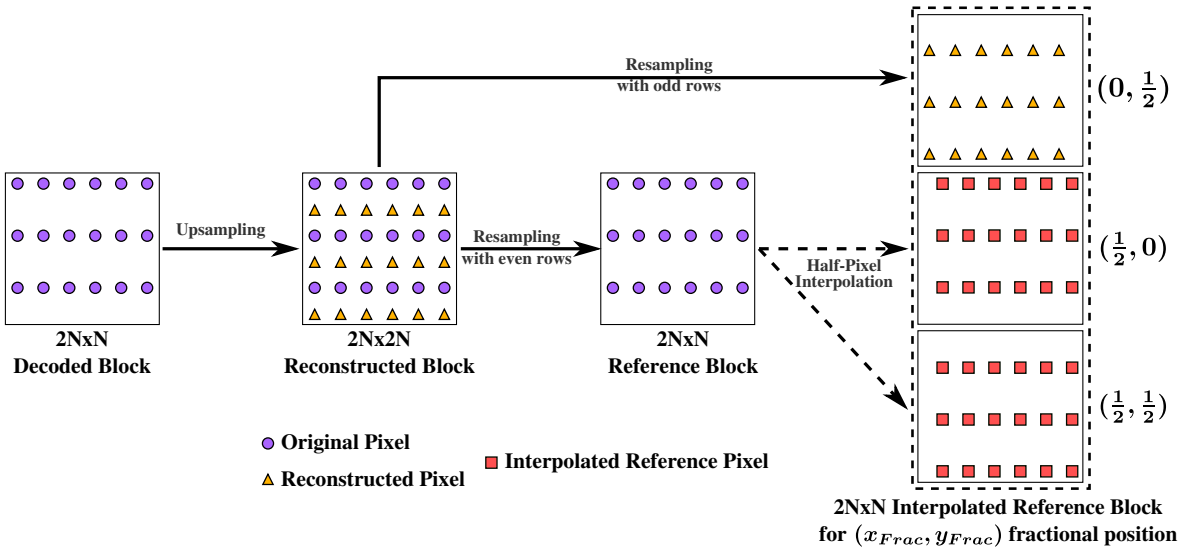


Fig. 6.11 Example fractional pixel interpolation for ASR-based inter-prediction with $2N \times N$ as chosen resolution.

is the same as the one used for upsampling, odd samples of the reconstructed reference EL block, which have been obtained by upsampling, can be used as reference for one of the three fractional sample positions, depending on the chosen resolution. For the remaining two fractional sample positions, the half-pixel interpolation is performed to obtain the correct reference pixels.

The other implementation decision made in the ASR-IP-MD encoder is the partitioning level at which the *inter_pred_flag* syntax element is signaled in the bitstream. As for the resolution mode signaling cost, investigated in Section 5.5.3, signaling the inter-picture decision for each EL block would result in a significant bitrate overhead. Instead, the implementation of the ASR-IP-MD encoder uses a signalization of the prediction mode at a PB level. Therefore, the chosen prediction mode is the same for all EL blocks corresponding to the same BL PU, i.e. for EL blocks sharing the same motion information. The RDO process in charge of selecting the best prediction mode is thus performed over all EL blocks sharing the same *inter_pred_flag* syntax element.

6.2.3 Objective Evaluation

The performance evaluation carried out in this section uses SHVC as reference, in RA configuration for both layers, thus following the JCT-VC CTC [53], except for the GOP size of 16 images. For the proposed ASR-IP-MD encoder, the HEVC BL encoder is also in RA configuration while the EL encoder also enables inter-picture predictions, as described in the previous sections. A QP offset of 4 has been used between the BL and EL encoder configurations of the proposed ASR-IP-MD scheme. Indeed, as it has been shown in [149], the

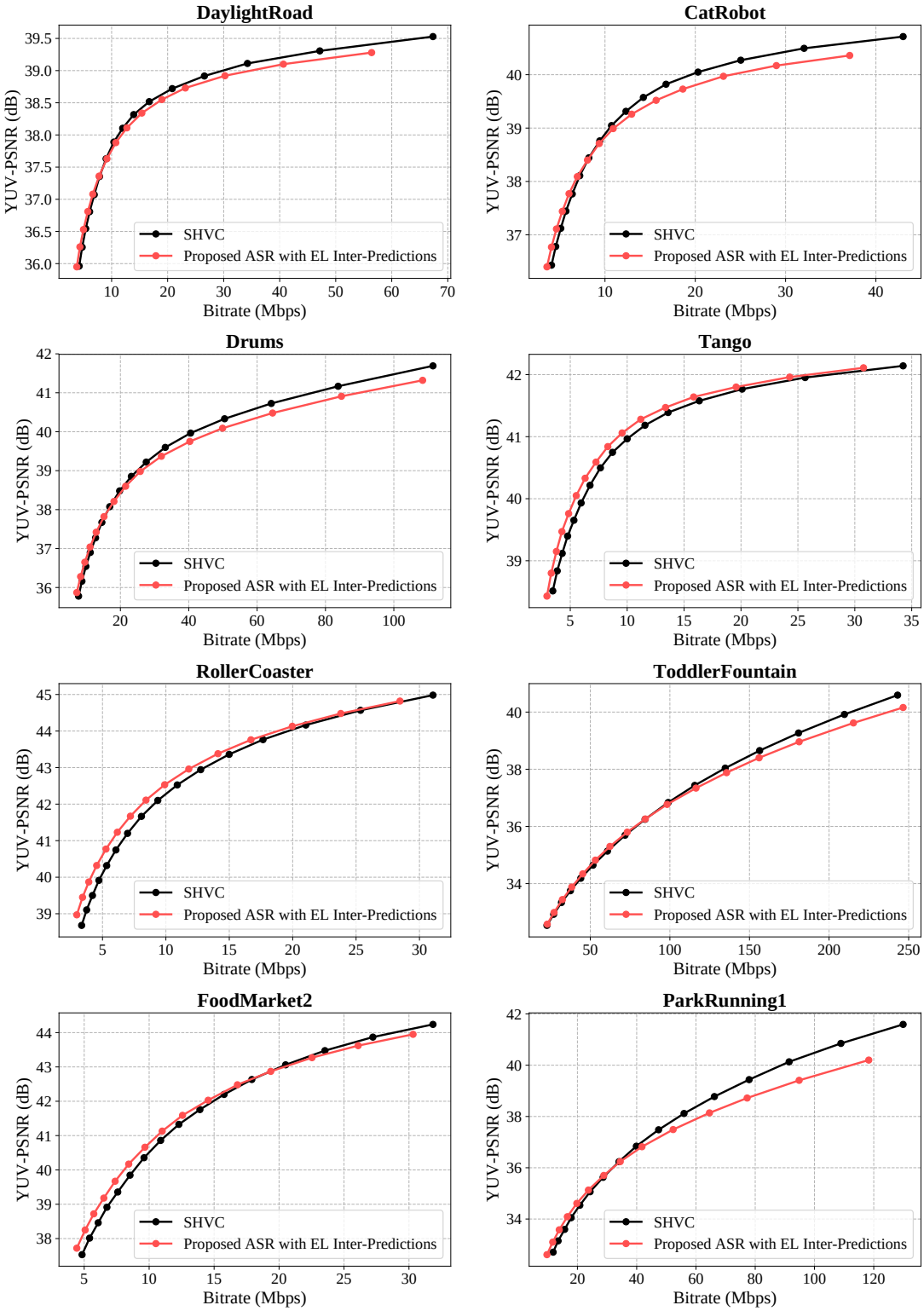


Fig. 6.12 Rate-Distortion curves for the proposed ASR-IP-MD encoder compared to SHVC (BL and EL in RA configuration).

Table 6.5 BD-Rate results (%) for proposed ASR-IP-MD scalable encoder with mode derivation compared to SHVC (BL and EL in RA configuration).

Sequence	**Y**	**U**	**V**	**YUV**
DaylightRoad	2.42	-2.44	3.09	2.08
CatRobot	-2.0	12.7	25.7	2.05
Drums	1.59	14.2	26.8	4.09
Tango	-11.5	-9.88	-9.93	-10.7
RollerCoaster	-11.4	-6.66	-3.48	-10.5
ToddlerFountain	2.27	-4.59	-2.36	1.74
FoodMarket2	-7.06	0.6	-1.47	-5.68
ParkRunning1	-4.38	55.9	51.4	3.77
Average	**-3.76**	**7.48**	**11.2**	**-1.64**

bitrate ratio between both layers can be optimized. Extensive experiments resulted in this optimal QP offset value of 4 for the proposed scalable coding scheme.

Table 6.5 summarizes the BD-rate results for the proposed ASR-IP-MD encoder compared to SHVC. Overall, bitrate savings of 1.64% compared to SHVC are achieved, for equal PSNR-YUV. The per-channel performance is quite variable, with a gain of 3.76% for the luma channel and average overhead of 7.5% and 11.2% for the U and V chroma channels, respectively. This difference could be resulting from the optimization performed mostly to improve the luma performance, as it is the channel that influences the perceived quality the most. A better chroma performance could be achieved by both changing the weight of chroma reconstructed errors in the R-D cost computation and adding chroma QP offsets.

As can be observed, the performance of the ASR-IP-MD encoder compared to SHVC is significantly varying depending on the test sequence. As observed for the previously presented ASR-IL and ASR-IL-MD solutions, the sequences with a lower amount of spatial details show subsequent gains over SHVC while the proposed solution struggles with sequences containing very highly textured areas. In addition, significantly higher chroma losses are observed for the *ParkRunning1* test sequence. This could be explained by the high amount of color information present in this content, which is not correctly recovered by the EL encoder with its less than optimal handling of chroma residuals.

The rate distortion curves depicted in **Fig. 6.12** show that the proposed ASR-IP-MD encoder and SHVC follow the same behavior as described in Section 6.1.4. The previously detailed analysis is thus still valid for proposed ASR-IP-MD encoder coding scheme.

Complexity-wise, the proposed ASR-IP-MD encoder scalable coding scheme achieves an overall average encoding time reduction of 78.6% compared to SHVC, as summarized in Table 6.6. This performance is higher than for the previous architecture due to the EL being coded in RA configuration. Therefore, the SHVC EL encoding time is greatly increased

Table 6.6 Complexity reduction for the proposed scalable encoder compared to SHVC (BL and EL in RA configuration).

	Enhancement Layer	**Overall (EL + BL)**
Encoding $TR_{\%}$	97.8 %	78.6 %

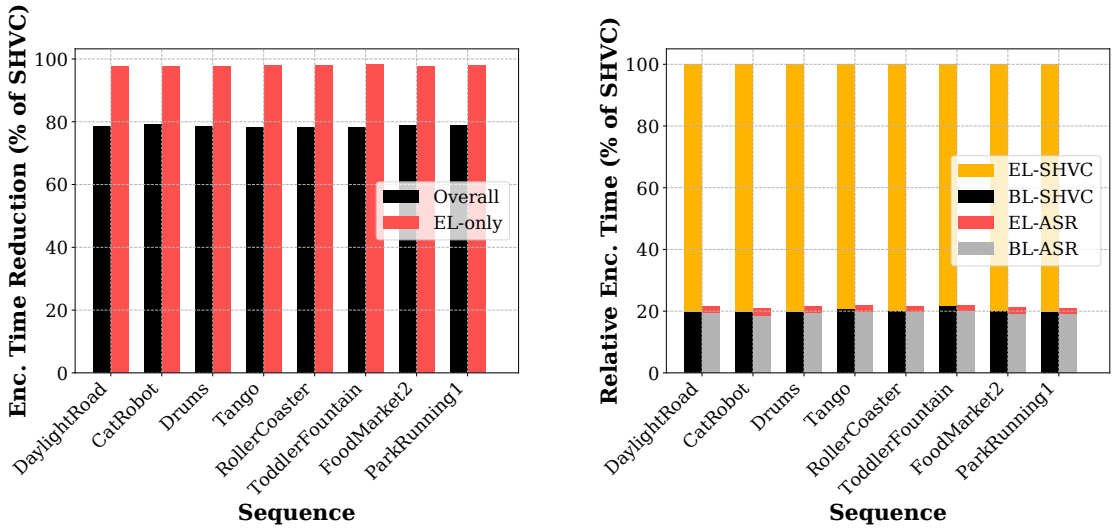


Fig. 6.13 Per-sequence encoding times comparison between proposed ASR-based encoder and SHVC in RA configuration.

compared to the previous study where only P_{BL}-only predictions were enabled in the EL encoder. If only the EL encoding time is taken into account, the ASR-IP-MD encoder shows an average complexity reduction of 97.8% compared to the HEVC EL encoder of the SHVC architecture. This means that a $x45$ speedup is achieved in the EL encoder.

The per-sequence average encoding time reduction results are depicted in **Fig. 6.13a**. As can be observed, the complexity gain is quite stable across all tested sequences and is thus not content-dependent. **Fig. 6.13b** shows the relative encoding time for each layer for both the SHVC and proposed ASR-IP-MD encoders. As can be seen, contrary to SHVC, spatial scalability is achieved with only a small complexity overhead for the ASR-IP-MD encoder. Indeed, the EL encoding time only represents a 11% encoding time average increase over the BL encoding process whereas the complexity overhead reaches 400% for the SHVC architecture.

Table 6.7 shows the performance of state-of-the-art SHVC complexity reduction algorithm compared to the ASR-IP-MD solution. As can be seen, our proposition outperforms state-of-the-art solutions both in terms of complexity reduction and coding efficiency. This improvement is enabled by the replacement of the EL HEVC encoder by an ASR-based very low complexity encoder, contrary to the state-of-the-art solutions that propose a speedup of

Table 6.7 Comparison of proposed coding scheme to state-of-the-art SHVC complexity reduction solutions.

	Tohidypour [143]	Shen [146]	Ours
EL Enc. time reduction	77 %	69 %	97.8 %
BD-rate value	4.2 %	2.3 %	-1.3 %

the HEVC encoding process.

6.3 Conclusion

In this chapter, the low-complexity scalable encoder based on the local adaptation of the spatial resolution proposed in Chapter 5 is improved with two complementary propositions.

The first solution, called ASR-IL-MD, aims at tackling the main issue of the ASR coding tool, i.e. the block-level spatial resolution signaling overhead. To this end, an extension of the ASR-IL algorithm is proposed to handle a BL encoder in RA configuration and thus exploit the BL MVs to limit the signaling cost of the coding tool. Indeed, a resolution mode derivation via motion compensation algorithm is proposed to remove the resolution signaling cost by predicted the chosen spatial resolution from previously encoded EL frames when MVs are available in the BL encoded bitstream. Experimental results showed that the proposed ASR-IL-MD solution can achieve a bitrate overhead reduced to 0.5% compared to SHVC in the same configuration - RA for the BL and P_{BL} only for the EL - for average encoding time reductions of 96% and 47% for the EL encoder and complete dual-layer encoder, respectively.

A second solution, called ASR-IP-MD, is proposed to add the support of RA configuration in the EL encoder, thus enabling the most popular coding configuration for broadcast and streaming use-cases. To this end, an ASR-based inter-prediction scheme is proposed, reusing the BL motion information and the previously proposed resolution mode derivation scheme. Through optimized fractional pixel interpolation and prediction mode signaling processes, a fast and efficient implementation of inter-predictions in the EL encoder is enabled. Experimental results showed that a complexity reduction of 78.6% and 97.8% can be achieved for the overall scalable ASR-IP-MD encoder and EL encoder, respectively. Concerning the coding efficiency, average bitrate savings of 1.3%, for equal PSNR-YUV quality, are shown compared to SHVC, despite some losses observed in the chroma channels.

The ASR-IP-MD solution thus outperforms state-of-the-art SHVC complexity reduction algorithms by a considerable margin. However, the proposed solution only tackles spatial resolution adaptation, thus further gains could be achieved by investigating the possible resolution adaptation in the temporal domain.

Part V

Spatio-Temporal Resolution Adaptation for Low-complexity Scalable Coding

Chapter 7

Variable Frame-Rate

As mentioned in Section 2.1.2, the introduction of frame-rates up to 100/120 fps, i.e. HFR, is one of the new features brought by the UHD video format. HFR video has been widely studied in the past decade [85, 150–155], showing a significantly better motion protrayal and thus greatly improving the perceived quality. However, these improvements are highly content-dependent and mostly confined to certain types of video content, notably those with high motion such as sporting events.

Besides, this increase in frame-rate doubles the amount of data to process compared to the traditional frame-rates of 50/60 fps currently used in the broadcast and streaming industries. This is one of the factors, in addition to the spatial resolution increase tackled in the previous chapter, that significantly increases the coding and decoding complexities using current solutions, thus greatly limiting the deployment of a complete UHDTV transmission chain to the consumer.

Following the same idea as for the adaptive spatial resolution presented in Chapter 5, the frame-rate of an input 120 fps video could be locally lowered when such a high frame-rate is not necessary to depict the movements of the scene. By doing so, the additional processing needed for HFR content would only be present when it provides visual gains.

In this chapter, a content-dependent VFR model capable of determining the ideal frame-rate of HFR videos is proposed. The ideal frame-rate is the lowest possible frame-rate that does not affect the perceived video quality of the original HFR video signal. The proposed solution is based on a machine learning algorithm that takes as input spatial and temporal features extracted from the HFR video content to determine the ideal frame-rate among an set of possible frame-rates, limited to 30, 60 and 120 fps in this study.

The chapter is structured as follows. Section 7.1 reviews the related works on HFR video and variable frame-rate. Section 7.2 describes the considered ML approach and the tackled VFR classification problem. Then, the ground truth generation process is detailed in Section 7.3 followed by the description of the ML model training process in Section 7.4. Finally, the model prediction performance is evaluated in Section 7.5 and the related real-

time demonstration of the VFR coding scheme is presented in Section 7.6.

7.1 Related Work

7.1.1 High Frame-Rate Video

The UHDTV signal, defined in the ITU-R BT. 2020 recommendation [1], introduces a number of improvements over the HDTV [2] aiming at providing a better visual experience to the user. Along with a wider color gamut and an increased bitdepth, which allow to depict real colors and avoid ringing artifacts respectively, the key features of the UHDTV signal enabling a better depiction of live content are the higher spatial resolution - up to 3840x2160 and 7680x4320 pixels - and increased frame-rate - up to 120 fps. The different experiments that lead to the definition of each characteristic of the UHDTV signal are summarized in [3, 156].

Particularly, high frame-rate video has been an active field of research in the last decade, with the goal of avoiding well-known motion-related artifacts, namely flickering, motion blur and stroboscopic effect, which are present in traditional HDTV frame-rates of 60 fps and lower. Flicker is a phenomenon in which unwanted visible fluctuations of luminance appear on a large part of the screen and occurs at low refresh rates on non hold-type displays (e.g. Cathode Ray Tubes (CRTs)). Several studies [4, 157] have shown that flicker can be eliminated, for UHDTV signals, by simply using a frame-rate higher than 80 fps. The stroboscopic effect is the result of temporal aliasing, where the frame-rate is insufficient to represent smooth motion of objects in a scene causing them to judder or appear multiple times. At a given frame-rate, strobing can be reduced by lowering the shutter speed of the camera. However, a lower shutter speed also increases motion blur, which is caused by the camera integration of an object position over time, while the shutter is opened. Thus, strobing artifacts and motion blur can not be optimized independently except by using higher frame-rate [158].

Based on previous studies by Barten [159] and Daly [160], Laird *et al.* [153] define a spatio-velocity Contrast Sensitivity Function (CSF) model of the HVS taking into account the effect of eye velocity on sensitivity to motion. In [152], Noland uses this model along with traditional sampling theory to demonstrate that the theoretical frame-rate required to eliminate motion-blur without any strobing effect is 140 fps for untracked motion and as high as 700 fps if eye movements are taken into account. Since this theoretical critical frame-rate is not yet achievable, several subjective studies have investigated the frame-rate for which motion-related artifacts are acceptable for the HVS. In [161], Selfridge *et al.* investigate the visibility of motion blur and strobing artifacts at various shutter angles and motion speeds for a frame-rate at 100 fps. Their subjective tests showed that even at such

a frame-rate, all motion-blur and strobing artifacts can not be both avoided simultaneously. Kuroki *et al.* [151] conducted a subjective test with frame-rates ranging from 60 to 480 fps, concluding that no further improvements of the visibility of blurring and strobing artifacts were visible above 250 fps. Recently, Mackin *et al.* [150] have performed subjective tests on the visibility of motion artifacts for frame-rates up to 2000 fps, achieved using a strobe light with controllable flash frequency. The study concluded that a minimum of 100 fps was required to reach tolerable motion artifacts.

For the purpose of the UHDTV signal definition, several studies further investigated the importance of HFR for television [85, 155, 162]. Emoto *et al.* [162] showed that increasing the frame-rate from the traditional 60 fps to high frame-rate of 120 fps provides a significant visual quality improvement. It is also stated that a further increase to 240 fps would also improve the motion portrayal but to a much lesser extent than the transition from 60 to 120 fps. Salmon *et al.* [155] have also studied HFR for television, showing that at least 100 fps is required for improvements over HDTV, especially for content with high motion such as sports. Recently, with one of the first 65 inches UHD HFR prototype displays, Hulusic *et al.* [85] studied the joint and independent contributions of 4K resolution and HFR. The subjective tests carried out showed that the 2160p100Hz format enables a significant increase in visual quality over other configurations - 1080p50Hz, 1080p100Hz and 2160p50Hz - but also that the improvements are strongly content dependent.

7.1.2 Compression of HFR content and Variable Frame-Rate

Since the adoption of HFR in the future television standard, through the second phase of the DVB UHD standard [163], several studies on the compression of HFR content have been carried out. Authors in [154] investigate the impact of high frame-rate on video compression, focusing on the perceptual quality of different motion types and frame-rates at several bitrates. Using the test sequences of the public HFR dataset described in [164] compressed using an HEVC encoder, it is shown that HFR is beneficial and desirable, especially at high bitrate and even at the current HDTV broadcast data rates for sequences containing camera and/or simple motion, for which the encoder can make use of the increased temporal correlation to predict adjacent frames. Sugito *et al.* [165] showed that the overhead, in terms of bitrate, introduced by the increase from 60 to 120 fps is reasonable, with an optimal bit allocation of 6-7% of the total bitrate for the additional frames needed to achieve HFR capability. However, one of the main limitation of doubling the frame-rate is the additional encoding complexity, with a near 40% increase in encoding time.

VFR, where the image frequency can be adapted based on the signal characteristics, is one of the solutions to cope with the complexity and bitrate increases. In [166], a variable frame-rate algorithm is designed to skip frames when a constant image quality can not be kept due to a limited bit budget. In [167], Song *et al.* add motion smoothness to the tradi-

tional rate control criterion to decide when a frame can be skipped to reduce the overall bitrate. Other rate control algorithms are based on the scene movements to drive the frame-rate adaptation, either through thresholds on simple frame differences [168] or motion vectors for transcoding applications [169]. Considering other use cases with real-time low latency and low bitrate constraints, such as video conferencing, a trade-off between spatial and temporal quality is studied in [170] by adapting the encoding frame-rate and predicted output quality based on the HVS. For mobile communications, Usach *et al.* [171] use a variable GOP size in addition to the variable frame-rate to meet the 3G network constraints. However, these variable frame-rate solutions have been designed for videos of up to 30 fps compressed with H.264 and older codecs and are thus not suitable for the UHDTV use case tackled in this chapter.

Recently, VFR for HFR content has been investigated, aiming at offering a perceptually indistinguishable temporally downsampled video. A Support Vector Regression (SVR) is used in [172] to predict a satisfied user ratio - percentage of people who do not see the difference between original and lower frame-rates - which is then used to dynamically select the appropriate image frequency at a GOP or sequence level. The trained SVR uses complex and computationally demanding features, notably a visual saliency map for each frame, and the training set is only composed of up to 60 fps content thus making it unsuitable for real-time dynamic frame-rate selection on 120 fps source contents. Katsenou *et al.* train Bagged Decision Trees to predict the critical frame-rate at a sequence level [173]. The selected feature set is only composed of Farnbäck's Optical Flow (OF) [174], for the temporal aspect, and Gray Level Co-occurrence Matrix (GLCM) [175] for the spatial details contribution. In addition, the considered dataset consists of 22 test sequences with critical frame-rates of 60 or 120 fps. Thus, a good generalization of the VFR decision problem is hard to achieve with such few data to train and validate the model.

7.1.3 HFR oriented evaluation metrics

To evaluate the quality of the frame-rate adaptation, several metrics, with the goal of assessing the motion related artifacts, can be used. A rate and perceptual quality model is proposed in [176] to evaluate the quality based on the frame-rate and quantization stepsize. Ou *et al.* [177] add the spatial resolution as a parameter of their perceptual quality model in addition to the temporal and amplitude resolutions. Both these models require content dependent parameters to be computed before using the metric.

With the introduction of HFR content, several metrics have been developed to specifically assess the perceptual impact of frame-rate reduction. In [178], Nasiri *et al.* designs a spatio-temporal perceptual aliasing factor based on traditional sampling theory and a CSF model of the HVS to measure the temporal aliasing introduced by lowering the frame-rate. Authors in [179] focus on the analysis of highpass coefficients of a temporal wavelet de-

composition to compare a temporally downsampled version to the original 120 fps source content, resulting in good correlation with MOS values on the Bristol Vision Institute High Frame-Rate (BVI-HFR) database. More recently, a metric based on the measure of temporal motion smoothness to evaluate the impact of frame-rate adaptation has been proposed in [180]. The temporal motion smoothness is calculated by analyzing the local phase correlation of complex temporal wavelets coefficients, which shows promising results compared to subjective tests, with some issues for content with local and highly variable spatial motion.

7.1.4 Motion Blur Rendering and Video Frame Interpolation

In a pipeline using a VFR video format to transport the video, several processing steps could be added to improve the perceptual quality of the output video. Indeed, on one hand, motion blur can be synthesized when the frame-rate is lowered to render a video close to what would have been captured with a camera at the lower frame-rate and its corresponding shutter speed. This would reduce the stroboscopic effect due to the frame decimation thus improving the visual quality of the VFR video. Motion blur synthesis has been extensively studied in an effort to render synthetic images as real as possible [181]. These techniques mostly rely on the perfect knowledge of the depth and motion of the scene and are thus not compatible for a live broadcast use-case. More recently, a motion blur rendering algorithm using only two consecutive images as inputs to produce a motion blurred output have been designed in [182]. These promising results are balanced by the computationally demanding algorithm, due to the underlying CNN architecture used to synthesize motion blur.

On the other hand, since most display devices do not support a variable frame-rate, the VFR video must be temporally upsampled to the original higher frame-rate before displaying it. Thus, frame interpolation methods can be used to improve the temporal upsampling step in order to obtain a visually better displayed video. Video frame interpolation is a well-studied field with several existing approaches to the problem. The classical approach interpolates intermediate frames from the optical flow field [183] of the scene. Interpolated frames, whose quality highly depends on the accuracy of the computationally expensive optical flow computation [184], typically suffer from motion boundaries and severe occlusions thus showing strong artifacts, even with state-of-the-art optical flow algorithms [185]. More recent promising works rely on neural networks to either predict convolution kernels for each pixel used to generate the interpolated frames [186, 187] or directly predict the intermediate image pixel values [188]. However, these techniques involve a large number of convolutions, sometimes with large kernels (up to 41x41 for each pixel) to cope with large motion, thus making the computational demand unsuitable for real-rime use-cases.

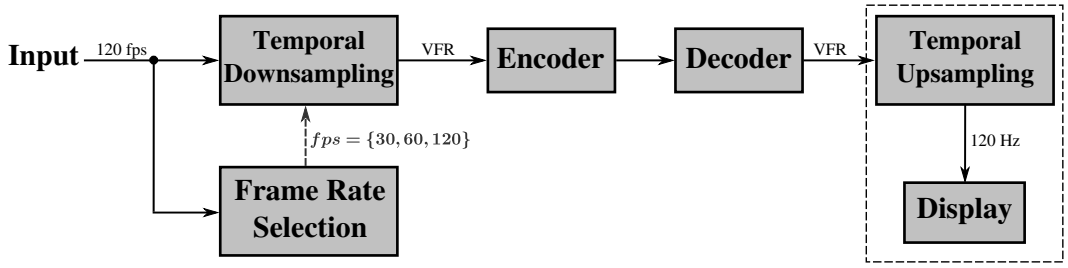


Fig. 7.1 Block diagram of the complete Variable Frame-Rate (VFR) coding scheme.

7.1.5 Objective and Motivation

Most existing algorithms containing variable frame-rate have been designed purely for rate control in 30 fps video encoding schemes, with the goal of skipping frames when the bit budget constraint can not be met. Such behavior does not take properly into account the impact on perceptual quality, making these solutions not suitable for HFR, which has been integrated to the UHDTV standard to improve motion portrayal.

Since the improvements brought by HFR capabilities are highly content dependent, several recent studies have based the frame-rate selection on perceptual factors to lower the frame-rate when there is no impact on visual quality. However, they use computationally expensive features and rely on a small dataset, not always composed of 120 fps content, to train and validate the variable frame-rate models. Thus, these solutions do not achieve good generalization of the problem. Moreover, they are not suitable for a real-time constraint, which is required for use-cases like live broadcast of, for example, sport events, that would periodically highly benefit from HFR.

In this chapter, real-time variable frame-rate for HFR source content is addressed. The proposed system, depicted in **Fig. 7.1**, relies on two RF classifiers to predict the critical frame-rate. It has been designed with the following objectives:

- Design a dynamic frame-rate selection at a GOP level with a perceptually invisible frame-rate changes.
- Real-time feature computation with low inference on the coding chain.
- Training and testing on a well dimensioned dataset containing various types of 120 fps content.
- Perceptual validation of the obtained objective performance through subjective evaluation test.
- Assess the bitrate and complexity savings within an HEVC encoding chain.

To meet the real-time constraint of live broadcast, state-of-the-art motion blur rendering and frame interpolation have not been integrated in the VFR pipeline. Instead, the proposed

system has been designed with simple frame decimation (resp. duplication) as a temporal downsampling (resp. upsampling) tool.

7.2 Random Forest Classifier for Variable Frame-Rate

This section briefly presents Random Forests (RFs) as a classification tool and introduces the method proposed in this work to reduce the frame-rate with no visual impact, i.e. the VFR decision problem, as a combination of two binary classification problems.

7.2.1 Background on Random Forests

Random Forests [189] are a common ML tool used to solve classification problems. A RF classifier is able to predict the value of a target variable, i.e. a class, based on a set of input variables, i.e. input features, using the majority vote of an ensemble of nearly independent decision trees.

A decision tree is constructed by first partitioning the training dataset, i.e. the features and the associated class of each sample, into two different subsets, called nodes. This process is performed recursively until either all the node samples belong to a single class or a tree constraint has been reached. At each node, each available input feature is evaluated for all its possible values, in order to achieve the best separation of the classes in the subsequent child-nodes.

In this work, the criterion used to quantify the quality of a split, given the feature F and its threshold value t, is based on the Gini impurity measure, a common metric for Decision Trees [190]. It is computed as follows

$$I_G(\mathbf{D}) = \sum_{c \in \mathbf{C}} \mathbb{P}(c \,|\, \mathbf{D}) \cdot (1 - \mathbb{P}(c \,|\, \mathbf{D})), \tag{7.1}$$

with $\mathbf{D}$ the sample set under consideration, $\mathbf{C}$ the set of possible class labels and $\mathbb{P}(c \,|\, \mathbf{D})$ the conditional probability of class c given the sample set $\mathbf{D}$.

The best split is then obtained by finding the pair (F,t) that maximizes the Mean Decrease Impurity (MDI) ΔI_G defined by Equation (7.2)

$$\Delta I_G(\mathbf{D},F,t) = I_G(D) - \frac{|\mathbf{D}_L|}{|\mathbf{D}|} \cdot I_G(\mathbf{D}_L) - \frac{|\mathbf{D}_R|}{|\mathbf{D}|} \cdot I_G(\mathbf{D}_R), \tag{7.2}$$

with $\mathbf{D}_L = \{x \in \mathbf{D}, F(x) < t\}$ (resp. $\mathbf{D}_R = \{x \in \mathbf{D}, F(x) \geq\}$) the subset of sample set D for which each sample x has a value of feature $F(x)$ smaller (resp. larger) than threshold t and $|\mathbf{D}|$ the cardinal of a set $\mathbf{D}$.

To minimize the correlation between trees of the RF, the bootstrap aggregating, or bag-

ging, technique [191] is used to construct the forest. This consists in training each tree T_i with a different subset $\mathbf{D}_i$ of the input data sample set $\mathbf{D}$. Each $\mathbf{D}_i$ is obtained by a uniform sampling of D with replacement, i.e. replacing discarded samples by duplicates of a selected one. In addition and to further reduce the correlation between trees, only a random subset of the features, here $\sqrt{n}$ features with n the total number of input features, are evaluated at each node to find the best available split.

7.2.2 VFR Classification Problem

The proposed solution aims at predicting when the frame-rate can be reduced, by discarding frames, without any perceptual impact on the quality of the original input HFR video. In an effort to keep the number of possible frame-rates reduced and obtain a regular frame decimation process, two frame-rates, *60 fps* and *30 fps*, were identified as potential candidates in addition to the original frame-rate of *120 fps*. The VFR decision problem thus becomes a three-class classification, as depicted in **Fig. 7.2**.

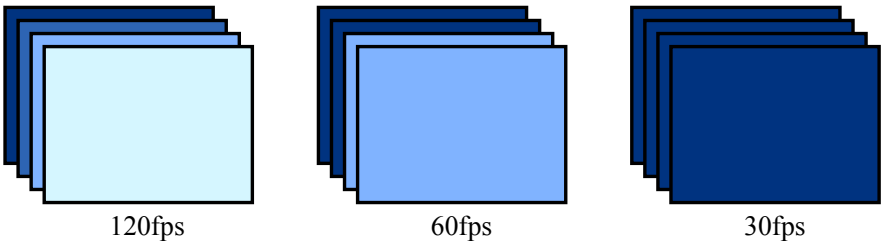


Fig. 7.2 Possible classes of the VFR classification problem. *Frames in the same color represent the same image repeated at multiple time-stamps to match the chosen frame-rate with the original 120fps one.*

In this work, a combination of two successive binary RF classifiers has been chosen to solve the classification problem, as depicted in **Fig. 7.3**, instead of directly training a forest with multi-class outputs. This decision leads to a better overall performance by training both classifiers independently on separate datasets and features. Indeed, in addition to the specialization of each binary classifier, almost all samples of the database for training either one or both classifiers while keeping balanced training datasets, as described in Section 7.3, thus increasing the accuracy of the overall model.

The first RF classifier, named *120fps - FD*, is specialized in deciding whether the frame-rate must remain *120 fps* or a Frame Decimation (FD) can be applied without impacting the visual quality. If the *120 fps* class is chosen by the first classifier, the frame-rate prediction process is stopped and no FD is applied. Otherwise, the prediction process continues by requesting the second classifier, named *60fps - 30fps*, which aims at selecting the appropriate lower frame-rate if a FD is applied on the input HFR video.

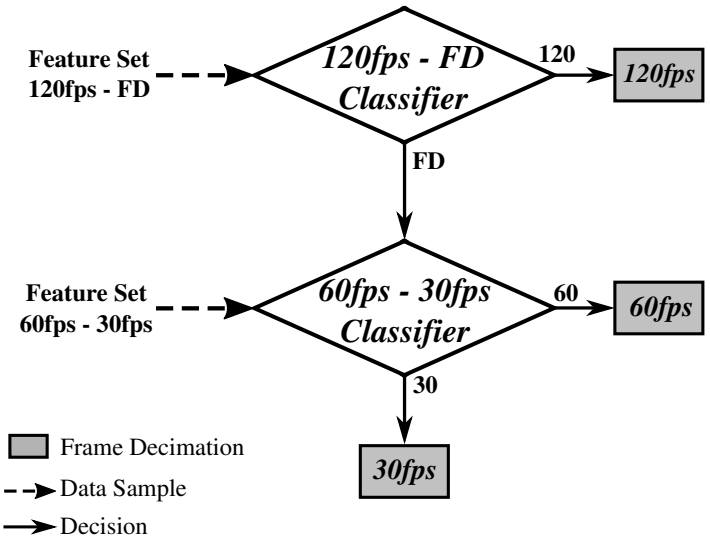


Fig. 7.3 Overall prediction scheme with cascaded binary RF classifiers.

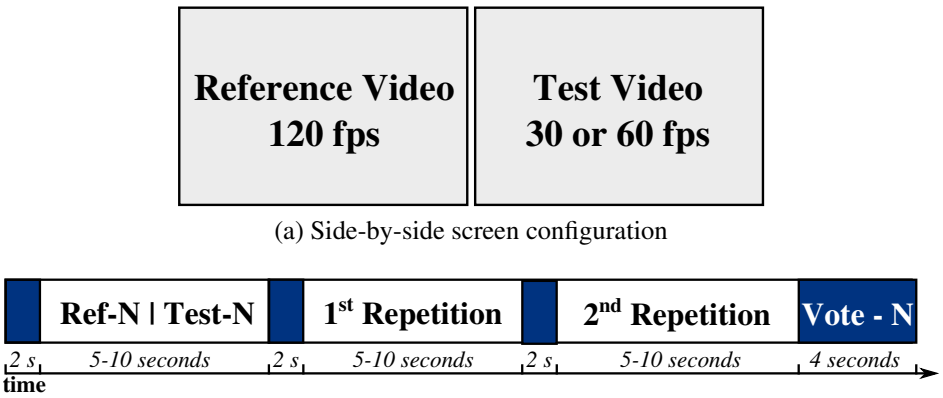
7.3 Ground Truth Generation

One of the most crucial steps towards training a supervised RF model is to gather a dataset which will be used as ground truth, i.e. examples - features representing a sample and the sample class label - used by the model for learning to predict. It is thus important to have a ground truth that contains a good representation of all the real cases the model could encounter in order to achieve a good generalization of the problem. This section focuses on detailing the ground truth generation process, necessary to obtain the datasets used to train both RF classifiers. The HFR database is first presented, followed by the detailed methodology for subjectively determining the critical frame-rate labels. Then, the creation of the dataset is described, with the composition of the balanced training set on one hand and the feature extraction process on the other hand.

7.3.1 HFR Video Database

To the best of our knowledge, there is no publicly available dataset for the VFR classification problem except for the BVI-HFR database [164]. However, this database only contains 22 videos, which is rather small to train a reliable model. In addition, the temporal downsampling technique used to create the lower frame-rates in this database is frame averaging, whereas the chosen method in this work is frame decimation, which could lead to different decisions on the appropriate frame-rate.

Therefore, a new database has been gathered, composed of 375 native HFR video clips of 5 to 10 second. These clips are all uncompressed and stored in YUV format with 4:2:0 chroma subsampling and 8-bit depth. Their original frame-rate is 120 fps and spatial reso-



(a) Side-by-side screen configuration

(b) Basic Test Cell

Fig. 7.4 Ground truth generation SDSCE evaluation method.

lution 1920x1080 pixels (HD) - a downsampling has been performed for source videos of higher spatial resolution using Lanczos3 filter banks [100].

The database includes video clips from the BVI-HFR dataset [164], together with b<>com, Harmonic and other non-publicly available test sequences. In order to later evaluate the trained model on unseen data samples, 15 sequences, with heterogeneous spatio-temporal characteristics, have been extracted from the database, leaving 360 video clips to annotate before training both RF classifiers.

7.3.2 Optimal Frame-rate Decision Methodology

The ground truth generation process requires each video of the database to be assigned a critical frame-rate chosen among the three considered ones in the VFR classification problem. A subjective test has thus been carried out with the objective of finding the lowest frame-rate for which no visual degradation can be observed compared to the original 120 fps video. Following the ITU-R BT.500-13 recommendation [28], the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) protocol has been used with a binary scale - either identical or visible difference. Therefore, subjects were asked, for each video of the database, if there was a visible difference between the known reference (left screen) and the lower frame-rate, either 30 or 60 fps, test video displayed on the right screen, as depicted in **Fig. 7.4a**.

The test was composed of 750 Basic Test Cells (BTCs), randomly divided into 30-minute test sessions. Each BTC is 40-second long and is composed of a 2-second message announcing the test video index followed by the side-by-side display of the reference and test videos with two repetitions.

A 2-second break displaying a mid-gray image has been added between each repetition. The BTC is concluded by a 4-second message asking the viewer to vote, as shown in **Fig. 7.4b**. Due to the large duration of the subjective test, only five viewers, experts in video processing, participated in the whole database annotation. The final frame-rate de-

Table 7.1 Database critical frame-rate distribution.

	Optimal frame-rate		
	30	60	120
# of shots w/ uniform motion	78	184	167
# of 4-frame chunks	3749	22327	19996

cision is the lowest frame-rate for which the majority of expert viewers did not notice any visible difference with the reference 120 fps video.

The tests were conducted in a controlled laboratory environment, following the ITU-R Rec BT.500-13 [28]. Two identical 27-inch screens capable of displaying 120 fps content (Asus RoG Swift PG278Q) were used side-by-side, aligned and placed at a distance from the viewer position of three times the screen height. Each participant has been screened to ensure (corrected to) normal visual acuity and normal color vision.

7.3.3 Balanced Dataset Composition

The results of the expert subjective test are summarized in **Table 7.1**. As can be seen, the sequences are not evenly distributed over the three possible frame-rates due to a large proportion of the available content being captured with HFR-capable devices containing high motion content, for which frame decimation downsampling to 30 fps is critical. Since the goal is to allow for a frame-rate adaptation at the lowest possible level, i.e. a frame-rate decision for every chunk of 4 frames to keep a regular frame decimation process, the sequence-level labels obtained via the subjective test have been extended to 4-frame chunk-level labels. Thus, for sequences with significant motion discontinuities, video shots with uniform motion can be identified and assigned different labels in a single sequence. Based on the observations made by the experts after the subjective test, a total of 429 video shots with uniform motion have been extracted from the 360 native HFR sequences. This refinement allows for a more accurate annotation of the database, thus avoiding chunks being annotated with inconsistent labels, for instance a chunk not containing any movement associated to the *120fps* class. However, such cases could still remain in the ground truth due to the difficulty to identify the motion discontinuities with a precision of less than four frames.

From this ground truth, two different datasets were created to train the *120fps-FD* and *60fps-30fps* RF classifiers. The first one contains all samples of the *120fps* class as well as those from the *FD* class. The *FD* class is composed of all *30fps* samples and a random subset of the *60fps* samples. The amount of selected *60fps* samples has been chosen to produce a balanced dataset, i.e. to roughly obtain the same sample size for both the *120fps* and *FD* classes. The second dataset, used to train the *60fps-30fps* classifier, is comprised of the

30fps samples and a random subset of the *60fps* samples, whose size has also been chosen to produce a balanced dataset. The choice of balancing datasets has been made because the unbalanced class distribution in the database does not necessarily represent the distribution of media content in a broadcast context, the chosen use-case for the proposed solution, but rather relates to the current difficulty to find HFR content with low motion. This is due to the fact that most of the currently available HFR content has been shot to demonstrate the gain in perceptual quality and motion portrayal brought by the technology.

7.3.4 Feature Extraction

The goal of a feature set is to gather the different metrics relevant to the considered classification problem that would help discriminate the output classes from one another. For the VFR classification problem, a first feature would intuitively be the motion information, e.g. the motion vectors between two consecutive frames. Indeed, high movement in a source HFR video will likely lead to visible temporal aliasing, i.e. stroboscopic effect, if a lower frame-rate is used after frame decimation. In addition, since motion blur is not added during the temporal downsampling process used in this work, lowering the frame-rate could introduce visible jerkiness in high motion videos.

For the three considered frame-rate classes, flickering, the other well-known motion-related artifact, can appear in highly textured areas where the local variation in luminance between two consecutively displayed frames would be visible at lower frame-rates. In an effort to capture this phenomenon in the feature set, the pixel luminance values and directional gradients can be used.

Based on the performed expert viewing sessions, it has been observed that small objects with high velocity, which would not necessarily be detected by the motion vectors depending on the used motion estimation algorithm, could induce visible artifacts at lower frame-rates. To take this observation into account, a simple metric capable of detecting both global displacements and small moving objects has been designed. This metric is based on the thresholding of the difference between two consecutive frames. First, the frame difference $\mathbf{D}_n(i,j)$, i.e. the difference in pixel value of the luminance plane at the same location in space (i,j) between the n^{th} frame $\mathbf{F}_n$ and the preceding one $\mathbf{F}_{n-1}$, is computed for each pixel using Equation (7.3)

$$\mathbf{D}_n(i,j) = |\,\mathbf{F}_n(i,j) - \mathbf{F}_{n-1}(i,j)\,|. \tag{7.3}$$

Then, a thresholding operation is performed on the frame difference image, defined as follows

$$\mathbf{A}_{n,Th}(i,j) = \begin{cases} 1 & if\ \mathbf{D}_n(i,j) \geq Th \\ 0 & if\ \mathbf{D}_n(i,j) < Th \end{cases}, \tag{7.4}$$

with $\mathbf{A}_{n,Th}$ the resulting thresholding activation map for the n^{th} frame and a threshold Th.

(a) (b)

Fig. 7.5 Example of thresholded motion difference with (a) the original image of the *Jokey* sequence and (b) the thresholding activation map with threshold $Th = 25$.

Fig. 7.5 depicts an example with both the original image and the resulting thresholded frame difference image.

The designed feature set is thus based on the following feature maps:

- **NormMV**, **HorMV**, **VerMV:** maps respectively representing the MVs norm, horizontal coordinate and vertical coordinate.
- **ThreshDiffMap:** thresholded frame difference map as defined in Equation (7.4).
- **GradMag**, **GradHor**, **GradVer:** maps respectively representing the Sobel gradient magnitude, horizontal gradient and vertical gradient.
- **Luma:** pixel luminance map.

For each map, several scores have been computed, namely the mean value, the standard deviation, the maximum value and the mean of the 10% highest values, to produce a total of 32 different features that will serve as an initial feature set for the training of both the considered RF models.

7.4 Random Forest Training Process

Once the ground truth is available and the features computed, the RF models can be trained to solve the VFR classification problem. This section focuses first on the performance evaluation process, necessary to assess and optimize the quality of the model critical frame-rate prediction task. Then, a feature selection process, used to reduce the initial feature set to only the relevant features for each binary classifier, is presented. Finally, a description of the optimization of several model hyper-parameters is given, followed by a presentation and analysis of the classification results.

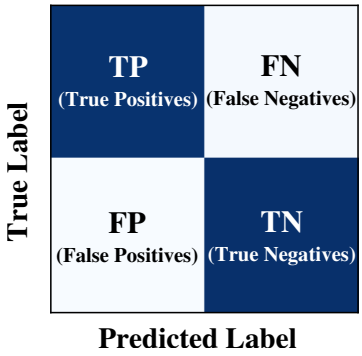


Fig. 7.6 Error definitions for binary classification.

7.4.1 Model Evaluation Process

In order to optimize a ML classifier, it is necessary to use a metric capable of evaluating the model classification performance to find the best parameters. To do so, several common metrics, namely precision, recall and F1-score [192], can be used. First, the trained model confusion matrix, as depicted in **Fig. 7.6**, has to be computed from the true and predicted labels of the dataset samples. Then, once the different quantities of the confusion matrix, namely True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN), are available either as number of samples or normalized probabilities, the precision, recall and F1-score can be computed as follows, with $\mathbf{C} = \{c_1, c_2\}$ the set of classes for the binary classifier under test

$$precision(\mathbf{C}) = \frac{1}{|\mathbf{C}|} \cdot \sum_{c_i \in \mathbf{C}} \frac{TP(c_i)}{TP(c_i) + FP(c_i)}, \tag{7.5}$$

$$recall(\mathbf{C}) \quad = \frac{1}{|\mathbf{C}|} \cdot \sum_{c_i \in \mathbf{C}} \frac{TP(c_i)}{TP(c_i) + FN(c_i)}, \tag{7.6}$$

$$F_1\text{-}score(\mathbf{C}) = \frac{2}{|\mathbf{C}|} \cdot \sum_{c_i \in \mathbf{C}} \frac{precision(c_i) \times recall(c_i)}{precision(c_i) + recall(c_i)}, \tag{7.7}$$

As for any binary classification problem, the goal is to maximize the confusion matrix main diagonal values, i.e. the number of TP and TN representing the correct predictions. This can be achieved by maximizing precision, recall, or F1-score during the training process, depending on the considered classification problem and the criticality of each error type. For the VFR classification problem, the main goal is also to minimize the critical errors - predicted frame-rate lower than the ground truth - which would potentially induce visible temporal artifacts thus greatly reducing the output visual quality. To emphasize these critical errors and avoid them in the final model, another performance evaluation metric M_{crit} has been designed, as a combination of the precision of the lower frame-rate class and the

recall of the higher frame-rate class, using Equation (7.8)

$$\begin{aligned} M_{crit}(\mathbf{C}) &= \frac{1}{|\mathbf{C}|} \cdot [precision(c_1) + recall(c_2)] \\ &= \frac{1}{|\mathbf{C}|} \cdot \left[\frac{TP(c_1)}{TP(c_1)+FP(c_1)} + \frac{TP(c_2)}{TP(c_2)+FN(c_2)} \right] \\ &= \frac{1}{|\mathbf{C}|} \cdot \left[\frac{TP(c_1)}{TP(c_1)+FP(c_1)} + \frac{TN(c_1)}{TN(c_1)+FP(c_1)} \right] \end{aligned} \tag{7.8}$$

with $\mathbf{C} = \{c_1, c_2\}$ the set of ordered classes - frame-rate of c_1 lower than the frame-rate of c_2. This metric has been used together with the F1-score to assess the quality of RF models for both the feature selection process and hyper-parameter tuning described in the next sections.

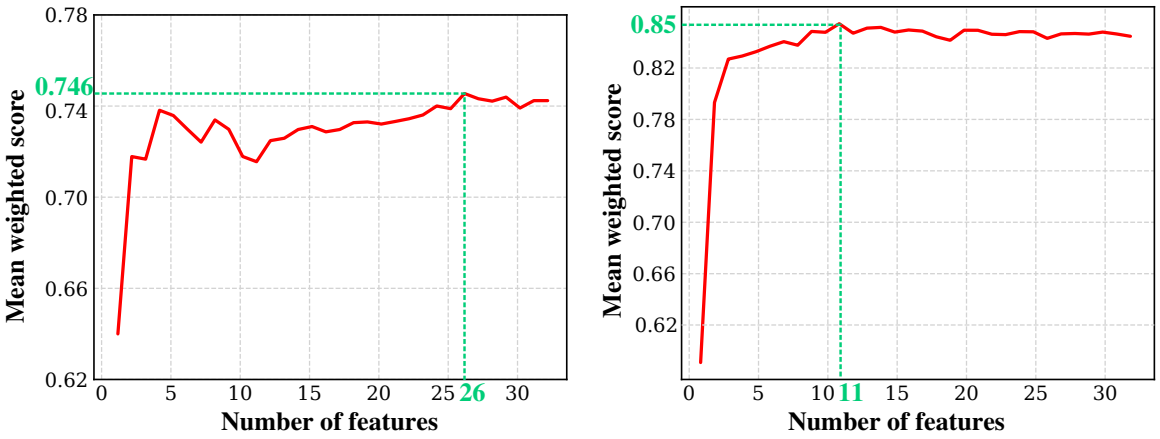
7.4.2 Feature Selection

In an effort to limit the model complexity and improve its performance, a dimensionality reduction algorithm has been used on the proposed initial feature set. Indeed, by only selecting the relevant features, thus removing features carrying useless information for the considered classification problem, both the feature computation time and training time are greatly reduced. Additionally, model over-fitting is also decreased when the size of the feature set is reduced due to the reduction of noise in the input data and the elimination of highly correlated features, i.e. features that would carry the same information about the target variable.

In this work, a Recursive Feature Elimination (RFE) process has been used to reduce the dimension of the initial feature set. It consists in recursively evaluating the model performance on a dataset and a feature in which the least important feature is removed after each iteration. The feature importance is computed in terms of mean decrease in Gini impurity, i.e. the average capacity of a feature to reduce the Gini impurity computed a given tree node using Equation (7.1). When the feature set size reaches the minimum tested dimension of 2, the feature set leading to the best model performance among all the tested dimensions is selected as the final feature set.

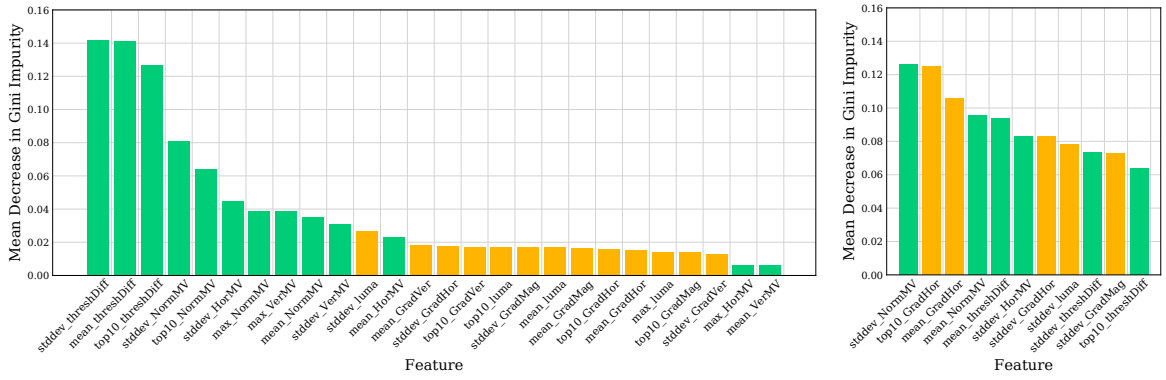
This process has been performed independently for both proposed RF models with the same initial feature set but leading to a different optimal feature set size for each binary RF classifier, respectively 26 and 11 features for the *120fps-FD* classifier and *60fps-30fps* classifier, as depicted in **Fig. 7.7**.

Fig. 7.8a shows the list of selected features for the *120fps-FD* RF classifier with their corresponding feature importance. As can be observed, the most relevant features to discriminate samples from both classes are based on the two motion features *ThreshDiffMap* and *NormMV*. This correlates well with the observations made by the experts during the ground truth annotation subjective tests. Indeed, it was pointed out that above a certain amount of movement, either from a moving camera or an object with high velocity - which can be captured by both metrics -, a stroboscopic effect as well as jerkiness, due to lack of motion blur, could easily be detected with frame-rates lower than 120 fps. Most of the fea-



(a) 120fps-FD RF model (b) 60fps-30fps RF model

Fig. 7.7 Recursive Feature Elimination process with weighted (F1,M_{crit}) score.



(a) 120fps-FD RF model (b) 60fps-30fps RF model

Fig. 7.8 Feature importance measured with Mean Decrease in Gini Impurity. *yellow: spatial features, green: motion features*

tures based on spatial measures are present in the optimized feature set, with a significantly lower importance compared to the aforementioned motion features. This tends to indicate that flickering becomes an important criterion to keep a high frame-rate when the amount of movement does not induce other motion artifacts.

For the *60fps-30fps* RF model, the selected features and their importance are depicted in **Fig. 7.8b**. As for the first RF model, the features based on the motion vectors, *NormMV*, have a high capacity to discriminate samples from both classes. However, spatial features, based on the *Luma* and *GradHor* feature maps, hold a significantly higher importance compared to the first model, indicating that flickering mostly occurs at 30 fps for most of the videos of the training dataset.

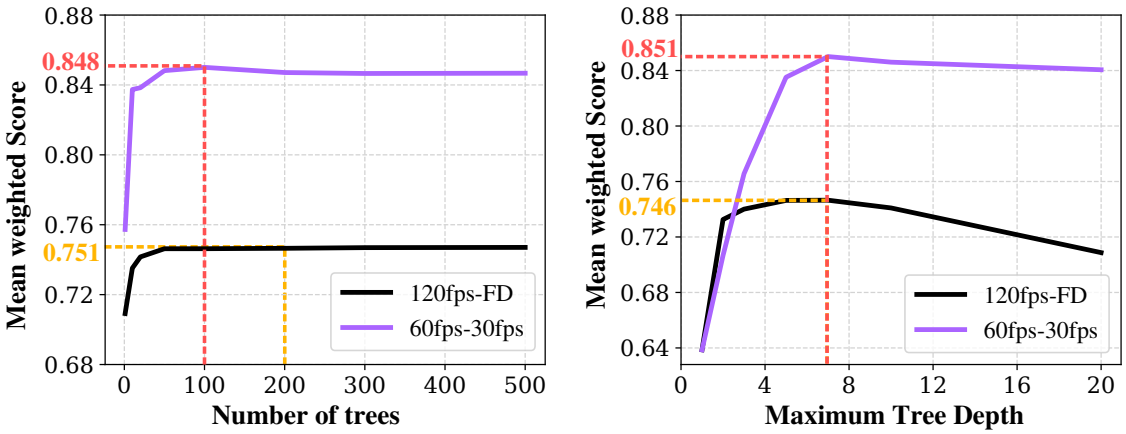


Fig. 7.9 Per-classifier hyper-parameter optimization.

7.4.3 Hyper-parameter Tuning

The performance of a ML model can also be improved by optimizing the different hyper-parameters characterizing it. Given these parameters, which are set beforehand, the training process can learn the other model parameters from the data. For random forests classifiers, several hyper-parameters are available to describe the forest and tree constructions, including the number of trees, the maximum depth of a tree, the minimum number of samples required to allow further splitting at a tree node or the minimum number of samples needed at a leaf node.

The optimization process aims at finding the combination of hyper-parameters that gives the best performance after training the model. This is done by exhaustively testing every hyper-parameter combination from a list of possible values for each parameter. In this work, only the number of trees comprised in the forest and the maximum tree depth have been jointly optimized.

Fig. 7.9a depicts the influence of the number of trees on both RF models. For the *120fps-FD* classifier, the prediction performance becomes nearly stable above 50 trees with a maximum value reached at 200 trees. The performance curve is steeper for the *60fps-30fps* classifier, and begin to slowly decrease when the number of trees becomes larger than 100. Since the aim is to use the models in a real-time application, the number of trees should be as small as possible while maximizing the prediction score. Hence, the chosen final values are 200 and 100 trees for the *120fps-FD* and *60fps-30fps* classifiers respectively.

For the depth hyper-parameter, depicted in **Fig. 7.9b**, the final chosen value is 7, the one maximizing the prediction performance for both classifiers. Indeed, a decrease in performance can be observed for larger than 7 maximum depth due to over-fitting the training data. For depth values lower than 5, the performance significantly drops for the *60fps-30fps*

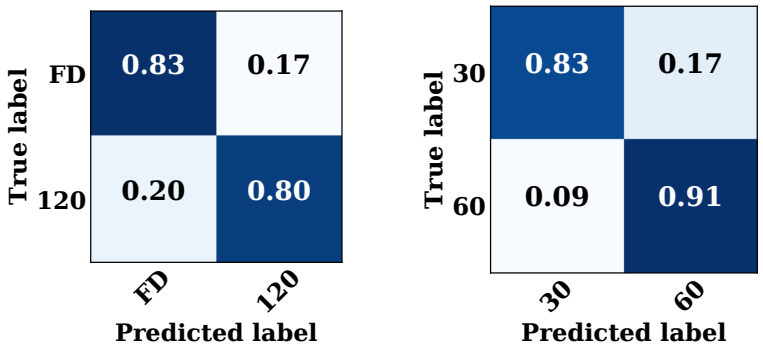
classifier due to the lower number of trees and more distributed feature importance.

7.4.4 Classification Results

With the final models and their associated feature sets, an in-depth analysis of the prediction capability of the models, both individually and combined to form the overall VFR prediction scheme, can be conducted. It is important to note that the models have been trained using a 10-fold cross-validation so that the considered performance is a combination of the results from the validation fold of each iteration. This means that each tested sample prediction presented in the different confusion matrices has been obtained without using the validation sample for training. Additionally, the training set has not been shuffled, so that chunks from a same sequence could not be in the training and validation folds at the same time, thus avoiding a highly biased performance evaluation.

Fig. 7.10 shows the resulting confusion matrices of both RF models, individually. For the *120fps-FD* classifier, error rates of 20% and 17% can be observed for the *120fps* and *Frame-rate Decimation (FD)* classes respectively, which represent a good performance considering the VFR classification problem and its imperfect groundtruth. Indeed, the frontier between annotating a sequence with a *120fps* label and a *60fps* label can be difficult to maintain consistent during the processing of the 360 videos of the training set. In addition, as detailed in Section 7.3.3, several sequences with high motion discontinuities have been separated into shots with different labels. Since this motion change frontiers could only be determined subjectively and not at a precise frame level, some dataset samples, i.e, 4-frame chunks, located at these frontiers could have been annotated with incorrect labels. Therefore, the error rate is likely to be over-estimated, leading to a visible motion artifacts rate lower than the observed 20%. For the *60fps-30fps* model, correct prediction rates are respectively 83% and 91% for the *30fps* and *60fps* classes. Critical errors, defined as critical frame-rate under-estimation, represent 9% of the *60fps* class samples. This rate can be problematic since the under-estimation with a frame-rate of 30 fps could lead to severe visible motion artifacts. However, the same aforementioned remark concerning imperfect ground truth applies to the training dataset of the *60fps-30fps* model. The proportion of frame-rate over-estimation errors, equal to 17%, does not impact the visual quality and is thus not as prejudicial as the critical errors.

The overall VFR prediction scheme confusion matrix, obtained by combining the cross-validation validation-fold predictions of both models, is depicted in **Fig. 7.11**. It is important to note that only a subset of randomly chosen *120fps* class samples has been used to compute the overall prediction scheme confusion matrix so that the three classes have the same number of samples. The observed performance is consistent with the individual RF model prediction results, with good probabilities of correct prediction. The only significant change is the 69% correct prediction rate for the *60fps* class, which can be explained by the fact



(a) 120fps-FD classifier (b) 60fps-30fps classifier

Fig. 7.10 Individual confusion matrices for a 10-fold cross-validation training of each proposed classifier with their respective dataset.

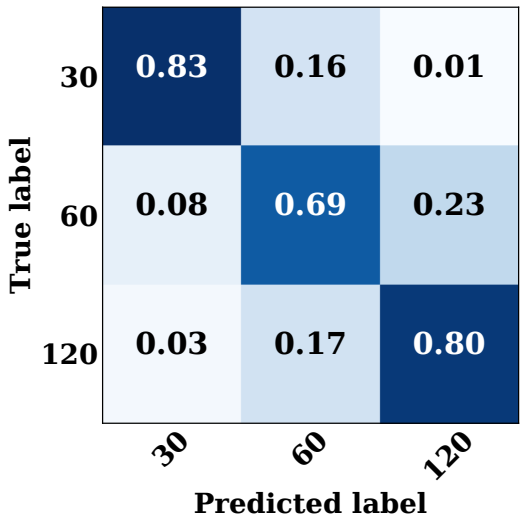


Fig. 7.11 Confusion matrix for the proposed overall prediction scheme.

that it is the intermediate class, thus sharing characteristics with the other two classes which makes the discrimination of the class samples harder to generalize. In addition, the *extreme* errors, i.e. the critical under-estimation of a *120fps* sample with a *30fps* predicted label or the exact inverse, are rarely occurring with rates of 3% and 1%, respectively. This tends to bolster the hypothesis on the ground truth being imperfect due to possibly unstable/blurry annotation frontiers between adjacent labels. If this hypothesis is correct, the combined prediction model should lead to VFR output video sequences visually identical to the HFR input. However, the compression and encoding complexity gains should be slightly lower than with ground truth labels. The next section aims at verifying this statement.

Table 7.2 Test set sequence characteristics.

Sequence	Resolution	Frame-Rate	#frames	Source
bouncyball	1920x1080	120	1200	BVI-HFR [164]
flowers	1920x1080	120	1200	
library	1920x1080	120	1200	
martial-arts	1920x1080	120	1200	
pour	1920x1080	120	1200	
Katana	3840x2160	120	1200	b<>com
Refuge1	3840x2160	120	1200	
Refuge2	3840x2160	120	1200	
Refuge3	3840x2160	120	1200	
Refuge4	3840x2160	120	1200	
Rowing1	3840x2160	120	1200	
Rowing2	3840x2160	120	1200	
NYCBike	3840x2160	120	1553	Harmonic
Rugby6	3840x2160	120	1113	
Rugby7	3840x2160	120	1561	

7.5 Results and Analysis

Before analyzing the coding performance of the VFR coding scheme, the visual quality of the output VFR video must be evaluated to assess whether the RF model frame-rate decisions are preserving the perceptual quality compared to the HFR source video. This section first describes the characteristics of the test set sequences and the chosen subjective evaluation methodology. Then, the results of the subjective tests are detailed and discussed for both uncompressed and compressed VFR videos. Finally, the coding performance of VFR coding scheme is presented in terms of bitrate savings and complexity reduction.

7.5.1 Specific Test Datasets and Subjective Tests Motivations

A total of 15 sequences has been selected to validate the performance of the VFR model. These sequences are unknown to the model, i.e. they have not been used during the cross-validation training of both binary RF classifiers. Table 7.2 summarizes the characteristics of the source sequences and **Fig. 7.12** depicts the first frame of each test sequence.. The input frame-rate is 120 fps for all test sequences and their duration ranges between 9 and 13 seconds. Source content with a higher resolution than 1920x1080 pixels has been downsampled with Lanczos3 filers [100] to ensure consistency during the subjective test.

(a) Refuge1

(b) Rowing1

(c) Rugby7

(d) library

(e) bouncyball

(f) Refuge4

(g) Rowing2

(h) Rugby6

(i) flowers

(j) martial_arts

Fig. 7.12 Example first frames of every sequence dataset used for the subjective tests.

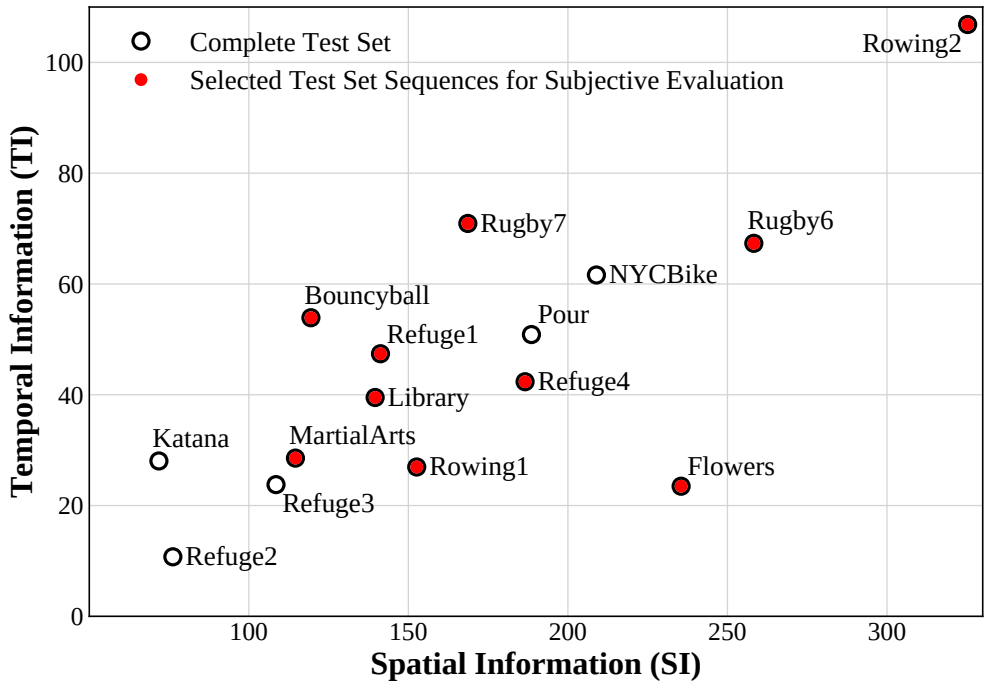


Fig. 7.13 SI-TI characteristics for test sequences of the three considered sets.

The test set sequences have been selected from various sources to cover a wide range of spatio-temporal characteristics both in terms of temporal and spatial information (SI and TI), as recommenced in [29] and shown in **Fig. 7.13**. They also depict several use-cases, including sporting events and movie-type clips, in addition to the more common natural video content.

In order to generate the final VFR model, both RF classifiers have been retrained with their respective whole datasets as well as their feature sets and hyper-parameters determined via cross-validation. The prediction results for the test set are depicted in **Fig. 7.14**. As can be observed, correct prediction rates reach 92%, 77% and 84% for the *30fps*, *60fps* and *120fps* classes, respectively, showing the capacity of the model to generalize the VFR classification problem to unknown data. The slightly better prediction results for the test set compared to the cross-validation predictions presented in Section 7.4.4 may be explained by the more accurate ground truth labels obtained for the test set, thus minimizing the labeling issue previously raised. In addition, the low amount of samples falsely predicted with a *30fps* class label should lead to a good perceptual quality of the VFR output videos, very close to the HFR source content.

To verify this statement, a subjective test comparing the uncompressed HFR and VFR videos has been designed. The *30fps* and *60fps* versions, obtained by frame decimation, have also been introduced in the subjective evaluation to assess the interest of variable frame-rate compared to systematic temporal downsampling in terms of perceived quality.

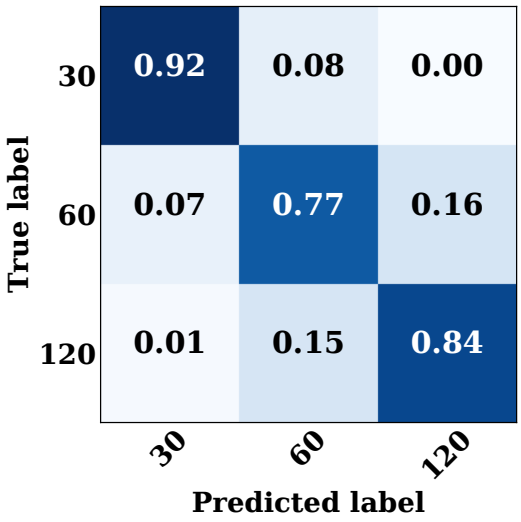


Fig. 7.14 Cascaded RF model prediction performance on test set as confusion matrices.

7.5.2 Subjective Evaluation Methodology

The considered subjective evaluation aims at assessing the effect of a system, here the VFR coding scheme, on the visual quality. For this kind of test, the ITU-R BT.500-13 recommendation [28] proposes the Double Stimulus Continuous Qualisty Scale (DSCQS) method, which consists in showing the observer pairs of videos - the un-processed source content and the same sequence processed with the system under test - and asking the observer to rate the quality of both sequences. The grading scale is a continuous vertical scale divided into 5 equal parts corresponding to the common 5-level ITU-R quality labels: *Excellent*, *Good*, *Fair*, *Poor* and *Bad*.

For each test session, a series of video pairs is presented to the observer in a random order, to distribute the degrees of quality impairments over the entire session. Each pair of videos is internally random, i.e. the observer is not aware of the position of the reference un-processed video (A or B), which is presented twice, successively. **Fig. 7.15** depicts the structure of a BTC presenting a pair of videos to assess. As can be observed, each BTC begins with a 2-second message indicating the id number of the current test point and ends with a message asking to vote. In addition, each display of a 10-second sequence is preceded by a 1-second message indicating if the following video is A or B on the answer sheet, making the total duration of a BTC equal to 50 seconds.

A total number of 10 sequences have been selected within the test set for the subjective test, as indicated in **Fig. 7.13**. The sequence set has been formed to cover a wide range of spatio-temporal characteristics and content types. For each sequence, 4 frame-rate pairs have been evaluated by the observers: *120fps vs 120fps*, *120fps vs VFR*, *120fps vs 60fps* and *120fps vs 30fps*. Therefore, a total of 40 BTCs were presented to each observer, randomly

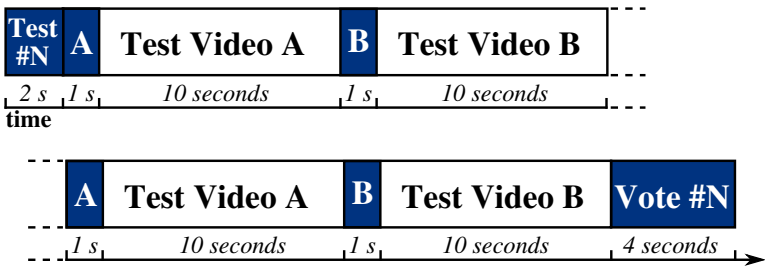


Fig. 7.15 Subjective test BTC presentation structure for DSCQS evaluation method.

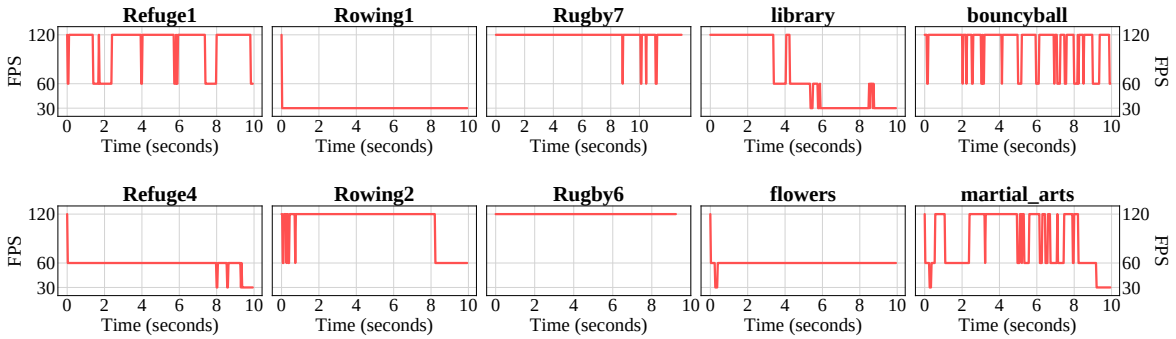


Fig. 7.16 Frame-rate decisions of the VFR algorithm for test set sequences.

divided into two 20-minute sessions separated by a 10-minute break. VFR video sequences have been obtained using the predicted frame-rates resulting from the proposed VFR model. **Fig. 7.16** depicts the evolution of frame-rate decisions over the duration of each sequence of the subjective evaluation test set. As can be observed, the predicted frame-rates are highly dependent on the test sequence, as expected considering the wide range of spatial and temporal information characteristics for the selected sequence set. In addition, the predicted frame-rates also vary over-time for most of the test sequences, demonstrating the interest of the 4-frame level of granularity proposed for frame-rate decisions.

The test was conducted in a controlled laboratory environment, with a viewing distance fixed to 3 times the screen height. A 65-inch LG OLED B6 display with HFR capabilities and peak luminance of 340 cd/m^2 has been used for both subjective tests. During the whole duration of the tests, all internal post-processing were disabled to avoid any impact on the perceived quality. Each test sequence in raw format (YUV 4:2:0 and 8-bit precision) has been encoded using the *libx265* encoder at 100 Mbps in order to be presented to the TV set via USB3 interface. Special care has been taken to ensure that the encoding needed for display did not introduce any 'coding' artifacts. A total of 19 participants took part in the subjective test. They were aged between 20 and 53 with (corrected-to) normal vision acuity and color vision. A post-screening analysis of the results has been carried out, according to the method described in ITU-R Rec. BT.500-13, to detect and reject the outliers before computing the MOS and Differential Mean Opinion Score (DMOS) values.

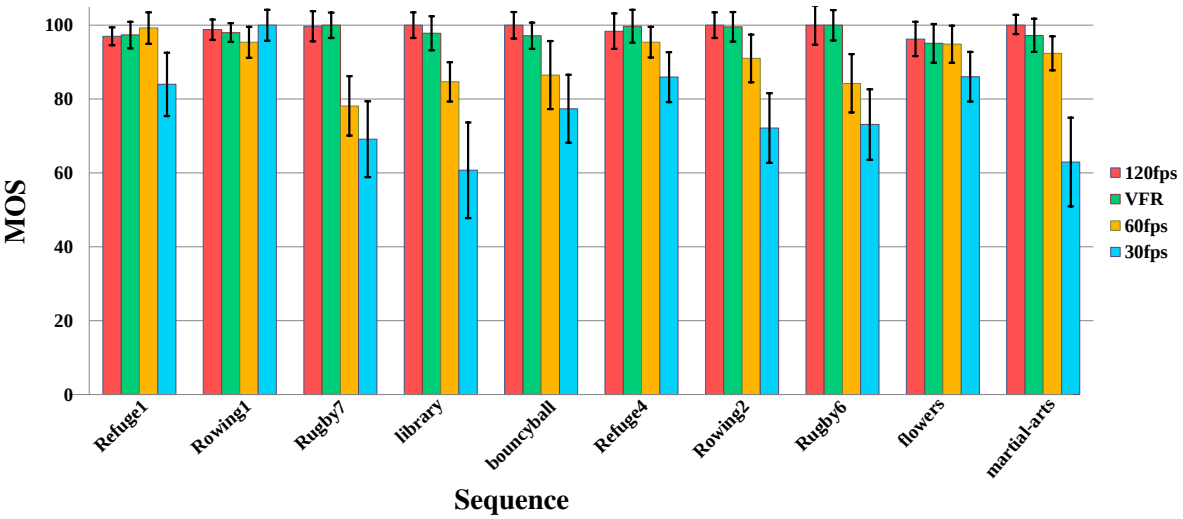


Fig. 7.17 Mean Opinion Score values with 95% confidence intervals for test set sequences and subjectively tested frame-rates.

7.5.3 Subjective Visual Quality Results

Fig. 7.17 shows the results of the subjective test carried out to demonstrate the interest of variable frame-rate and evaluate the perceived quality of the proposed VFR model output. For each sequence of the test set previously presented, the DMOS values, computed using Equation (7.9), of each tested frame-rate are depicted together with their associated 95% Confidence Intervals (CIs). Since none of the participants were flagged as outliers after the post-screening analysis, the presented DMOS values have been obtained using the results from the 19 participants.

$$DMOS_f(s) = 100 - \frac{1}{N} \times \sum_{n=1}^{N} S_{n,120fps}(s) - S_{n,f}(s), \tag{7.9}$$

with N the total number of valid participants, $N = 19$ in this test, $DMOS_f(i)$ the DMOS value for sequence s at the tested frame-rate f, $f \in \{120fps, VFR, 60fps, 30fps\}$. The pair $(S_{n,120fps}(s), S_{n,f}(s))$ represents the scores attributed to sequence s at respectively the hidden reference *120fps* frame-rate and tested frame-rate f, i.e. both videos of a given BTC, by the n^{th} participant, $n \in \{1,..,N\}$.

The first statement that can be made by analyzing the results of the subjective test is that, as previously stated, the benefit brought by a frame-rate of 120 images per second compared to lower frame-rates is highly content-dependent. Indeed, for the sequences *Rugby7*, *library* and *Rugby6*, there is a significant difference between the DMOS values associated to the *120fps* frame-rate and those of the *60fps* and *30fps* frame-rates. The same trend can be observed for the *bouncyball*, *Rowing2* and *martials_arts* sequences. However, for these se-

quences, the CIs of the *120fps* and *60fps* DMOS are overlapping, thus a significant difference between the perceived quality of the two frame-rates cannot be confidently guaranteed for these sequences. For other sequences, namely *Refuge1*, *Refuge4* and *flowers*, the perceived quality of the *120fps* and *60fps* seem equivalent, with similar DMOS and highly overlapping CIs. Finally, the sequence *Rowing1* shows no difference even with a frame decimation down to *30fps*.

Comparing the perceived qualities of the VFR model outputs with their source HFR 120 fps counterpart, the DMOS values of both configurations appear to be equivalent for every sequence. This trend highlights the interest of variable frame-rate with its capacity to adapt to the quantity of movement possibly varying over time. For instance, the *library* sequence opens on a camera panning with a gradually slowing speed which then stops at the middle of the sequence on a stationary top spinning at high speed. The first part of the video requires 120 fps to correctly portray the camera panning, while lower frame-rates can be used without introducing artifacts as the speed of the camera gradually drops. For this sequence, participants attributed significantly lower scores to the *60fps* and *30fps* frame-rates due to the important motion artifacts present in the first part of the video at these frame-rates. On the contrary, the VFR model correctly lowers the frame-rate when the content permits it, resulting in a score identical to the one attributed to the source HFR video. However, despite highly correlated DMOS values and overlapping CIs, there is still a chance that the perceived qualities of the compared frame-rates are actually different.

To confirm these observations and confidently state that the VFR model output perceived quality is the same as for the source HFR content, a more rigorous analysis can be performed using a two-sample unequal variance Student's t-test with a two-tailed distribution (also called Welsch's t-test). This test allows to determine if indeed the perceived qualities given by the MOS values of each pair of tested frame-rates are "significantly" different or not. In this case, the null hypothesis, H_0, would be that the tested frame-rate f_{test} has the same perceived quality as the considered reference frame-rate f_{ref}. The alternate hypothesis, H_a, would be that there is a difference between the perceived qualities of f_{test} and f_{ref}. In order to test the similarity for each possible pair of frame-rates, the possible values for both frame-rates are: $f_{test} \in \{VFR, 60fps, 30fps\}$ and $f_{ref} \in \{120fps, VFR, 60fps\}$.

First, considering the sample populations from the scores attributed to a sequence s at the two frame-rates f_{test} and f_{ref} being compared, the t-statistic $t_{f_{test},f_{ref}}(s)$ can be used, expressed as follows

$$t_{f_{test},f_{ref}}(s) = \frac{\overline{S}_{f_{test}}(s) - \overline{S}_{f_{ref}}(s)}{\sqrt{\frac{\sigma^2_{f_{test}}(s)}{N_{f_{test}}} + \frac{\sigma^2_{f_{ref}}(s)}{N_{f_{ref}}}}}, \tag{7.10}$$

with $\overline{S}_{f_i}(s)$, $\sigma^2_{f_i}(s)$, N_{f_i} the sample mean, sample variance and sample population size for

frame-rate f_i, $i \in \{test, ref\}$. In this test, $N_{f_{test}} = N_{f_{ref}} = N$, the number of observers that took part in the subjective test.

Then, by approximating the t-statistic with a Student's t-distribution, a value p, which indicates the degree of correlation between the means of the two sample populations, can be computed from the t-statistic. The higher the *p-value* is, the more significant the similarity between the distributions of the two populations is. A *p-value* lower than 0.05 indicates that there is statistical significance that the tested frame-rate f_{test} has a different perceived quality compared to the considered reference frame-rate f_{ref}. Indeed, in this case, there is a low probability of committing a type-I error, i.e. rejecting the null hypothesis when it is true, meaning that the null hypothesis can be confidently rejected. On the contrary, if the *p-value* is greater than or equal to 0.05, the null hypothesis cannot be safely rejected and both frame-rates can be considered to have the same perceived quality.

Finally, the *p-value* does not give information on the probability of committing a type-II error, i.e. a failure to reject the null hypothesis when the alternate hypothesis is true, which is thus still a possibility. To ensure a low type-II error probability, and thus a statistically powerful test, the power β of the statistical test [193] must be lower than 0.2. The power β has been computed for each possible pair of tested and reference frame-rates, resulting in an average β value of 0.044, showing that there is a lower than 5% chance, on average, to commit a type-II error. Therefore, the similarity assessment for each pair of possible frame-rates can be only based on the *p-values* resulting from the Student's t-test.

Fig. 7.18 depicts the *p-values* computed for each sequence and each possible frame-rate combination. Green-colored cells show the frame-rate pairs for which the associated *p-value* is greater than 0.05. Since every *VFR* vs *120fps* comparison falls within this category, it can be confidently concluded than the perceived quality of the VFR model output video is always the same as the original 120 fps frame-rate. This confirms that the under-estimated frame-rate predictions, identified in the confusion matrix depicted in **Fig. 7.14**, do not impact the perceived quality of the VFR videos. This also tends to validate the hypothesis made while analyzing training errors, stating that the ground truth is imperfect due to the coarse-grained nature of the ground truth annotations. Indeed, with its fine-grained decisions, the VFR model is capable of capturing smaller variations of critical frame-rates, thus resulting in predictions different from the ground truth, which are identified as prediction errors.

7.5.4 Compression Efficiency and Complexity Reduction

In order to evaluate the impact of VFR on coding performance, both the source HFR videos and VFR model outputs have been encoded using the HEVC reference software encoder HM16.12 [41]. The encoder was configured to use the HEVC CTC in RA configuration with a GOP size of 16 pictures and an intra-period of approximately 1 second to match with the considered broadcasting use-case. The quantization parameter was set to $QP =$

FPS	120	VFR	60
VFR	**0.77**		
60	**0.26**	**0.41**	
30	**0.02**	**0.03**	**0.00**

(a) Refuge1

FPS	120	VFR	60
VFR	**0.96**		
60	**0.14**	**0.14**	
30	**0.42**	**0.44**	**0.12**

(b) Rowing1

FPS	120	VFR	60
VFR	**0.56**		
60	**0.00**	**0.00**	
30	**0.00**	**0.00**	**0.01**

(c) Rugby7

FPS	120	VFR	60
VFR	**0.56**		
60	**0.00**	**0.00**	
30	**0.00**	**0.00**	**0.00**

(d) library

FPS	120	VFR	60
VFR	**0.41**		
60	**0.01**	**0.00**	
30	**0.00**	**0.00**	**0.01**

(e) bouncyball

FPS	120	VFR	60
VFR	**0.85**		
60	**0.27**	**0.27**	
30	**0.02**	**0.01**	**0.03**

(f) Refuge4

FPS	120	VFR	60
VFR	**0.45**		
60	**0.02**	**0.01**	
30	**0.00**	**0.00**	**0.00**

(g) Rowing2

FPS	120	VFR	60
VFR	**0.26**		
60	**0.00**	**0.00**	
30	**0.00**	**0.00**	**0.00**

(h) Rugby6

FPS	120	VFR	60
VFR	**0.99**		
60	**0.56**	**0.58**	
30	**0.00**	**0.00**	**0.01**

(i) flowers

FPS	120	VFR	60
VFR	**0.13**		
60	**0.00**	**0.03**	
30	**0.00**	**0.00**	**0.00**

(j) martial_arts

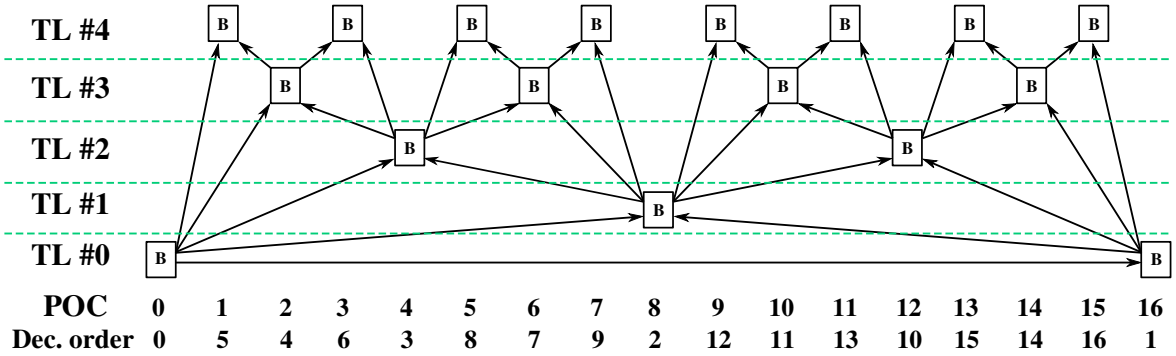
Fig. 7.18 *p-value* probabilities resulting from two-sample unequal variance bilateral Student's t-test on MOS values for each pair of tested frame-rates and each test set sequence. $p \geq 0.05$ *(green) means there is no significant difference between the MOS value of the row and column frame-rate labels while* $p < 0.05$ *(red) indicates that the MOS value of the row frame-rate label is significantly lower than the MOS value of the column frame-rate label.*

$\{22, 27, 32, 37\}$ to cover a wide range of bitrates and applications.

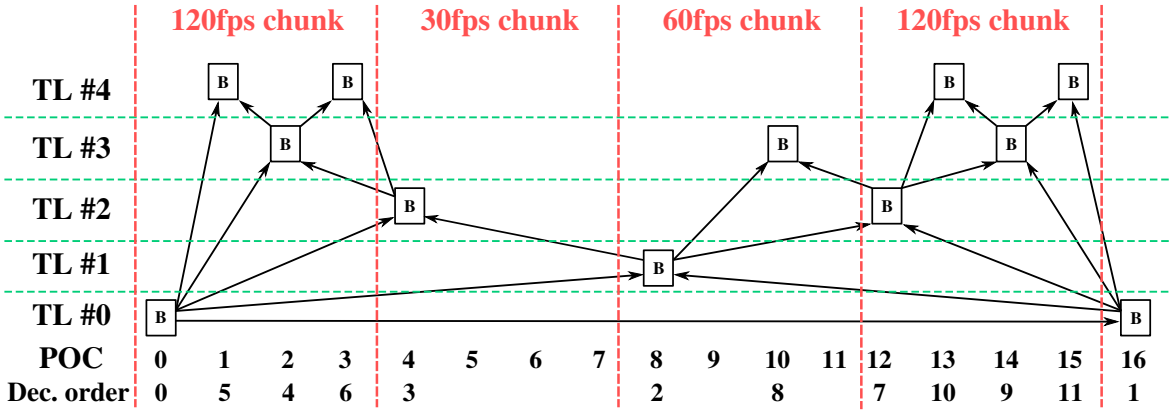
For the VFR encodings, the reference software encoder has been modified to handle the list of 4-frame chunk-level frame-rates decisions as input. **Fig. 7.19** depicts the GOP structures for both the *120fps* and VFR encodings. As can be seen, only the frames kept by the VFR model are processed by the encoder, allowing for a lower resulting bitrate and a reduced coding complexity. Thanks to the built-in support of temporal scalability, removing the frames from upper TLs does not break the decoding dependencies, making the VFR bitstream decodable by the reference software decoder without modifications.

The bitrate savings presented in this performance evaluation assume that, with the same QP, the perceived quality of the VFR decoded video is the same as the decoded original *120fps* video, as was demonstrated for the VFR and *120fps* uncompressed inputs. Indeed, the frames decoded from the VFR bitstream are exactly the same as their corresponding picture in the *120fps* decoded video due to the use of identical GOP structures. This statement has been verified by video coding experts during a viewing session comparing the decoded videos of both systems, showing that the VFR decisions made on the uncompressed content are still valid, and thus do not introduce additional artifacts after encoding and decoding.

Table 7.3 summarizes the performance results for the VFR encodings compared to regular *120fps* HEVC encodings, in terms of both bitrate savings and encoding complexity



(a) 120 fps



(b) VFR

Fig. 7.19 Example of GOP structures of size 16 for a) source HFR 120 fps content and b) VFR with different frame-rates for each 4-frame chunk.

reduction for the 15 sequences of the objective evaluation test set. The proportion of frames dropped by the VFR coding scheme are also added for information. Results for both the VFR model and ground truth decisions are presented to compare the performance at two different levels of granularity.

With the VFR model decisions, the VFR coding scheme offers 4.3% bitrate savings on average, ranging from 0% to 15.4% for sequences where 120 and 30 frames per second are chosen for the whole sequence, respectively. For sequences with mostly *60fps* chosen or with temporally varying decisions, the bitrate savings are generally around 4%. These bitrate savings are not equal to the proportion of frames dropped by the VFR model due to the significantly lower amount of bits used to encode the frames of the upper TLs. Indeed, upper TL frames are coded using higher quantization steps and greatly benefit from the inter-picture predictions of the RA coding configuration. Thus, the amount of transmitted quantized residuals is lower for these frames, especially if the motion is easily predictable and if high spatial details are not present in the source content. For the complexity reduction brought by the VFR coding scheme, the results are close to the amount of frames dropped,

Table 7.3 VFR HEVC encoding performance compared to 120fps HEVC encodings for VFR predicted labels (Model) and ground truth (G-T) labels on the test set.

Sequence	Bitrate savings		Encoding $TR_{\%}$		Frames Dropped		Decoding $TR_{\%}$	
	Model	G-T	Model	G-T	Model	G-T	Model	G-T
Refuge1	-1.1 %	-4.9 %	7.8 %	39 %	10 %	50 %	8.4 %	42 %
Rowing1	-9.3 %	-9.3 %	60 %	60 %	75 %	75 %	58 %	58%
Rugby7	-0.1 %	0.0 %	0.6 %	0.0 %	0.9 %	0.0 %	0.6 %	0.0 %
library	-5.0 %	-4.8 %	39 %	37 %	42 %	41 %	35 %	33 %
bouncyball	-2.5 %	0.0 %	6.7 %	0.0 %	8.7 %	0.0 %	7.4 %	0.0 %
Refuge4	-3.2 %	-3.5 %	47 %	49 %	52 %	55 %	46 %	48 %
Rowing2	-0.6 %	-0.4 %	6.1 %	5.4 %	9.7 %	8.7 %	6.4 %	5.6 %
Rugby6	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
flowers	-4.1 %	-4.1 %	41 %	40 %	50 %	50 %	37 %	37 %
martial_arts	-4.0 %	-0.5 %	23 %	5.9 %	28 %	7.6 %	24 %	5.7 %
Katana	-5.9 %	-1.2 %	35 %	11 %	38 %	13 %	34 %	11 %
NYCBike	-6.2 %	-5.6 %	27 %	25 %	32 %	29 %	25 %	23 %
pour	-1.6 %	-1.6 %	11 %	11 %	16 %	15 %	11 %	11 %
Refuge2	-15 %	-14.6 %	70 %	68 %	74 %	71 %	66 %	63 %
Refuge3	-5.8 %	-5.6 %	53 %	52 %	58 %	56 %	52 %	50 %
Average	-4.3 %	-3.7 %	28 %	27 %	33 %	32 %	27 %	26 %

with an average encoding complexity reduction of 28%, ranging from 0% to 70%. The per-sequence results follows the same trend as for bitrate savings but with higher gain variations. The difference between the complexity reduction and frames dropped results mainly comes from the slightly reduced coding complexity of the upper TL frames compared to the kept frames of the lower TLs. Indeed, a higher number of residual coefficient to binarize and process with the entropy coding engine increases the encoding time. For the decoding complexity, the observed gains are highly similar to those observed for the encoding side, with an average decoding complexity reduction of 27.4% for the VFR coding scheme.

With the ground truth annotated frame-rates, the bitrate savings and complexity reduction results are very close to the performance with the predicted frame-rates. This can be explained by the high correct prediction rate of the VFR model on the test set. The results are only significantly different for some sequences. *Refuge1* shows lower gains for the VFR model output due to the over-estimation of the required frame-rate, i.e. alternation between *120fps* and *60fps* prediction while the annotated ground truth frame-rate is *60fps* for the major part of the sequence. The opposite situation can be observed for the sequences *bouncyball*, *martial_arts* and *Katana*, where the VFR model allows for lower frame-rates more frequently than the ground truth, thus resulting in higher gains.

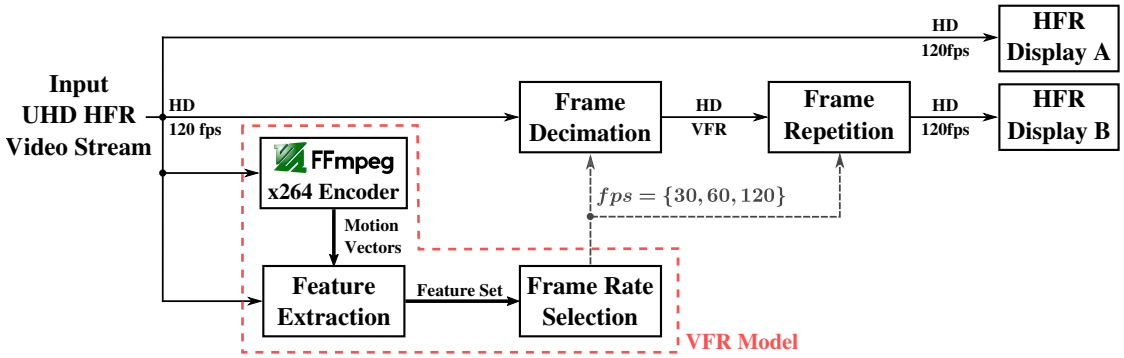


Fig. 7.20 Block diagram of the real-time UHD VFR demonstration implementation

7.6 Real-time VFR Demonstration

The proposed VFR RF-based model has been implemented and integrated into the demonstration pipeline depicted in **Fig. 7.20**. This pipeline aims at demonstrating that the VFR model can process HD HFR video streams in real-time while delivering an output quality visually equivalent to the source HFR content but with less processed frames.

The input HD HFR video stream, in raw YUV 4:2:0 8-bit format, is fed to three different processes: the *VFR Model*, a *Frame Decimation* process and a first HFR-capable screen, called *HFR Display A*, directly displaying the source HFR content.

The *VFR Model* implementation is composed of several separate processes. The HFR stream is first fed the *Feature Extraction* responsible for generating the feature set vector of each input frame. An external *Ffmpeg x264 Encoder* [194] is used in parallel in ultrafast configuration to generate the MVs required by the *Feature Extraction* process. This x264 encoding is only used as a motion estimation algorithm as only the MVs are subsequently used, during the temporal feature extraction process, by the VFR model. It has been chosen for its very low computational demand and fast implementation while delivering a good enough motion field. It is important to note that the VFR prediction results presented in this chapter have been obtained using the proposed RF classifiers trained with the MVs resulting from the same x264 encoding process. Once the feature set of each frame of the current 4-frame chunk are available, the *Frame-Rate Selection* process, i.e. the proposed cascaded binary RF classifiers, are used to predict the critical frame-rate.

The *Frame Decimation* process generates the VFR video from the input HFR stream and the frame-rate instruction given by the *Frame-Rate Selection* process. Each 4-frame chunk is processed as soon as its associated frame-rate decision is available. A *Frame Repetition* process is added to format the VFR stream into a regular HFR stream by copying the previous kept frame each time a frame is dropped by the VFR model. Thus, the video displayed on the second HFR-capable screen, *HFR Display B*, is kept at a constant frame-rate. This aims to mirror the behavior of conventional decoders that would copy the last decoded frame if

the one expected at a given time-stamp, derived from the given 120 fps frame-rate, is not available. Both the software video players used to feed the HFR displays are synchronized to demonstrate the identical perceived quality between the legacy HFR content and its VFR counterpart.

The display used for the demo are the same as for the expert database annotation viewing sessions, i.e. two identical Asus RoG Swift PG278Q (27" TN panel, QHD resolution, 144Hz max refresh rate, 350 cd/m^2 peak luminance, 1000:1 contrast ratio). Both displays are interfaced with a DisplayPort cable to the graphic card of the computer running all the pipeline processes in parallel and in real-time. The computer configuration used for this real-time demonstration is as follows: Intel i7-4930K, 3.4 GHz, 16GB RAM and Nvidia GeForce GTX 1060 6GB GPU card.

7.7 Conclusion

In this chapter, a new variable frame-rate coding scheme is proposed for broadcast delivery of HFR (120 fps) contents. The proposed scheme incorporates a machine learning based VFR model capable of dynamically adapting the frame-rate of the video before encoding and transmitting it to the end receiver.

The VFR model relies on several spatio-temporal features extracted from each frame of the input video to predict the optimal lowest artifact-free frame-rate through two cascaded binary RF trained classifiers. The considered frame-rate adaptation is performed dynamically by choosing for each chunk of 4 consecutive input frames its associated critical frame-rate, among the three possible values: *30fps*, *60fps* or *120fps*. The model achieves an average critical frame-rate correct prediction rate of 84%, while keeping the frame-rate under-estimations error rate below 8%. The visual quality of the generated VFR videos has been carefully evaluated through formal subjective tests showing an identical perceived quality compared to the source HFR content.

From a coding performance perspective, the proposed VFR coding scheme provides average bitrate savings of 4.3% in addition to average complexity reductions of 28% and 27.4% at the encoding and decoding sides, respectively.

A real-time demonstration of the uncompressed VFR video generation, shown at both the International Broadcasting Convention (IBC) 2019 and National Association of Broadcasters (NAB) Show 2019, is also proposed. It includes a real-time software implementation of the VFR prediction model running in parallel with two software raw video players each feeding an HFR screen. Both the input legacy HFR and processed VFR videos are displayed synchronously to demonstrate the equivalence in perceived quality.

The proposed solution is a practical candidate to lower the requirements for the broadcast delivery of the upcoming HFR services of the DVB UHD second deployment phase.

Chapter 8

Locally Adaptive Spatio-Temporal Resolution

As demonstrated in Chapters 5 and 6, the spatial resolution can be locally changed based on the video content to achieve an efficient scalable encoding of UHD video clips, both in terms of coding efficiency and processing time. On another hand, it has been shown in Chapter 7 that, depending on the content of the video, the frame-rate can also be dynamically lowered to achieve an efficient coding of HFR contents without impacting the perceived quality.

Both these approaches have been developed independently, each showing a significant reduction in encoding complexity and an interesting coding efficiency. This chapter aims at combining the ASR-based scalable encoder and the VFR model to propose a low-complexity scalable coding scheme based on Locally Adaptive Spatio-Temporal Resolution (LASTR).

This chapter is organized as follows. The proposed scalable coding scheme combining the different algorithms proposed in this thesis is first described in Section 8.1. Then, the preliminary performance evaluation of the proposed solution compared to SHVC is presented in Section 8.2. Finally, Section 8.3 concludes this chapter.

8.1 Description of the Technique

Since the VFR model has been designed as an entirely pre-processing tool, the combination with the ASR-based dual-layer scalable encoder is straightforward. Indeed, the proposed LASTR scalable coding simply uses these two methods sequentially, with the VFR analysis performed before the ASR-based scalable encoding.

The detailed block diagram of the proposed solution is depicted in **Fig. 8.1**. As can be observed, the frame-rate selection process of the VFR model is performed on the spatially downsampled input video. By doing this, the input to the VFR model has the same HD HFR format as the content used to train the model. As soon as the frame-rate of a 4-frame chunk

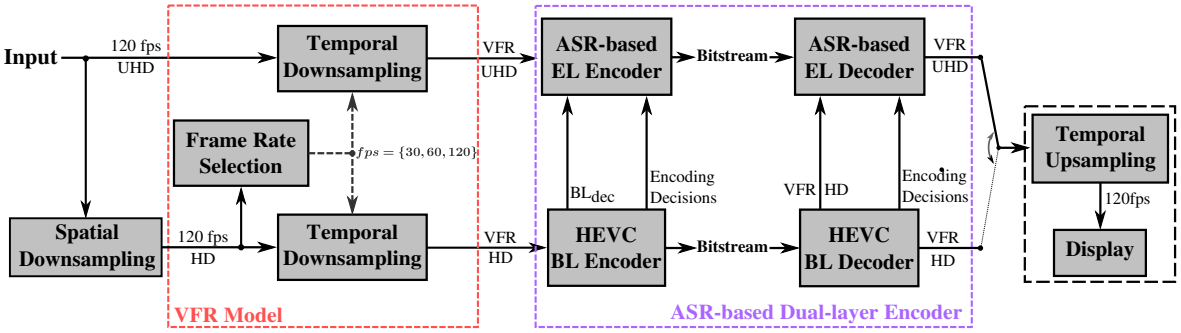


Fig. 8.1 Block diagram of the LASTR-based scalable encoder combining the VFR Model and the ASR-based EL encoder.

is available, both the input UHD HFR and its spatially downsampled version are temporally downsampled to the chosen frame-rate to respectively serve as VFR inputs to the ASR-based EL encoder and HEVC BL encoder.

Due to the RA configuration used, and its inherent out-of-order coding of the video frames, there is a small latency introduced before the scalable encoding process can be performed. Indeed, all images of a GOP must be available to begin the encoding process, thus inducing a 16-frame (4 chunks) latency. Since the ASR-based encoder follows the same GOP structure and decoding dependencies as HEVC, the VFR encoding of the EL is enabled in the same way as for HEVC, as described in Section 7.5.4. After decoding, the decoded video of the selected layer is temporally upsampled to 120 fps before being sent to the display device.

8.2 Performance Evaluation

Due to the highly limited amount of video content available in UHD HFR format, the performance of the proposed solution has only been evaluated on a small video dataset. Three videos of the test set presented in Section 7.5.1, namely *Rowing1*, *Rowing2* and *Refuge4* have been used in their UHD version. The spatio-temporal characteristics of these sequences, for their HD version, are shown in **Fig. 7.13**. Although these values slightly vary for the UHD resolution, the distribution in the SI-TI space remains similar. One of the sequences used for training, *YachtRide* has also been used to complete the performance evaluation dataset. This sequence contains an horizontal camera panning following a moving boat. It thus contains fairly high motion with a moderate amount of spatial details.

Even with this limited number of test sequences, a good insight on the performance of the LASTR-based scalable coding scheme should be obtained since the frame-rate labels chosen by the VFR model for these sequences reflect quite well the critical frame-rate distribution expected for broadcast and streaming use-cases. Indeed, *Rowing2* and *YachtRide* have the *120fps* label attributed to the vast majority of their chunks, while *Refuge4* and *Rowing1* have

Table 8.1 Performance comparison between the ASR-based, LASTR-based and SHVC encoders for 2x spatial scalability in RA configuration.

(a) BD-Rate for equal PSNR-YUV

Sequence	LASTR vs SHVC	LASTR vs ASR	ASR vs SHVC
YachtRide	-5.58 %	0.0 %	-5.58 %
Rowing1	3.88 %	-20.8 %	31.4 %
Rowing2	0.35 %	-1.29 %	1.7 %
Refuge4	-16.2 %	-3.09 %	-13.5 %
Average	**-3.52 %**	**-7.17 %**	**3.49 %**

(b) Encoding Time Reduction - $TR_{\%}$ - (EL+BL)

Sequence	LASTR vs SHVC	LASTR vs ASR	ASR vs SHVC
YachtRide	74.7 %	0.0 %	74.7 %
Rowing1	91.3 %	65.1 %	75.2 %
Rowing2	76.6 %	7.16 %	74.8 %
Refuge4	87.8 %	45.3 %	76.9 %
Average	**82.5 %**	**29.4 %**	**75.4 %**

majority frame-rate labels of *60fps* and *30fps*, respectively.

Table 8.1a summarizes the coding performance results for the proposed LASTR-based scalable encoder and SHVC. The comparison with the ASR-based encoder is also shown to assess the impact of the VFR solution in a scalable context. As can be observed, the LASTR-based solution achieves 3.5% average bitrate savings compared to SHVC, while the ASR-based solution shows a bitrate overhead of 3.5% compared to SHVC for this dataset.

As expected form the analysis of the results the ASR and VFR schemes respectively presented in Sections 6.2.3 and 7.5.4, the coding efficiency is content-dependent. Indeed, a high coding gain is obtained for the *Refuge4* test sequence, which mainly comes from the ASR scheme due to the easily recovered spatial details through the upsampling of the BL. For the *YachtRide* sequence, since a *120fps* frame-rate is chosen for the whole sequence by the VFR model, the observed bitrate savings only come from the better coding efficiency of the ASR-based encoder for such sequences with a moderate amount of spatial details which can be well recovered by the internal upsampling process.

For the remaining two sequences, the LASTR scheme shows a bitrate overhead compared to SHVC. This is especially the case for the *Rowing1* sequence which contains a considerable amount of very high spatial details in the water, thus leading to a poor performance of the ASR-based scalable encoder - 31.4% losses compared to SHVC. However, this bitrate overhead is almost entirely compensated by the coding gain achieved by the VFR model for this sequence due to the choice of a *30fps* frame-rate for the whole sequence. It is important to note that the *Rowing1* test sequence is an unusually critical sequence for video coding with its very high spatial details that temporally vary, in an unpredictable manner, which can explain the poor performance of the ASR-IP-MD. Indeed, for another sequence with very high spatio-temporal details, such as *Rowing2*, the ASR-IP-MD algorithm only shows reasonable losses compared to SHVC.

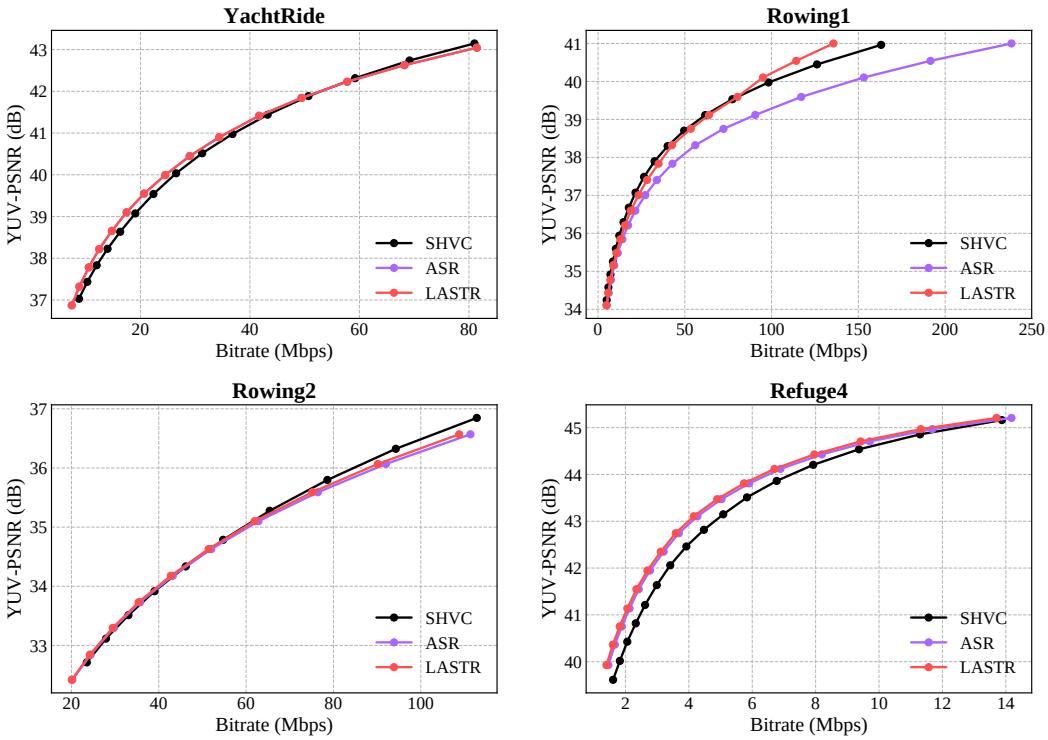


Fig. 8.2 Rate-Distortion curves for the proposed LASTR-based scalable encoder compared to the ASR-based encoder and SHVC (BL and EL in RA configuration).

Overall, the VFR model results in average bitrate savings of 7% when added before the ASR-based scalable encoder, with the per-sequence performance directly depending on the proportion of frames removed when lowering the frame-rate, which is consistent with the results obtained in Section 7.5.4. The rate-distortion curves depicted in **Fig. 8.2** confirm this analysis.

Concerning the complexity reduction brought by the proposed solution, the obtained encoding time reduction measures are summarized in Table 8.1b. As can be seen, the average encoding time reduction achieved by the LASTR-based scalable coding scheme, for the complete encoder (BL + EL), compared to SHVC reaches 82.5%. This gain comes from both the ASR scheme, with its 75% complexity reduction constant over all sequences, and the VFR model, with its average 29% encoding time reduction. The per-sequence gains mainly depend on the VFR model frame-rate decisions, with an encoding time reduction as high as 91.3% for the *Rowing1* coded with a frame-rate lowered to *30fps*.

8.3 Conclusion

In this chapter, the LASTR algorithm, a complete low-complexity scalable coding scheme based on the local adaptation of both the spatial and temporal resolutions has been proposed.

To this end, the ASR-based scalable coding scheme, presented in Chapters 5 and 6, is coupled to the VFR model, described in Chapter 7, performed prior to encoding.

The proposed LASTR solution leverages the advantages of adapting the spatial resolution for the scalable coding of UHD contents and the benefits of dynamically lowering the frame-rate for an efficient coding of HFR videos, with the goal of achieving further improvements in terms of coding efficiency and complexity reduction compared to the previously proposed algorithms.

The preliminary evaluation shows that the proposed LASTR coding scheme achieves 3.5% bitrate savings compared to SHVC, thus outperforming the ASR-based scalable scheme by 7% in terms of resulting bitrate for an equal PSNR-YUV value. In addition, the complexity is further reduced compared to the ASR-based encoder, with an average encoding time reduction of 82.5% over SHVC, with both the BL and EL encoders taken into account.

The proposed solution thus appear to significantly outperform the state-of-the-art scalable SHVC encoder in terms of both coding efficiency and encoding complexity. The LASTR-based scalable coding scheme thus appears to be an efficient solution for the deployment of UHDTV-compliant services necessitating the coding of UHD HFR video contents.

However, due to the limited amount of UHD HFR available to evaluate the performance of the LASTR scalable coding scheme, all kinds of video contents have not been tested in the preliminary performance evaluation carried out in this chapter. To confirm the observed performance gains, the LASTR scalable coding scheme should thus be further evaluated on a larger dataset to cover a more fine-grained scale of spatio-temporal characteristics. This will be made possible in the upcoming months with the arrival of UHD HFR display devices in the consumer market, which will surely increase the availability of such UHD HFR contents.

Part VI

Conclusion

Chapter 9

Conclusion

Due to the numerous improvements of the video signal characteristics defined in the UHDTV standard, the deployment of UHDTV compliant services remains a challenging task for the broadcasting industry and OTT service providers. Indeed, the introduction of technologies such as 4K and beyond, HFR, HDR and WCG in the encoding pipeline considerably enlarges the amount of data to process. The already very high computational demand of state-of-the-art video codecs is thus further increased, especially for scalable use-cases.

The objective of this thesis was thus to investigate lightweight scalable encoding approaches based on the adaptation of the spatial and temporal resolutions, either through pre-processing algorithms or coding tools implemented inside end-to-end codecs. The final goal was to propose a low-complexity scalable method capable of efficiently encoding 4K HFR contents while providing a backward compatibility for existing HDTV compliant infrastructures. Section 9.1 summarizes the main contributions of this thesis and Section 9.2 addresses future works through straightforward and more long-term potential improvements and extensions of the proposed solutions.

9.1 Achieved Work

The contributions of this thesis are organized into three parts. The first one focused on the design of pre- and post-processing tools to achieve spatial scalability with little complexity overhead. To this end, two different signal decompositions performed prior to encoding have been proposed.

First, a low-complexity scalable coding scheme, based on an improved polyphase-based decomposition followed by a regular HEVC encoding process, has been proposed. The state-of-the-art polyphase decomposition allows for a decomposition of each video frame into four sub-resolution images, which can then be used to achieve spatial scalability with the built-in decoding properties of the HEVC GOP structure. However, the polyphase decomposition contains inherent flaws, such as the introduction of a phase shift difference between the luma

and chroma pixel of each sub-resolution image. A filtering process has thus been proposed to compensate this shift and thus obtain better inter-picture prediction of the chroma pixels. The proposed solution offered average bitrate savings of 3.6% compared to the state-of-the-art polyphase decomposition, leading to a lowered bitrate overhead of 6.6% compared to SHVC while reducing the coding complexity by 55%. This work has been presented during the 2018 IEEE Data Compression Conference (DCC) [195].

Then, a wavelet-based decomposition has been designed to replace the polyphase-based decomposition in the proposed low-complexity scalable coding scheme, aiming at avoiding the aliasing introduced by the polyphase decomposition. A modification to the conventional wavelet transform was proposed to enable the use of both the wavelet transformed signal and the HEVC GOP properties to achieve spatial scalability. Several integer-to-integer wavelet transforms have been considered, with a proposed modification of the R-D cost computation based on the wavelet sub-band type for bi-orthogonal wavelets. Experimental results showed that the simple orthogonal Haar wavelet achieved the best results, with an average bitrate overhead of 1.9% compared to SHVC with a complexity reduction of 54%, thus outperforming the previously proposed improved polyphase decomposition. This work has been presented in the Picture Coding Symposium (PCS) in 2018 [196].

The second part of this thesis addressed the design of a more conventional dual-layer scalable architecture using an HEVC encoder in the BL for backward compatibility and a proposed low-complexity encoder based on several novel algorithms for the EL.

First, the ASR-IL algorithm, based on the local adaptation of the spatial resolution for inter-layer predictions has been designed. It consists in enabling a R-D optimized and content-based selection of the spatial resolution for each block of the EL, which is then used for the inter-layer prediction stage and remaining core encoding processes. Experimental results showed that the proposed dual-layer architecture achieved a complexity reduction of 72% compared to SHVC for a bitrate overhead of 4.8%.

Then, two improvements to the ASR-IL scalable coding scheme are proposed. The first solution, called ASR-IL-MD, has been proposed to reduce the signaling cost of the ASR-IL algorithm using a prediction scheme for the block-level optimal resolution. This resolution mode prediction algorithm has been designed to reuse the MVs of the BL when motion compensation was used by the BL encoder to predict the co-located BL block. After an appropriate scaling of the MVs, the resolution of the EL blocks can be predicted from previously encoded EL frames through a proposed resolution derivation process. Experimental results showed that the proposed solution enabled an average bitrate overhead of 0.5% compared to SHVC, with the same prediction constraints, for complexity reductions of 96% and 47% for the EL encoder and dual-layer encoder, respectively. This work has been presented at the 2019 International Conference on Image Processing (ICIP) [197] and protected by patent [198]. The second improvement, called ASR-IP-MD, has been proposed to enable inter-picture predictions in the EL encoder. Coupled with the resolution mode derivation

scheme, this solution allows for an efficient support of the commonly used, in broadcast and streaming use-cases, RA configuration. Experimental results showed that the complexity reductions are further improved for both the EL encoder and overall dual-layer scheme with respectively 97.8% and 78.6% average gains. In addition, the proposed ASR-IP-MD solution achieved average bitrate savings of 1.3% compared to SHVC, despite some losses observed in the chroma channels of several test sequences.

The last part of this dissertation investigated the spatio-temporal resolution adaptation to propose a complete low-complexity scalable coding scheme.

First, a variable frame rate solution has been designed to locally detect the critical frame-rate, i.e. the lowest frame-rate that does not introduce motion artifacts. Based on a ML algorithm trained on a proposed large dataset of annotated HFR videos, the VFR model is capable of dynamically changing the frame-rate, among 3 possible values, for each chunk of 4 frames. Subjective tests demonstrated that the VFR model output was visually identical to the source HFR contents and experimental results showed that average bitrate savings of 4.3% and complexity reduction of 28% could be achieved using VFR videos instead of their original HFR counterparts. This work, protected by patent [199], is currently under review for a publication in the IEEE Transactions on Broadcasting journal. The real-time demonstration of this work has also been presented in two international industrial conventions, the NAB Show 2019 and IBC 2019.

Then, the LASTR solution has been proposed to locally adapt both the spatial and temporal resolutions in order to achieve a lightweight scalable encoding of UHD HFR contents. This solution is based on the combination of the critical frame-rate detection offered by the VFR model with the ASR-IP-MD scalable coding scheme. Preliminary experimental results showed that average bitrate savings of 3.5% can be achieved while reducing the overall dual-layer scalable scheme encoding time by 82.5% compared to the state-of-the-art SHVC encoder, thus outperforming by a significant margin state-of-the-art SHVC complexity reduction algorithms.

To summarize, three main contributions were proposed in this thesis. The first one is based on pre-processing tools to achieve low-complexity scalable coding with a single HEVC encoder instance. The second contribution investigated a dual-layer scalable coding scheme using a low-complexity encoder based on adaptive spatial resolution for the enhancement layer while ensuring backward compatibility through the HEVC base layer. Finally, the third contribution of this thesis focused on the spatio-temporal resolution adaptation with the design of a variable frame-rate algorithm performed prior to encoding combined with the previously proposed scalable encoder. Through these contributions, the deployment of scalable coding services compliant to the upcoming UHDTV standard could be considerably eased, thanks to low computational demand of the proposed solutions. Furthermore, this work led to several publications and patents, listed in Appendix A.

9.2 Prospects and Future Works

The contributions presented in this dissertation opens opportunities for future works on low-complexity scalable coding. This section presents the proposed research directions and prospects.

9.2.1 Exploitation

Open-source implementations of the HEVC/SHVC standards, such as the Kvazaar encoder [200] and the openHEVC decoder [201], could be used to implement the proposed solutions in real-time encoders/decoders. This would allow for an increased coding performance and potentially enable the real-time processing of additional new emerging formats compliant with the UHDTV standard. In addition, real-time implementations of the proposed solutions would ease their deployment by industrial broadcasters and OTT service providers.

9.2.2 Performance Improvement

Several contributions presented in this document could achieve a better performance by improving certain processes of the proposed scalable schemes.

First, the ASR-IP-MD dual-layer scalable encoder presented in Chapters 5 and 6 rely on a straightforward EL block partitioning process derived from the BL decisions to greatly limit the computational demand of the proposed EL encoder. By enabling a limited search space for optimal partitioning around the block size derived from the BL CU depth, the coding efficiency could be significantly increased thanks to a better adaptation to the EL content. To keep a low-complexity constraint, this improved EL partitioning could be inferred using ML approaches as in [202]. Additionally, other HEVC features that have not been integrated in the proposed encoder, such as a deblocking filter or other loop filters. This would potentially significantly increase the output perceived quality for a reasonable complexity overhead. Besides, the performance of the proposed should be further evaluated with perceptual video quality metrics, such as VMAF or SSIM, and more importantly through formal subjective tests.

Then, the complexity reduction of the proposed ASR-IP-MD encoder could also be increased by replacing the R-D optimized selection of the spatial resolution by a ML-based prediction of the R-D optimal resolution. Indeed, by using the BL and EL input pixels as well as the decisions from previously encoded neighboring EL blocks, a RF regressor could be trained to predict the R-D cost of each possible spatial resolution, thus avoiding the costly RDO stage. Furthermore, if the regression model accuracy can achieve good accuracy without the EL block pixel as input, the process could be reproduced at the decoder side, which would remove the need to signal the chosen resolution, thus further improving the coding

efficiency.

In addition, if such a resolution prediction is implemented, the kernel-based upsampling process used in the ASR could be replaced by learned SR algorithms, for example using shallow CNNs. This would considerably lower the inter-layer prediction residuals as well as significantly improved the reconstruction quality of the EL output, especially for blocks coded in one of the lower resolutions. However, this solution would most likely only be viable for devices equipped with specific hardware, such as GPUs.

On another hand, for the VFR model presented in Chapter 7, a possible improvement of the algorithm would be to train it with the labels annotated based on the compressed versions of the HFR dataset, at several compression ratios. By doing this, the frame-rate could potentially be lowered more often due to the loss of spatial details induced by the encoding process, thus lowering the possible flickering artifacts at lower frame-rates. This would require a considerable time investment on dataset annotation but could lead to significant compression and complexity reduction gains.

Finally, it would be highly beneficial to confirm the observed performance of the LASTR algorithm presented in Chapter 8 by further evaluating the performance of this method through both objective and subjective tests using a larger set of test sequences. This will soon be made possible thanks to the upcoming arrival of UHD HFR displays in the consumer market, which will surely increase the availability of UHD HFR contents. Besides, all the potential improvements to the ASR-IP-MD and VFR solutions discussed in this section would also impact the LASTR scalable encoder presented in Chapter 8.

9.2.3 Extension of Proposed Solutions

In addition to the performance improvements prospects presented in the previous section, the contributions proposed in this document could be extended, either by straightforward modifications to handle more use-cases or more long-term changes that modify the core design of the solutions.

Since the beginning of this thesis, the JVET committee has been developing the next generation video codec to be standardized by the ISO/IEC and ITU consortium as VVC, the successor of HEVC. With its 40% increased coding efficiency, it could be really interesting to adapt the proposed dual-layer encoder with a VVC encoder as BL, and a low-complexity EL encoder modified to handle the new partitioning structures of VVC.

With a more long-term perspective, and without the low-complexity constraint, the proposed solutions could be extended to use state-of-the-art artificial intelligence technologies. For example, on one hand, Generative Adversarial Networks (GANs) [203] could be used as a replacement of the EL to generate the supplemental enhancement information of the scalable coding scheme, potentially achieving substantial performance improvements considering the recent advances in the domain. On the other hand, the VFR model could be modified

to handle additional cases in which the frame-rate of the encoded video would be lowered but a CNN-based frame interpolation process would be introduced as a post-processing. This would enable the reduction of the frame-rate in more cases, when a *30fps* decimation would be too severe but the missing frames could be recoverable by the frame-interpolation process without visible artifacts.

Part VII

References and Appendix

References

[1] ITU-R, "Recommendation BT.2020-1: Parameters Values of Ultra-High Definition Television Systems for Production and International Programme Exchange."

[2] ITU-R, "Recommendation BT.709-5: Parameter Values for the HDTV Standards for Production and International Programme Exchange."

[3] M. Nilsson, "Ultra high definition video formats and standardisation," *BT Media and Broadcast Research Paper*, 2015.

[4] M. Sugawara and K. Masaoka, "UHDTV Image Format for Better Visual Experience," *Proceedings of the IEEE*, vol. 101, no. 1, pp. 8–17, 2013.

[5] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," *White Paper*, 2019.

[6] "High efficiency video coding," *ITU-T Rec.H.265 and ISO/IEC 23008-2 (HEVC), ITU-T and ISO/IEC JTC 1*, April 2013 (and subsequent editions).

[7] G. J. Sullivan, J.-R. 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, 2012.

[8] E. Alshina, A. Alshin, Y. Cho, J. Park, W. Pu, J. Chen, X. Li, V. Seregin, and M. Karczewicz, "Inter-layer filtering for scalable extension of hevc," in *2013 Picture Coding Symposium (PCS)*, pp. 382–385, IEEE, 2013.

[9] ITU-R, "Recommendation BT.601-7: Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios."

[10] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, "A model of visual adaptation for realistic image synthesis," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 249–258, ACM, 1996.

[11] D. F. Sheet, "Introduction to the dvb project. creating global standards for digital television," 2008.

[12] K. Andersson, P. Hermansson, J. Samuelsson, J. Strom, and M. Petersson, "Report for CE1.a (Chroma QP)," *ISO/IEC JTC1/SC29/WG11 MPEG2015/m37179*, Geneva, Switzerland, October 2015.

[13] J. Strom, J. Samuelsson, and K. Andersson, "Report for CE1.b(luma adjustment)," *ISO/IEC JTC1/SC29/WG11 MPEG2015/m37272*, Geneva, Switzerland, October 2015.

[14] J. Strom, J. Sole, and Y. He, "Report of HDR CE1," *ISO/IEC JTC1/SC29/WG11 MPEG2016/m37605*, San Diego, USA, February 2016.

[15] C. Bist, R. Cozot, G. Madec, and X. Ducloux, "Tone expansion using lighting style aesthetics," *Computers & Graphics*, vol. 62, pp. 77–86, 2017.

[16] C. Bist, *Combining aesthetics and perception for display retargeting*. PhD thesis, Rennes 1, 2017.

[17] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[18] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[19] "Video codec for audiovisual services at px64 kbit/s," *ITU-T Rec. H.261*, version 1: Nov. 1990, version 2: Mar. 1993.

[20] "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s — part 2: Video," *ISO/IEC 11172-2(MPEG-1), ISO/IEC JTC 1*, 1993.

[21] "Generic coding of moving pictures and associated audio information—part 2: Video," *ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG 2 Video), ITU-T and ISO/IEC JTC 1*, Nov. 1994.

[22] "Video coding for low bit rate communication," *ITU-T Rec. H.263*, Nov.1995 (and subsequent editions).

[23] "Coding of audio-visual objects — part 2: Visual," *ISO/IEC 14496-2(MPEG-4 Visual version 1), ISO/IEC JTC 1*, Apr. 1999 (and subsequent editions).

[24] "Advanced video coding for generic audio-visual services," *ITU-T Rec.H.264 and ISO/IEC 14496-10 (AVC), ITU-T and ISO/IEC JTC 1*, May 2003 (and subsequent editions).

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] F. Xiao *et al.*, "Dct-based video quality evaluation," *Final Project for EE392J*, vol. 769, 2000.

[27] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *Netflix Tech Blog*, 2016.

[28] ITU-R, "Recommendation BT.500-13:Methodology for the Subjective Assessment of the Quality of Television Pictures."

[29] "Subjective video quality assessment methods for multimedia applications," *ITU-T Rec. P.910*, September 1999.

[30] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," *VCEG-M33*, 2001.

[31] G. Bjontegaard, "Improvements of the bd-psnr model," in *ITU-T SG16/Q6, 35th VCEG Meeting, Berlin, Germany, July, 2008*, 2008.

[32] 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, 2012.

[33] 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.

[34] A. Alshin, E. Alshina, J. H. Park, and W.-J. Han, "Dct based interpolation filter for motion compensation in hevc," in *Proc. SPIE 8499, Applications of Digital Image Processing XXXV*, vol. 8499, 2012.

[35] P. Helle, S. Oudin, B. Bross, D. Marpe, M. O. Bici, K. Ugur, J. Jung, G. Clare, and T. Wiegand, "Block merging for quadtree-based partitioning in hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1720–1731, 2012.

[36] J. Sole, R. Joshi, N. Nguyen, T. Ji, M. Karczewicz, G. Clare, F. Henry, and A. Duenas, "Transform coefficient coding in hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1765–1777, 2012.

[37] A. Norkin, G. Bjontegaard, A. Fuldseth, M. Narroschke, M. Ikeda, K. Andersson, M. Zhou, and G. Van der Auwera, "Hevc deblocking filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1746–1754, 2012.

[38] C.-M. Fu, E. Alshina, A. Alshin, Y.-W. Huang, C.-Y. Chen, C.-Y. Tsai, C.-W. Hsu, S.-M. Lei, J.-H. Park, and W.-J. Han, "Sample adaptive offset in the hevc standard," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 22, no. 12, pp. 1755–1764, 2012.

[39] R. Sjoberg, Y. Chen, A. Fujibayashi, M. 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, 2012.

[40] V. Sze and M. Budagavi, "High throughput cabac entropy coding in hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1778–1791, 2012.

[41] "HEVC reference software version 16.12."

[42] C. Rosewarne, B. Bross, K. Sharman, and G. Sullivan, "Jctvcu1002: High efficiency video coding (hevc) test model 16 (hm 16) improved encoder description," *Tech. Rep., Joint Collaborative Team on Video Coding (JCT-VC)*, 2015.

[43] J. Vanne, M. Viitanen, T. D. Hamalainen, and A. Hallapuro, "Comparative rate-distortion-complexity analysis of hevc and avc video codecs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1885–1898, 2012.

[44] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE signal processing magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[45] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, "Overview of shvc: Scalable extensions of the high efficiency video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34, 2015.

[46] 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, 2007.

[47] "SHVC reference software version 9.0."

[48] R. Parois, W. Hamidouche, J. Vieron, M. Raulet, and O. Deforges, "Efficient parallel architecture for a real-time uhd scalable hevc encoder," in *Signal Processing Conference (EUSIPCO), 2017 25th European*, pp. 1465–1469, IEEE, 2017.

[49] E. Thomas, "Polyphase subsampled signal for spatial scalability," *ISO/IEC JTC1/SC29/ WG11 JVET-B0043*, San Diego, USA, February 2016.

[50] E. Thomas, "Experiments on polyphase subsampled sequence coding," *ISO/IEC JTC1/SC29/WG11 JVET-C0032*, Geneva, Switzerland, May 2016.

[51] P. Philippe, T. Biatek, and V. Lorcy, "Cross-check of c0032 (polyphase subsampled sequence coding)," *ISO/IEC JTC1/SC29/WG11 JVET-C0078*, Geneva, Switzerland, May 2016.

[52] E. Thomas, "Subjective quality analysis of polyphase subsampled sequences," *ISO/IEC JTC1/SC29/WG11 JVET-D0034*, Chengdu, China, October 2016.

[53] V. Seregin and Y. He, "Common test conditions for shvc," *ISO/IEC JTC1/SC29/WG11 JCTVC-X1009*, Geneva, Switzerland, May 2016.

[54] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on pure and applied mathematics*, vol. 41, no. 7, pp. 909–996, 1988.

[55] G. Strang, "Wavelets and dilation equations: A brief introduction," *SIAM review*, vol. 31, no. 4, pp. 614–627, 1989.

[56] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE signal processing magazine*, vol. 8, pp. 14–38, 1991.

[57] I. Daubechies, *Ten lectures on wavelets*, vol. 61. Siam, 1992.

[58] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[59] G. Strang and T. Nguyen, *Wavelets and filter banks*. SIAM, 1996.

[60] C. S. Burrus, R. A. Gopinath, H. Guo, J. E. Odegard, and I. W. Selesnick, *Introduction to wavelets and wavelet transforms: a primer*, vol. 1. Prentice hall New Jersey, 1998.

[61] A. N. Akansu, P. A. Haddad, R. A. Haddad, and P. R. Haddad, *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic press, 2001.

[62] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 674–693, 1989.

[63] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on image processing*, vol. 1, no. 2, pp. 205–220, 1992.

[64] W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," *Applied and computational harmonic analysis*, vol. 3, no. 2, pp. 186–200, 1996.

[65] B. E. Usevitch, "A tutorial on modern lossy wavelet image compression: foundations of jpeg 2000," *IEEE signal processing magazine*, vol. 18, no. 5, pp. 22–35, 2001.

[66] A. E. Coding, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Transactions on signal processing*, vol. 41, no. 12, 1993.

[67] D. Taubman, "High performance scalable image compression with ebcot," *IEEE Transactions on image processing*, vol. 9, no. 7, pp. 1158–1170, 2000.

[68] A. Said, W. A. Pearlman, *et al.*, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on circuits and systems for video technology*, vol. 6, no. 3, pp. 243–250, 1996.

[69] Z. Xiong, K. Ramchandran, and M. T. Orchard, "Space-frequency quantization for wavelet image coding," in *Wavelet Image and Video Compression*, pp. 171–197, Springer, 2002.

[70] D. Taubman, E. Ordentlich, M. Weinberger, and G. Seroussi, "Embedded block coding in jpeg 2000," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 49–72, 2002.

[71] G. K. Wallace, "The jpeg still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[72] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal processing magazine*, vol. 18, no. 5, pp. 36–58, 2001.

[73] "Itu-t recommendation t.800| iso/iec 15444-1:2002, information technology -jpeg 2000 image encoding system -part : Core coding system."

[74] "Iso/iec 15444-3:2002, information technology -jpeg 2000 image encoding system -part3: Motion jpeg2000."

[75] S.-J. Choi and J. W. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on image processing*, vol. 8, no. 2, pp. 155–167, 1999.

[76] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 3, pp. 1793–1796, IEEE, 2001.

[77] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (limat) framework for highly scalable video compression," *IEEE transactions on image processing*, vol. 12, no. 12, pp. 1530–1542, 2003.

[78] T. André, M. Cagnazzo, M. Antonini, M. Barlaud, N. Bozinovic, and J. Konrad, "(n, 0) motion-compensated lifting-based wavelet transform," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. iii–121, IEEE, 2004.

[79] M. Cagnazzo, F. Castaldo, T. André, M. Antonini, and M. Barlaud, "Optimal motion estimation for wavelet motion compensated video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 7, pp. 907–911, 2007.

[80] T. Borer and T. Davies, "Dirac video compression using open technology," *BBC EBU technical review*, 2005.

[81] "St 2042-1:2009 - smpte standard - vc-2 video compression," *ST 2042-1:2009*, pp. 1–130, Nov 2009.

[82] D. Le Gall and A. Tabatabai, "Sub-band coding of digital images using symmetric short kernel filters and arithmetic coding techniques," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 761–764, IEEE, 1988.

[83] B. Usevitch, "Optimal bit allocation for biorthogonal wavelet coding," in *Data Compression Conference, 1996. DCC'96. Proceedings*, pp. 387–395, IEEE, 1996.

[84] M. Karczewicz, Y. Ye, and I. Chong, "Rate distortion optimized quantization," *Doc. VCEG-AH21*, Antalya, January 2008.

[85] V. Hulusic, G. Valenzise, J.-C. Gicquel, J. Fournier, and F. Dufaux, "Quality of experience in uhd-1 phase 2 television: the contribution of uhd+ hfr technology," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*, pp. 1–6, IEEE, 2017.

[86] A. M. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1132–1144, 2003.

[87] A. Segall, M. Elad, P. Milanfar, R. Webb, and C. Fogg, "Improved high-definition video by encoding at an intermediate resolution," in *Visual Communications and Image Processing 2004*, vol. 5308, pp. 1007–1018, International Society for Optics and Photonics, 2004.

[88] G. Georgis, G. Lentaris, and D. Reisis, "Reduced complexity superresolution for low-bitrate video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 2, pp. 332–345, 2015.

[89] M. Shen, P. Xue, and C. Wang, "Down-sampling based video coding with super-resolution technique," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, May 2010.

[90] J. Dong and Y. Ye, "Adaptive downsampling for high-definition video coding," in *2012 19th IEEE International Conference on Image Processing*, pp. 2925–2928, Sep. 2012.

[91] J. Dong and Y. Ye, "Adaptive downsampling for high-definition video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 3, pp. 480–488, 2013.

[92] R. Wang, C. Huang, and P. Chang, "Adaptive downsampling video coding with spatially scalable rate-distortion modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 11, pp. 1957–1968, 2014.

[93] M. Afonso, F. Zhang, A. Katsenou, D. Agrafiotis, and D. Bull, "Low complexity video coding based on spatial resolution adaptation," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017.

[94] M. Afonso, F. Zhang, and D. R. Bull, "Spatial resolution adaptation framework for video compression," in *Applications of Digital Image Processing XLI*, vol. 10752, p. 107520L, International Society for Optics and Photonics, 2018.

[95] W. Lin and L. Dong, "Adaptive downsampling to improve image compression at low bit rates," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2513–2521, 2006.

[96] Viet-Anh Nguyen, Yap-Peng Tan, and Weisi Lin, "Adaptive downsampling/upsampling for better video compression at low bit rate," in *2008 IEEE International Symposium on Circuits and Systems*, May 2008.

[97] J. Lin, D. Liu, H. Yang, H. Li, and F. Wu, "Convolutional neural network-based block up-sampling for hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.

[98] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, and H. Yang, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2316–2330, 2018.

[99] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 1153–1160, December 1981.

[100] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of applied meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[101] J. Agbinya, "Interpolation using the discrete cosine transform," *Electronics letters*, vol. 28, no. 20, pp. 1927–1928, 1992.

[102] H. Ur and D. Gross, "Improved resolution from subpixel shifted pictures," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 2, pp. 181–186, 1992.

[103] N. R. Shah and A. Zakhor, "Resolution enhancement of color video sequences," *IEEE transactions on Image Processing*, vol. 8, no. 6, pp. 879–885, 1999.

[104] N. Nguyen and P. Milanfar, "An efficient wavelet-based algorithm for image superresolution," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 2, pp. 351–354, IEEE, 2000.

[105] R. Tsai, "Multiframe image restoration and registration," *Advance Computer Visual and Image Processing*, vol. 1, pp. 317–339, 1984.

[106] S. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 1013–1027, 1990.

[107] S. P. Kim and W.-Y. Su, "Recursive high-resolution reconstruction of blurred multiframe images," in *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pp. 2977–2980, IEEE, 1991.

[108] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE transactions on image processing*, vol. 5, no. 6, pp. 996–1011, 1996.

[109] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high resolution image estimation using a sequence of undersampled images," *IEEE transactions on Image Processing*, vol. 6, no. 12, 1997.

[110] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *JOSA A*, vol. 6, no. 11, pp. 1715–1726, 1989.

[111] A. J. Patti, M. I. Sezan, and A. M. Tekalp, "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1064–1076, 1997.

[112] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[113] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *Journal of Visual Communication and Image Representation*, vol. 4, no. 4, pp. 324–335, 1993.

[114] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International journal of computer vision*, vol. 40, no. 1, pp. 25–47, 2000.

[115] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, no. 2, pp. 56–65, 2002.

[116] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, IEEE, 2004.

[117] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2008.

[118] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[119] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

[120] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, pp. 184–199, Springer, 2014.

[121] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

[122] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[123] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4778–4787, 2017.

[124] K. Turkowski, "Filters for common resampling tasks," in *Graphics gems*, pp. 147–165, Academic Press Professional, Inc., 1990.

[125] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P. Le Callet, "Comparing upscaling algorithms from hd to ultra hd by evaluating preference of experience," in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 208–213, IEEE, 2014.

[126] K. Ugur, A. Alshin, E. Alshina, F. Bossen, W. Han, J. Park, and J. Lainema, "Interpolation filter design in hevc and its coding efficiency - complexity analysis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1704–1708, May 2013.

[127] S. Chaudhuri, *Super-resolution imaging*, vol. 632. Springer Science & Business Media, 2001.

[128] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[129] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014.

[130] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[131] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

[132] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[133] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos, "Content adaptive complexity reduction scheme for quality/fidelity scalable hevc," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1744–1748, IEEE, 2013.

[134] R. Bailleul, J. De Cock, and R. Van De Walle, "Fast mode decision for snr scalability in shvc digest of technical papers," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 193–194, IEEE, 2014.

[135] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos, "An encoder complexity reduction scheme for quality/fidelity scalable hevc," *IEEE Transactions on Broadcasting*, vol. 62, no. 3, pp. 664–674, 2016.

[136] C.-C. Wang, Y.-S. Chang, and K.-N. Huang, "Efficient coding tree unit (ctu) decision method for scalable high-efficiency video coding (shvc) encoder," *Recent Advances in Image and Video Coding*, p. 247, 2016.

[137] X. Li, M. Chen, Z. Qu, J. Xiao, and M. Gabbouj, "An effective cu size decision method for quality scalability in shvc," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 8011–8030, 2017.

[138] W.-J. Chiang, J.-J. Chen, and Y.-H. Tsai, "A fast shvc coding scheme based on base layer co-located cu and cross-layer pu mode information," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 381–386, IEEE, 2017.

[139] Q. Li, B. Liu, and D. Wang, "Fast cu size decision and pu mode decision algorithm for quality shvc inter coding," *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 7819–7839, 2019.

[140] H. R. Tohidypour, H. Bashashati, M. T. Pourazad, and P. Nasiopoulos, "Online-learning-based mode prediction method for quality scalable extension of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 2204–2215, 2016.

[141] X. Zuo and L. Yu, "Fast mode decision method for all intra spatial scalability in shvc," in *2014 IEEE Visual Communications and Image Processing Conference*, pp. 394–397, IEEE, 2014.

[142] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos, "Adaptive search range method for spatial scalable hevc," in *2014 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 191–192, IEEE, 2014.

[143] H. R. Tohidypour, M. T. Pourazad, and P. Nasiopoulos, "Probabilistic approach for predicting the size of coding units in the quad-tree structure of the quality and spatial scalable hevc," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 182–195, 2015.

[144] X. Lu, C. Yu, Y. Gu, and G. Martin, "A fast intra coding algorithm for spatial scalability in shvc," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1792–1796, IEEE, 2018.

[145] I. Wali, A. Kessentini, M. A. B. Ayed, and N. Masmoudi, "Fast inter-prediction algorithms for spatial scalable high efficiency video coding shvc," *Signal, Image and Video Processing*, vol. 13, no. 1, pp. 145–153, 2019.

[146] L. Shen, P. An, and G. Feng, "Low-complexity scalable extension of the high-efficiency video coding (shvc) encoding system," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 2, p. 44, 2019.

[147] Q. Ge and D. Hu, "Fast encoding method using cu depth for quality scalable hevc," in *2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA)*, pp. 1366–1370, IEEE, 2014.

[148] J. Chen, J. Boyce, Y. Ye, and M. Hannuksela, "Shvc test model 9 (shm 9) introduction and encoder description," *Document JCTVC-T1007. Geneva, Switzerland*, 2015.

[149] T. Biatek, W. Hamidouche, J.-F. Travers, and O. Deforges, "Optimal bitrate allocation in the scalable hevc extension for the deployment of uhd services," *IEEE Transactions on Broadcasting*, vol. 62, no. 4, pp. 826–841, 2016.

[150] A. Mackin, K. C. Noland, and D. R. Bull, "High frame rates and the visibility of motion artifacts," *SMPTE Motion Imaging Journal*, vol. 126, no. 5, pp. 41–51, 2017.

[151] Y. Kuroki, T. Nishi, S. Kobayashi, H. Oyaizu, and S. Yoshimura, "A psychophysical study of improvements in motion-image quality by using high frame rates," *Journal of the Society for Information Display*, vol. 15, no. 1, pp. 61–68, 2007.

[152] K. Noland, "The application of sampling theory to television frame rate requirements," *BBC Research & Development White Paper*, vol. 282, 2014.

[153] J. Laird, M. Rosen, J. Pelz, E. Montag, and S. Daly, "Spatio-velocity csf as a function of retinal velocity using unstabilized stimuli," in *Human Vision and Electronic Imaging XI*, vol. 6057, p. 605705, International Society for Optics and Photonics, 2006.

[154] A. Mackin, F. Zhang, M. A. Papadopoulos, and D. Bull, "Investigating the impact of high frame rates on video compression," in *Image Processing (ICIP), 2017 IEEE International Conference on*, pp. 295–299, IEEE, 2017.

[155] R. Salmon, T. Borer, M. Pindoria, M. Price, and A. Sheikh, "Higher frame rates for television," 2013.

[156] "The present state of ultra-high definition television," *ITU-R Report BT.2246-6*, March 2017.

[157] M. Emoto and M. Sugawara, "Critical fusion frequency for bright and wide field-of-view image display," *Journal of Display Technology*, vol. 8, no. 7, pp. 424–429, 2012.

[158] R. Salmon, M. Armstrong, and S. Jolly, "Higher frame rates for more immersive video and television," *BBC White Paper WHP*, vol. 209, 2011.

[159] P. G. Barten, *Contrast sensitivity of the human eye and its effects on image quality*, vol. 19. Spie optical engineering press Bellingham, WA, 1999.

[160] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Vision Models and Applications to Image and Video Processing*, pp. 179–200, Springer, 2001.

[161] R. Selfridge, K. C. Noland, and M. Hansard, "Visibility of motion blur and strobing artefacts in video at 100 frames per second," in *Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016)*, p. 3, ACM, 2016.

[162] M. Emoto, Y. Kusakabe, and M. Sugawara, "High-frame-rate motion picture quality and its independence of viewing distance," *Journal of Display Technology*, vol. 10, no. 8, pp. 635–641, 2014.

[163] EBU, "Ebu policy statement on ultra high definition television," in *European Broadcasting Union, Grand-Saconnex, Switzerland.*

[164] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 3407–3411, IEEE, 2015.

[165] Y. Sugito, S. Iwasaki, K. Chida, K. Iguchi, K. Kanda, X. Lei, H. Miyoshi, and K. Kazui, "A study on the required video bit-rate for 8k 120hz hevc temporal scalable coding," in *Picture Coding Symposium (PCS), 2018*, pp. 1–5, IEEE, 2018.

[166] J.-J. Chen and H.-M. Hang, "Source model for transform video coder and its application. ii. variable frame rate coding," *IEEE transactions on circuits and systems for video technology*, vol. 7, no. 2, pp. 299–311, 1997.

[167] H. Song and C.-C. Kuo, "Rate control for low-bit-rate video via variable-encoding frame rates," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 512–521, 2001.

[168] Y. Yuan, D. Feng, and Y. Zhong, "Fast adaptive variable frame-rate coding," in *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*, vol. 5, pp. 2734–2738, IEEE, 2004.

[169] H. Shu and L.-P. Chau, "Variable frame rate transcoding considering motion information [video transcoding]," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, pp. 2144–2147, IEEE, 2005.

[170] Z. Bojkovic and A. Samcovic, "Variable frame rates control strategy based on human visual system," in *Computer as a Tool, 2005. EUROCON 2005. The International Conference on*, vol. 1, pp. 179–182, IEEE, 2005.

[171] P. Usach, J. Sastre, and J. Lopez, "Variable frame rate and gop size h. 264 rate control for mobile communications," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 1772–1775, IEEE, 2009.

[172] Q. Huang, S. Y. Jeong, S. Yang, D. Zhang, S. Hu, H. Y. Kim, J. S. Choi, and C.-C. J. Kuo, "Perceptual quality driven frame-rate selection (pqd-frs) for high-frame-rate video," *IEEE Transactions on Broadcasting*, vol. 62, no. 3, pp. 640–653, 2016.

[173] A. V. Katsenou, D. Ma, and D. R. Bull, "Perceptually aligned frame rate selection using spatio temporal features," in *Picture Coding Symposium (PCS), 2018*, pp. 1–5, IEEE, 2018.

[174] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*, pp. 363–370, Springer, 2003.

[175] R. M. Haralick, K. Shanmugam, I. Dinstein, *et al.*, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[176] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, "Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 5, pp. 671–682, 2012.

[177] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-star: a perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2473–2486, 2014.

[178] R. M. Nasiri and Z. Wang, "Perceptual aliasing factors and the impact of frame rate on video quality," in *Image Processing (ICIP), 2017 IEEE International Conference on*, pp. 3475–3479, IEEE, 2017.

[179] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *Image Processing (ICIP), 2017 IEEE International Conference on*, pp. 300–304, IEEE, 2017.

[180] R. M. Nasiri, Z. Duanmu, and Z. Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *Image Processing (ICIP), 2018 IEEE International Conference on*, IEEE, 2018.

[181] F. Navarro, F. J. Serón, and D. Gutierrez, "Motion blur rendering: State of the art," in *Computer Graphics Forum*, vol. 30, pp. 3–26, Wiley Online Library, 2011.

[182] T. Brooks and J. T. Barron, "Learning to synthesize motion blur," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6840–6848, 2019.

[183] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[184] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014.

[185] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

[186] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 670–679, 2017.

[187] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 261–270, 2017.

[188] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9000–9008, 2018.

[189] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[190] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," 1984.

[191] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[192] N. Chinchor, "Muc-4 evaluation metrics," in *Proceedings of the 4th conference on Message understanding*, pp. 22–29, Association for Computational Linguistics, 1992.

[193] Wikipedia Foundation, "Power (statistics)," 2015. [Online; accessed 05-September-2019].

[194] Ffmpeg Developers, "Ffmpeg tool."

[195] G. Herrou, W. Hamidouche, and L. Morin, "Low-complexity spatial scalability scheme using hevc for 4k and vr videos," in *2018 Data Compression Conference*, pp. 411–411, IEEE, 2018.

[196] G. Herrou, W. Hamidouche, and L. Morin, "Wavelet decomposition pre-processing for spatial scalability video compression scheme," in *2018 Picture Coding Symposium (PCS)*, pp. 149–153, IEEE, 2018.

[197] G. Herrou, W. Hamidouche, and L. Morin, "Low-complexity scalable encoder based on local adaptation of the spatial resolution," in *2019 International Conference on Image Processing (ICIP)*, IEEE, 2019.

[198] G. Herrou, J.-Y. Aubié, W. Hamidouche, and L. Morin, "Procédé et dispositif de codage et de décodage de données correspondant à une séquence vidéo," *French patent application FR1900950, filed on January 1st 2019*.

[199] J.-Y. Aubié, P. Duménil, W. Hamidouche, and G. Herrou, "Procédé de formation d'une séquence d'images de sortie à partir d'une séquence d'images d'entrée, procédé de reconstruction d'une séquence d'images d'entrée à partir d'une séquence d'images de sortie, dispositifs, équipement serveur, équipement client et programmes d'ordinateurs associés," *PCT application PCT/EP2019/070290, filed on July 26th 2019, under priority of French patent application FR1857078, filed on July 30th 2018*.

[200] M. Viitanen, A. Koivula, A. Lemmetti, A. Ylä-Outinen, J. Vanne, and T. D. Hämäläinen, "Kvazaar: open-source hevc/h. 265 encoder," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1179–1182, ACM, 2016.

[201] W. Hamidouche, M. Raulet, and O. Déforges, "Parallel shvc decoder: Implementation and analysis," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2014.

[202] A. Mercat, A. Lemmetti, M. Viitanen, and J. Vanne, "Acceleration of kvazaar hevc intra encoder with machine learning," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2676–2680, IEEE, 2019.

[203] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Appendix A

Publications and Patents

A.1 Scientific journal

[J1] G. Herrou, C. Bonnineau, W. Hamidouche, P. Duménil and L. Morin,
"Machine Learning driven Variable Frame-Rate in Video Broadcast Applications",
submitted and under-review for publication in the *IEEE Transactions on Broadcasting* journal, September 2019.

A.2 International Conferences

[C1] G. Herrou, W. Hamidouche and L. Morin,
"Low-Complexity Spatial Scalability Scheme using HEVC for 4K and VR Videos",
presented at the *Data Compression Conference (DCC), March 2018.*

[C2] G. Herrou, W. Hamidouche and L. Morin,
"Wavelet Decomposition Pre-processing for Spatial Scalability Video Compression Scheme",
presented at the *Picture Coding Symposium (PCS) 2018, June 2018.*

[C3] G. Herrou, W. Hamidouche and L. Morin,
"Low-complexity Scalable Encoder based on Local Adapatation of the Spatial Resolution",
presented at the *International Conference on Image Processing (ICIP) 2019, September 2019.*

A.3 Patents

[P1] G. Herrou, J-Y. Aubié, W. Hamidouche and L. Morin,
"Procédé et dispositif de codage et de décodage de données correspondant à une séquence vidéo",
French patent application FR1900950, filed on January 31st 2019.

[P2] J-Y Aubié, P. Duménil, W. Hamidouche and **G. Herrou**,
"Procédé de formation d'une séquence d'images de sortie à partir d'une séquence d'images d'entrée, procédé de reconstruction d'une séquence d'images d'entrée à partir d'une séquence d'images de sortie, dispositifs, équipement serveur, équipement client et programmes d'ordinateurs associés",
PCT application PCT/EP2019/070290, filed on July 26th 2019, under priority of French patent application FR1857078, filed on July 30th 2018.

INSA de RENNES
Service des Formations

AVIS DU JURY SUR LA REPRODUCTION DE LA THESE SOUTENUE

Titre de la thèse:
Résolution Spatio-temporelle Adaptative pour un Codage Léger des Formats Vidéo Émergents

Nom Prénom de l'auteur : HERROU GLENN

Membres du jury :
- Monsieur BULL David
- Monsieur VANNE Jarno
- Madame MORIN Luce
- Monsieur HAMIDOUCHE Wassim
- Monsieur RICORDEL Vincent
- Monsieur CAGNAZZO Marco
- Monsieur DUFAUX Frédéric

Président du jury : Frédéric Dufaux

Date de la soutenance : 26 Novembre 2019

Reproduction de la these soutenue

☒ Thèse pouvant être reproduite en l'état
☐ Thèse pouvant être reproduite après corrections suggérées

Fait à Rennes, le 26 Novembre 2019

Signature du président de jury

Le Directeur,

M'hamed DRISSI

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

INSA RENNES

Titre: Résolution Spatio-temporelle Adaptative pour un Codage à Faible Complexité des Formats Vidéo Émergents

Mot clés : Compression vidéo, UHDTV, résolution adaptative, scalabilité, réduction de compléxité

Resumé : La standardisation du dernier format vidéo en date, appelé *Ultra-High Definition TV* (UHDTV), vise à améliorer la qualité d'expérience des utilisateurs en introduisant de nouvelles technologies telles que la 4K ou le *High Frame-Rate* (HFR). Cependant, ces améliorations multiplient la quantité de données à traiter avant transmission du signal par un facteur 8.

En plus de ce nouveau format, les fournisseurs de contenu doivent aussi encoder les vidéos dans des formats et à des débits différents du fait de la grande variété des systèmes et réseaux utilisés par les consommateurs.

SHVC, l'extension scalable du dernier standard de compression video *High Efficiency Video Coding* (HEVC) est une solution prometteuse pour adresser ces problématiques. En revanche, son architecture, très demandeuse en termes de calculs, atteint ses limites lors de l'encodage des nouveaux formats vidéo immersifs tels que le standard UHDTV.

L'objectif de cette thèse est donc d'étudier des approches de codage scalables et légères basées sur l'adaptation de la résolution spatio-temporelle des vidéos. La première partie de cette thèse propose deux algorithmes de pré-traitement, utilisant respectivement des approches polyphase et ondelette basées image, afin de permettre la scalabilité spatiale avec une faible augmentation de la complexité.

Ensuite, dans un second lieu, le design d'une architecture scalable à deux couches, plus conventionnelle, est étudié. Celle-ci est composée d'un encodeur HEVC standard dans la couche de base pour assurer la compatibilité avec les systèmes existants. Pour la couche d'amélioration, un encodeur basse complexité, se basant sur l'adaptation locale de la résolution spatiale, est proposé.

Enfin, la dernière partie de cette thèse se focalise sur l'adaptation de la résolution spatio-temporelle. Un algorithme faisant varier la fréquence image est d'abord proposé. Cet algorithme est capable de détecter localement et de façon dynamique la fréquence image la plus basse n'introduisant pas d'artefacts visibles liés au mouvement. Les algorithmes de fréquence image variable et de résolution spatiale adaptative sont ensuite combinés afin d'offrir un codage scalable à faible complexité des contenus 4K HFR.

Title: Adaptive Spatio-temporal Resolution for Lightweight Coding of Emerging Video Formats

Keywords : Video coding, UHDTV, adaptive resolution, scalability, complexity reduction

Abstract : The definition of the latest Ultra-High Definition TV (UHDTV) standard aims to increase the user's quality of experience by introducing new video signal features such as 4K and High Frame-Rate (HFR). However, these new features multiply by a factor 8 the amount of data to be processed before transmission to the end user.

In addition to this new format, broadcasters and Over-The-Top (OTT) content providers have to encode videos in different formats and at different bitrates due to the wide variety of devices with heterogeneous video format and network capacities used by consumers.

SHVC, the scalable extension of the latest video coding standard High Efficiency Video Coding (HEVC) is a promising solution to address these issues but its computationally demanding architecture reaches its limit with the encoding and decoding of the data-heavy newly introduced immersive video features of the UHDTV video format.

The objective of this thesis is thus to investigate lightweight scalable encoding approaches based on the adaptation of the spatio-temporal resolution. The first part of this document proposes two pre-processing tools, respectively using polyphase and wavelet frame-based approaches, to achieve spatial scalability with a slight complexity overhead.

Then, the second part of this thesis addresses the design of a more conventional dual-layer scalable architecture using an HEVC encoder in the Base Layer (BL) for backward compatibility and a proposed low-complexity encoder, based on the local adaptation of the spatial resolution, for the Enhancement Layer (EL).

Finally, the last part of this thesis investigates spatio-temporal resolution adaptation. A variable frame-rate algorithm is first proposed as pre-processing. This solution has been designed to locally and dynamically detect the lowest frame-rate that does not introduce visible motion artifacts. The proposed variable frame-rate and adaptive spatial resolution algorithms are then combined to offer a lightweight scalable coding of 4K HFR video contents.