Tampere University

RAMIN GHAZNAVI YOUVALARI

Encoding and Streaming Solutions for Immersive Virtual Reality Video

Tampere University Dissertations 375

Tampere University Dissertations 375

RAMIN GHAZNAVI YOUVALARI

Encoding and Streaming Solutions for Immersive Virtual Reality Video

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for online public discussion
on Friday, 12 February 2021, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos
Professor Moncef Gabbouj
Tampere University
Finland

Pre-examiners
Dr. Gwendal Simon
Huawei Technologies
France

Dr.-Ing. Cornelius Hellge
Fraunhofer Heinrich Hertz Institute
Germany

Opponent
Priv. Doz. Dr.-Ing. habil. Mathias Wien
RWTH Aachen University
Germany

Cover design: Roihu Inc.

ISBN 978-952-03-1851-2 (print)
ISBN 978-952-03-1852-9 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1852-9

PunaMusta Oy – Yliopistopaino
Joensuu 2021

ABSTRACT

Immersive virtual reality (VR) technology is becoming mainstream nowadays. This technology makes use of omnidirectional content in order to create immersion in the virtual environment. Omnidirectional content is captured in a way that it covers the entire 360° field-of-view (FOV) around the capturing device. Thus, it is able to create the three Degrees-of-Freedom (3-DoF) experience in VR. In order to create an immersive experience, VR technology is required to use stereoscopic omnidirectional video in high resolution, quality and frame rates. Such requirements introduce significant challenges in the encoding and streaming stages of this technology.

The most common way of compressing omnidirectional video is by means of existing 2D image/video codecs such as High Efficiency Video Coding (HEVC/H.265) and Versatile Video Coding (VVC/H.266) standards. Therefore, this spherical content is projected over 2D image planes to be used in the 2D chain of the codec. However, the projection process introduces different sampling characteristics to the content compared to the spherical version. These characteristics can be represented as over-sampling of the content in different parts of the projected image. This over-sampling results in content stretching, deformations and non-linear motion behaviors. On the other hand, existing codecs are not optimized for such behaviors in the content, consequently, the resulting compression performance is sub-optimal for the projected video.

This thesis investigates and proposes new approaches for improving the motion estimation and compensation performances for non-linear motion of the projected omnidirectional videos in HEVC and VVC standards. The first contribution for this purpose is the motion vector scaling method, which attempts to provide uniform motion vector predictors for the coding block. The scaling factor is derived based on the geometry characteristics of the projection plane and the position of the blocks in that plane. In the second and third contributions, a novel method is proposed for adaptively and efficiently predicting the motion information of the block

based on a learning process from the neighboring motion information in full block and sub-block levels. The performances of the proposed methods have been assessed over diverse video datasets which are commonly used in the standardization activities and by following the standard simulation protocols and were shown to provide high compression improvements while retaining codec's complexity in a reasonable range.

In recent years, tile-based viewport-adaptive streaming (VAS) methods have been considered for delivering omnidirectional content, where a portion of the content, i.e. viewport, is transmitted in the highest resolution and the remaining parts, i.e. non-viewport, are sent in lower resolutions. The reason being that VR content is mainly consumed via Head-Mounted Display (HMD) devices that have limited FOVs for example, 110°×90°. Since a user can see only a portion of the 360° video at each time instance, transmitting the whole VR video at the highest resolution requires a large bandwidth. Even though, tile-based VAS methods provide significantly better streaming performances compared to traditional streaming, but these methods use frequent Intra Random Access Points (IRAPs) for viewport switching. These IRAPs are intra-coded pictures in the bitstream, thus, they include higher bitrates compared to the inter-coded pictures. The frequent IRAPs in the bitstream make the VAS method sub-optimal for VR video streaming.

For resolving the sub-optimal performance of VAS, this thesis develops novel solutions for enabling viewport switching operations without using frequent IRAP pictures in the bitstream. In the first contribution, a multi-layer SHVC-ROI scheme is proposed. The SHVC-ROI method utilizes the inter-layer prediction (ILP) functionality of the codec for coding the high-quality switching points as inter-coded pictures. The use of ILP requires the method to stream the whole 360° low-quality video, hence, no switching occurs for this content. Thus, longer IRAP intervals than conventional ones are used for low-quality content. This streaming configuration resolves the frequent IRAP need in both high- and low-quality content. In the second contribution, a single-layer Simulcast HEVC method is proposed for using infrequent IRAPs in low-quality content. This method follows the same logic as the low-quality coding scheme of the SHVC-ROI where longer IRAP periods are considered and the whole 360° low-quality content is sent to the user. In addition to the mentioned advantages, both of these contributions benefit from not using tiling in the low-quality content, thus, avoid the compression overhead of tiling schemes in

encoding and streaming of such content. Finally, the Shared Coded Picture (SCP) technique is proposed for enabling the viewport switching without frequent IRAPs in both quality versions of the content while using the standard single-layer coding scheme. To this end, certain pictures (i.e. SCPs) in the video are coded in a way that they are identical in both quality versions of the content. Consequently, these identically-coded pictures are used for switching from one version of the bitstream to another. Furthermore, the SCPs are inter predicted from the previous SCP picture in the bitstream. Thus, they require significantly lower bitrates than the intra-coded switching point pictures. The performances of the proposed methods have shown significant streaming bitrate reductions compared to the existing state-of-the-art methods.

PREFACE

First and foremost, I wish to express my deepest gratitude to my supervisor Prof. Moncef Gabbouj for the excellent opportunity that he provided me during my M.Sc. and Ph.D. studies. He gave me valuable insights, support, feedback and encouragement throughout the duration of my studies. I would also like to thank the pre-examiners of this thesis, Prof. Gwendal Simon and Dr. Cornelius Hellge, for their valuable and insightful comments.

I am deeply thankful to my colleagues Dr. Miska Hannuksela and Dr. Alireza Aminlou for their help and amazing support throughout the years. I would also like to thank Miska for his careful review of this thesis and his constructive comments.

I am also thankful to Nokia Technologies for providing the opportunity to conduct my research in the Media Technologies Research LAB. Special thanks to Jari Hagqvist and Jani Lainema for their encouragement and support to do my Ph.D. studies. My gratitude also goes to my colleagues at Nokia, Igor Curcio, Kashyap Kammachi-Sreedhar, Justin Ridge, Antti Hallapuro, Deepa Naik, Henri Toukomaa, Miikka Vilermo, Mikko Pekkarinen, Maryam Homayouni, Emre Aksu, Francesco Cricri and Honglei Zhang for the great work environment.

I would like to thank my great friends Saber Kordestanchi, Masoud Malekzadeh, Saman Bahrampour, Sina Kordestanchi, Armin Bazrafkan, Sounak Bhattacharya, Khazar Khorrami, Zeinab Rezaei, Pouria Hajiani, Umair Ahmed, Sajjad Nouri and Nima Sheikhipour for their friendship and support.

I am thankful also to my family, in particular my sister Fariba and my brother Ahmad, for their significant support throughout my life.

Finally, I wish to thank the love of my life Nahid Sheikhipour. This thesis would not be possible without your endless love, support and encouragement.

Tampere, January 2021
Ramin Ghaznavi Youvalari

CONTENTS

ABBREVIATIONS

2D	Two Dimensional
3-DoF	Three Degrees-of-Freedom
3D	Three Dimensional
6-DoF	Six Degrees-of-Freedom
AMC	Affine Motion Compensation
AMVP	Advanced Motion Vector Prediction
AR	Augmented Reality
ASC	Asymmetric Circular Projection
ATMVP	Alternative Temporal Motion Vector Prediction
AVC/H.264	Advanced Video Coding Standard
BD-Rate	Bjontegaard Delta Bitrate
BL	Base-Layer
CILP	Constraint Inter-layer Prediction
CMP	Cubemap Projection
CPP-PSNR	Crasters Parabolic Projection Peak-Signal-to-Noise Ratio
CTC	Common Test Condition
CU	Coding Unit
DASH	Dynamic Adaptive Streaming over HTTP
DP	Duplicated Picture
EL	Enhancement-Layer
ERP	Equirectangular Projection

FILP	Full Inter-layer Prediction
FOV	Field-of-View
HEVC/H.265	High Efficiency Video Coding Standard
HMD	Head-Mounted Display
ILP	Inter-Layer Prediction
IRAP	Intra Random Access Point
ISOBMFF	ISO Base Media File Format
ISP	Icosahedral projection
JVET	Joint Video Experts Team
MC	Motion Compensation
MCTS	Motion-constraint Tile Set
ME	Motion Estimation
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
MV	Motion Vector
MV-HEVC	Multi-view Extension of High Efficiency Video Coding
MVD	Motion Vector Difference
OHP	Octahedron Projection
OMAF	Omnidirectional MediA Format
PSNR	Peak-Signal-to-Noise Ratio
QAV	Quality Assessment View
QEC	Quality Emphasis Center
QP	Quantization Parameter
RA	Random Access
RMVF	Regression-based Motion Vector Field
ROI	Region-of-Interest
S-PSNR	Spherical Peak-Signal-to-Noise Ratio

SCP	Shared Coded Picture
SE	Syntax Element
SHVC	Scalable Extension of High Efficiency Video Coding
SSP	Segmented Sphere Projection
TSP	Truncated Square Pyramid
VAS	Viewport-adaptive Streaming
VR	Virtual Reality
VVC/H.266	Versatile Video Coding Standard
WS-PSNR	Weighted to Spherically uniform Peak-Signal-to-Noise Ratio
XR	Extended Reality

ORIGINAL PUBLICATIONS

[P1] R. Ghaznavi-Youvalari and A. Aminlou. Geometry-based motion vector scaling for omnidirectional video coding. IEEE International Symposium on Multimedia (ISM). Dec. 2018, 127–130. DOI: 10.1109/ISM.2018.00030.

[P2] R. Ghaznavi-Youvalari and A. Aminlou. Adaptive motion vector prediction for omnidirectional video. IEEE Visual Communications and Image Processing (VCIP). Dec. 2018, 1–4. DOI: 10.1109/VCIP.2018.8698614.

[P3] R. Ghaznavi-Youvalari, A. Aminlou and J. Lainema. Regression-based motion vector field for video coding. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). Sep. 2019. DOI: 10.1109/TCSVT.2019.2942086.

[P4] R. Ghaznavi-Youvalari, A. Zare, H. Fang, A. Aminlou, Q. Xie, M. M. Hannuksela and M. Gabbouj. Comparison of HEVC coding schemes for tile-based viewport-adaptive streaming of omnidirectional video. IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). Oct. 2017, 1–6. DOI: 10.1109/MMSP.2017.8122227.

[P5] R. Ghaznavi-Youvalari, A. Zare, A. Aminlou, M. M. Hannuksela and M. Gabbouj. Shared coded picture technique for tile-based viewport-adaptive streaming of omnidirectional video. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT). Oct. 2018, 3106–3120. DOI: 10.1109/TCSVT.2018.2874179.

1 INTRODUCTION

1.1 General Context

Technological advancements in the past decades have revolutionized the way we consume information. To name a few, there has been significant industrial developments in multimedia acquisition, processing, storage and display domains. On top of these, the wide availability of WiFi, 4G/5G networks provided us a medium that we can easily consume and share more data every day. Such advancements enable technologies which earlier were science fiction and could not be implemented due to limitations of the existing technologies back then. Among them is immersive media technology where the user can create a realistic model of the real-world in a virtual environment in way that he/she can interact with and/or navigate through it. Examples of immersive technologies are virtual reality (VR), augmented reality (AR) and extended reality (XR).

Virtual reality (VR) is the most feasible and promising emerging technology nowadays. There are vast enablers of this technology in the consumer market, such as commercial capturing cameras that are able to record high-quality VR content (e.g., Ricoh Theta V [67], GoPro Max 360 [30], and Gear360 [69]), as well as widely available display devices for VR (e.g., Gear VR [70], Oculus VR headsets [60, 61, 62], and Vive [92]), increasing support in content sharing platforms (e.g., Facebook [17] and YouTube [100]). These means resulted in the fast adoption of VR in the consumer market, with a variety of use cases. Examples of VR applications are, gaming, streaming live events (sports, concerts, etc.), as well as health and education.

The immersive experience through VR can be defined as the degrees-of-freedom (DoF) that this technology provides to the users. Hence, depending on the degrees-of-freedom, the immersive experience could be separated into different phases. For example, Figure 1.1 illustrates the standardization roadmap [11, 57] by the Moving Picture Experts Group (MPEG) which divides the standardization of this technol-

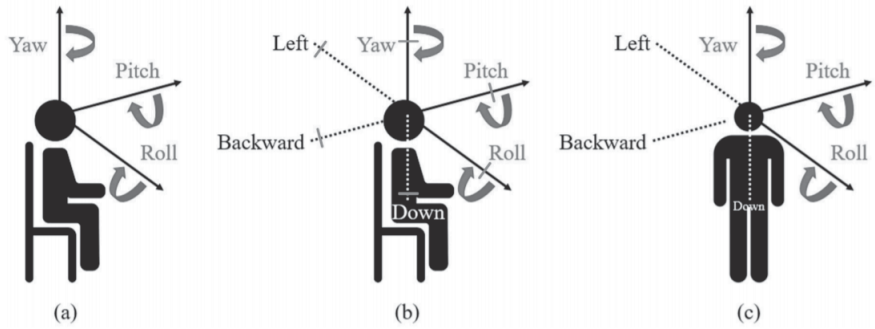


Figure 1.1 Viewing angle and degrees of freedom in VR: (a) 3-DoF, (b) 3-DoF+, (c) 6-DoF [11]

ogy into three phases, where each phase enables certain degrees of immersive capabilities in VR.

- 3-DoF (Figure 1.1.a): Provides limited degrees of freedom to the user in VR by allowing only rotational motion (i.e. yaw, pitch and roll). This is achieved by watching 360° omnidirectional content through e.g. Head Mounted Display devices.
- 3-DoF+ (Figure 1.1.b): Provides limited movements to the 3-DoF VR. If the viewing orientation that the viewer is attempting to watch does not exist, a synthesis process is executed in order to synthesize the virtual view by making use of depth information.
- 6-DoF (Figure 1.1.c): Provides fully immersive experience by allowing three rotational degrees of freedom as well as three translations degrees of freedom to the user. In this technology, the user can move within and interact with the virtual environment.

This thesis focuses on Three Degrees-of-Freedom (3-DoF) VR. Consequently, throughout the thesis, the terms 360° video, VR and immersive video are used interchangeably in the context of 3-DoF VR technology and its use cases.

In order to create the feel of immersion in the virtual environment, the 3-DoF VR technology makes use of omnidirectional content. The omnidirectional content cover 360° field-of-view (FOV) around the capturing device. The 360° coverage of the scene, can enable VR technology to create the 3-DoF experience to the user in the virtual environment. Because of the 360° FOV of these content, they are often

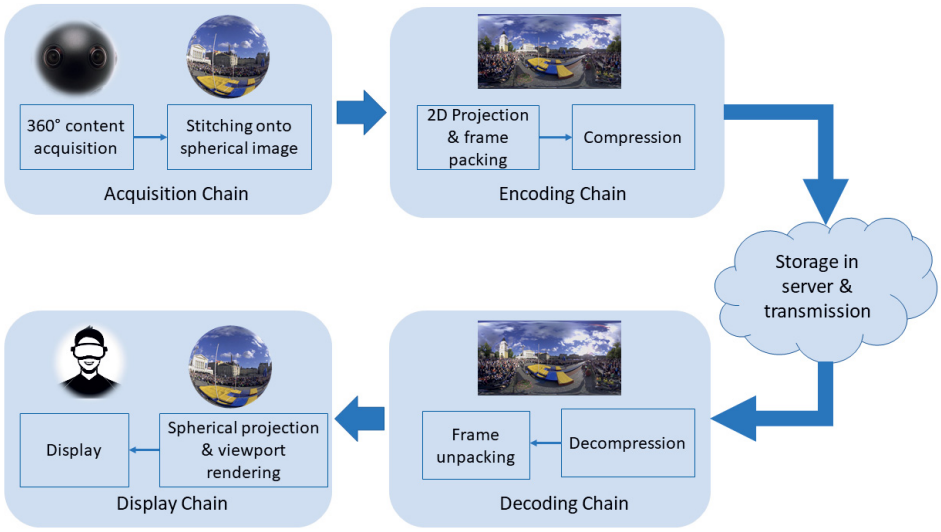


Figure 1.2 An illustration of end-to-end processing pipeline for VR content

referred to as spherical, 360° and panoramic images and videos.

Figure 1.2 illustrates an example of the end-to-end processing chain, from content acquisition to display, for VR. In the following, there is a brief description of this chain.

Omnidirectional content can be captured through different means. It can be captured by multi-camera setup or a single camera with multiple lenses. There are many commercial cameras for this purpose in the market [30, 67, 69]. For generating the 360° scene, the captured views with multi-camera setup, are stitched together on the spherical domain. In the compression stage, this content is projected into two-dimensional (2D) image formats. The reason for such conversion is that the existing state-of-the-art compression standards, such as High Efficiency Video Coding (HEVC/H.265) [36] and Versatile Video Coding (VVC/H.266) [84], are designed to operate only over 2D content. The compressed bitstreams are stored in the server side and based on the request from the end-user, the corresponding bitstream in a certain resolution/quality is selected and transmitted. At the user side, the received bitstream is decoded and projected back into a spherical domain. Finally, the portion of the content that is in the user's viewing orientation is rendered and displayed.

The typical way of consuming VR content is via Head-Mounted Display (HMD) devices. In order to have a proper immersive experience in VR, the technology requires to have stereoscopic content with very high resolution, quality and frame rates. The viewport resolution of the existing HMD devices in the consumer market is in the range of 1080×1200 to 1440×1600 pixels per eye with the FOVs of around $110° \times 90°$ [55, 56]. For enabling such viewport resolutions, the omnidirectional video should be around 6K (6144×3072 pixels) resolution, per eye. Another important factor for the immersive VR is the low motion-to-photon delay (around 11 ms for 90 Hz) requirement. This is important for avoiding commonly encountered motion sickness while using the HMD devices. Considering all these requirements, streaming 6K stereoscopic omnidirectional video at 90 Hz frame rate would bring significant issues in storage and transmission of this content. In recent years, the Moving Pictures Experts Group (MPEG) of ISO/IEC has started working on developing standards in order to cope with the immersive VR content more efficiently [74, 96]. The aim of this thesis is to provide solutions for further improving the compression and streaming aspects of the described chain for VR technology.

1.2 Objectives

The common way of compressing spherical omnidirectional content is to use existing state-of-the-art compression standards such as Advanced Video Coding (AVC/H.264) [1], High Efficiency Video Coding (HEVC/H.265) [36], and Versatile Video Coding (VVC/H.266) [84]. However, these codecs are designed to compress only 2D content and are not able to operate on 3D omnidirectional signals. Hence, these spherical images and videos are projected into 2D representations. Different projection methods have been studied in recent years and some of them are described in Chapter 2. Among the studied formats, equirectangular projection (ERP) and cubemap projection (CMP) formats are the most popular ones for sphere to 2D projection purposes. The projection formats that are used for 2D conversion do not represent the same spherical characteristics of the content in the projection plane. For example, the ERP projection suffers from over-sampling and deformations of the content in areas near the poles. The CMP format has content discontinuity and over-sampling problems in the 2D image plane. These characteristics of the content are not very suitable for the existing codecs. Existing codecs are tuned in way that they perform

well on typical 2D planar content, hence, the compression performance with these codecs would be sub-optimal for omnidirectional content coding.

In recent years, viewport-adaptive streaming (VAS) methods have been considered for VR applications, instead of traditional streaming, where only the portion of the content that falls into the viewing orientation of the user is transmitted in a high resolution and/or quality, while the remaining parts are streamed in a lower resolution and/or quality. Even though these methods improve the streaming performance compared to transmitting the omnidirectional video in the traditional way, they use frequent Intra Random Access Points (IRAPs) for switching from one viewing orientation to another. The IRAP switching points are intra-coded pictures and they consume significantly higher bitrates compared to the inter-coded pictures. In order to have seamless viewport switching in the content, the streaming method must have frequent IRAP pictures. Thus, the frequent IRAPs make the viewport-adaptive streaming operation sub-optimal.

The first research question we pose in this thesis is as follows: can we develop more efficient coding tools to improve the compression performance of current codecs in order to deal with immersive virtual reality video? When such a video is transmitted to a viewer, the second research question is how can we stream a high quality immersive video in a more efficient way to the viewer? The research work in the thesis targets to answer these two main research questions and sets the following as specific objectives:

Objective 1: investigate and develop new coding tools in order to improve the sub-optimal compression performance of the HEVC and VVC codecs for immersive video. To this end, the thesis focuses on improving the motion compensated prediction process while maintaining the complexity of the codec at a reasonable level.

Objective 2: improve the viewport-adaptive streaming performance of VR video. As mentioned before, the end-user watches only a portion of the 360° scene at each time instance. Thus, streaming the whole 360° content in the highest resolution and quality is not considered as an efficient approach. The thesis investigates novel methods in order to enable seamless viewport switching operations without the need for frequent IRAP pictures in the tile-based viewport-adaptive streaming practices. The objective for streaming targets viewport-adaptive streaming of VR video over HTTP Adaptive Streaming (DASH) [41] for On-Demand VR content use cases.

1.3 Thesis Outline

The rest of the thesis is organized as follows.

Chapter 2 provides a review of the recent works for improving the compression and streaming performances of omnidirectional video. Section 2.1, reviews the 2D projection formats that are proposed for 360° content. The compression algorithms are described in Section 2.2, whereas the streaming methods are reviewed in Section 2.3. This is an active research area and many of the methods reviewed in the literature continue to be developed along with the thesis work.

Chapter 3 describes the coding tools that are studied in this thesis for improving the inter prediction performance of omnidirectional video. Section 3.1 provides the testing conditions and performance evaluation methodologies for the proposed methods. Three methods are proposed in this chapter. The first method, in Section 3.2, proposes a motion vector scaling technique based on the projection geometry of the content in order to provide a more uniform motion vector predictors. In Section 3.3, an adaptive motion vector prediction scheme is proposed, where a 6-parameter motion model is used for predicting the motion vectors of block based on the neighboring motion information. Finally in Section 3.4, the motion model of Section 3.3 is extended to operate in 4×4 and 8×8 sub-block levels, in order to improve the compression performance.

Chapter 4 describes the tile-based viewport-adaptive streaming solutions that are proposed in this work. In Section 4.1, the testing conditions and performance evaluation methodologies for the proposed methods are provided. In Section 4.2, two methods are used for improving the streaming bitrate by avoiding the IRAP pictures fully or partially in the switching points. In Section 4.3, the novel Shared Coded Picture (SCP) technique is proposed. The SCP-based method removes the necessity of IRAPs in viewport switching points in a way that the generated bitstream is decodable with standard single-layer decoders.

Finally, Chapter 5 provides the conclusion of the thesis and points to possible future work.

1.4 Author's Contribution

[P1] This publication provides a geometry-based motion vector scaling method in order to improve the motion compensated prediction efficiency of the omnidirectional content. The candidate is the sole responsible for creating and implementing the idea, as well as writing the paper.

[P2] This publication introduces an adaptive motion vector prediction scheme based on a linear regression model. The candidate contributed significantly in creating and implementing the algorithm and writing the paper.

[P3] This publication extends the adaptive motion vector prediction method of the previous publication to be used in sub-block levels. The author is the main person for the implementing the algorithm and writing the publication.

[P4] This paper provides two methods for improving the streaming performance of omnidirectional video compared to the existing methods. The candidate is contributed to creating and implementing the algorithms. The majority of the paper is written by the candidate.

[P5] This publication presents a novel streaming method for VR video that does not require intra-coded pictures for viewport switching operations. The candidate is contributed to the algorithm creation and developments. Moreover, the implementation of the methods are also done by the candidate. The majority of the publication is written by the candidate.

2 BACKGROUND AND RELATED WORK

This section provides an overview of the related works which have been conducted in recent years on omnidirectional images and video from the compression and streaming perspectives. Several methods and systems described in the related works have been developed concurrently with the dissertation, hence, links and overlaps will be pointed out.

The chapter is organized as follows. The common projection formats for omnidirectional content are described in Section 2.1. The recent developments of compression algorithms for omnidirectional video is reviewed in Section 2.2, while existing streaming solutions of such content are described in Section 2.3.

2.1 Omnidirectional Projection Formats

Omnidirectional images and video represent the 360° surrounding in 3D domain, hence it is also referred to as spherical or 360° content. Such content can be characterized by a sphere where the 360° field-of-view (FOV) of the surrounding capturing device is projected on the surface of a sphere, while the viewer or the virtual camera may be located in the center of the sphere. However, this spherical representation of omnidirectional content is not suitable for existing image and video compression standards. The state-of-the-art compression standards such as Advanced Video Coding (AVC/H.264) [1, 22, 54, 80], High Efficiency Video Coding (HEVC/H.265) [36, 79, 82] and Versatile Video Coding (VVC/H.266) [10], are designed to operate on 2D image and video formats. Thus, in order to make use of the legacy codecs for compressing such data formats, omnidirectional content must be converted into 2D representations of the data using sphere to plane projections [78]. The most popular formats for this purpose are equirectangular projection (ERP) and cubemap projection (CMP) formats. Figure 2.1 shows an example of the spherical to 2D ERP projection process. In this format, the spherical content is projected onto a cylinder

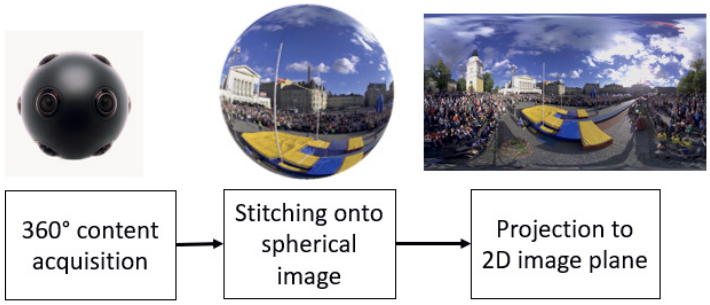


Figure 2.1 An illustration of the spherical to ERP projection

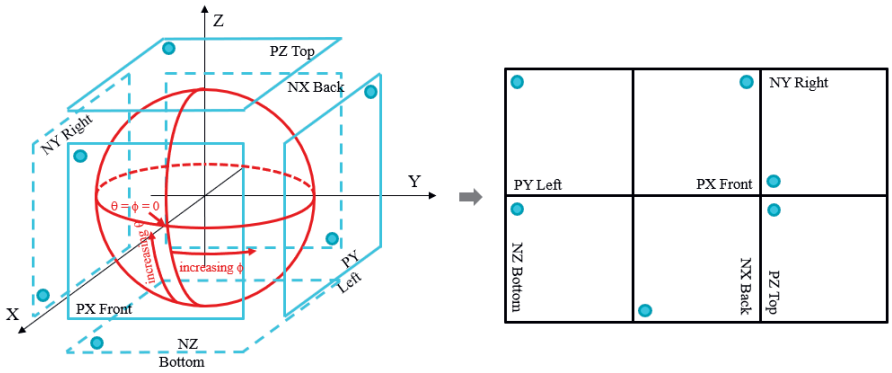


Figure 2.2 An illustration of the cubemap projection [94]

and then it is unfolded to a rectangular form. The horizontal and vertical axes in the ERP correspond to the longitude (360°) and latitude (180°) in the spherical domain. As can be observed from the figure, the ERP projection suffers from over-sampling of the content in the polar areas, resulting in content deformations on the projected plane.

Figure 2.2 illustrates the mapping for the cubemap projection (CMP) format. In this case, it is assumed that the sphere is bounded by a cube, then the 3D content is projected onto the six faces of the cube. For 2D representation purpose, the cube faces are unfolded and packed into a rectangular format as shown in the figure. Different arrangements for cube faces may be considered, such as 2×3 configuration (as shown in Figure 2.2), 3×2 or 6×1. Moreover, the arrangement of the cube faces may result in different compression performances [111] due to the continuity

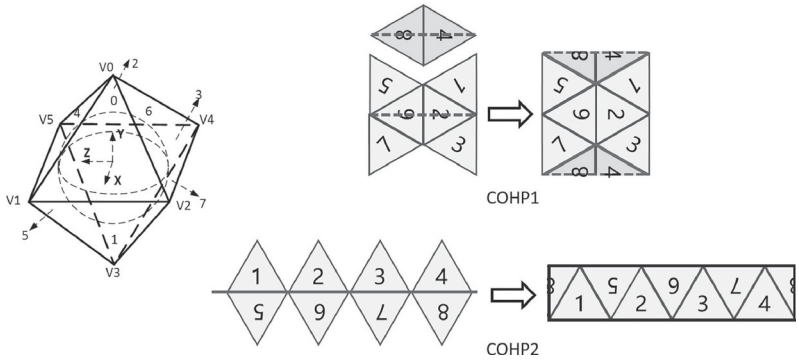


Figure 2.3 An illustration of OHP projection [52, 53]

or discontinuity of the content between cube faces.

The Octahedron projection (OHP) [52, 53], illustrated in Figure 2.3, is another projection that is used for encoding the 360° content. The OHP format consists of 8 triangle faces and 6 vertices. For projection, the 360° sphere is assumed to be inside the OHP shape and then the spherical content is mapped onto each of the triangular faces of the OHP. Later, these faces are unfolded and packed into a rectangular frame for encoding. Different packing methods, referred to as compact layouts, have been studied for arranging the triangular faces in the rectangular plane. An extended version of OHP, proposed in [2, 102], is called Icosahedral projection (ISP). The ISP projection uses 20 triangular faces and 12 vertices in its format. Similar to OHP, the triangular faces are packed into a rectangular frame for the compression stage. An example of ISP format is shown in Figure 2.4.

Segmented sphere projection (SSP) [109, 110] is yet another format for omnidirectional video projection. The SSP method divides the 360° content into north pole, south pole and equator segments, where the polar areas are mapped into circular segments while the equator is mapped into a rectangular areas. Figure 2.5 shows an example of this format with the corresponding packing of the segments in the frame.

2.2 Compression Methods for Omnidirectional Video

As explained above, the 3D spherical content must be projected into 2D image planes in order to utilize the current state-of-the-art codecs for compressing such content.

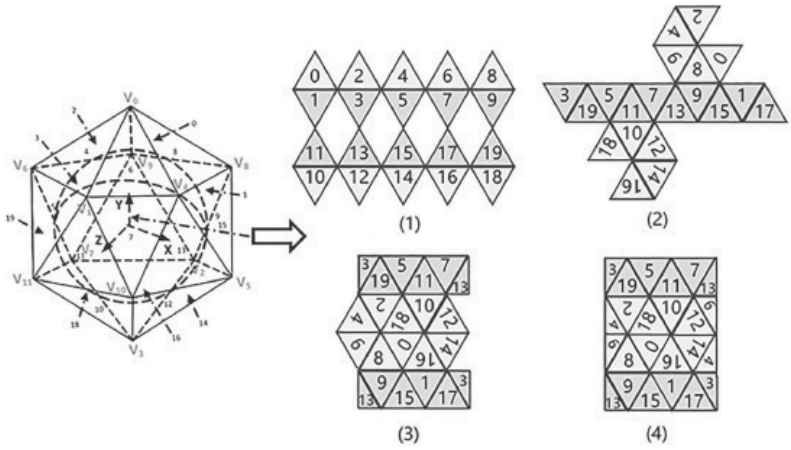


Figure 2.4 An illustration of ISP projection [2]

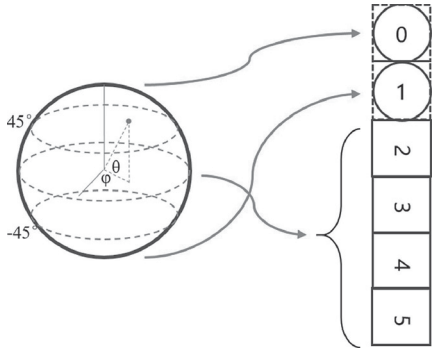


Figure 2.5 An illustration of SSP projection [109]

The most common projection formats that are used for omnidirectional video are equirectangular projection (ERP) [78] and cubemap projection (CMP) formats. The main reason for the popularity of these formats are because of their ease of use, wide support in content authoring tools, particularly for ERP, and wide support in software development environments. However, these 2D formats include particular sampling characteristics, compared to conventional 2D planar formats, which makes the current codecs not so efficient for compression purposes. For example, the ERP format suffers from severe deformations of content caused by different sampling characteristics of its projection plane. To be more precise, the sampling density varies in the spherical domain in a way that toward the polar areas, there are less

number of samples compared to equator areas. However, the projection plane that is used for mapping from 3D to 2D, allocates equal sampling in every part of the 2D image plane. This results in over-samplings and deformations of the content in the 2D domain, especially in the polar areas. This sampling characteristics of ERP can be represented as a non-linear sampling behavior of objects in different parts of the image. In case of video, it can be represented as a non-linear motion in objects from one frame to another. Moreover, the current compression standards, e.g., AVC and HEVC, are not ideal for predicting such non-linear samples distribution in intra prediction and motion compensated prediction processes. For the cubemap projection format, even though the content deformations are not as severe as in the ERP format, there are still some deformations and stretching of content close to cube face boundaries. Moreover, since the 360° content is projected in different cube faces, there exists some discontinuity of pixels from one face to another which makes the intra prediction across the cube faces inefficient. Such discontinuity makes the inter prediction process sub-optimal as well. Especially, when the motion in the content is from one face of the cube in a frame to another face of the cube in a different frame, then the existing motion compensated prediction methods are not capable of performing such predictions efficiently.

In recent years, many studies have proposed methods for improving the compression efficiency of the 360° images and video. Since each projection format includes unique sampling characteristics, the conducted studies usually develop projection-specific algorithms for improving the compression efficiency of each format. Since this work mainly focuses on the ERP projection format, most of the related studies in the sequel describe the compression approaches for this format.

The work in [83], proposes to apply the motion estimation process directly in the spherical domain instead of the 2D domain. To this end, multi-resolution decomposition of spherical images is used to increase the consistency of motion estimation. The work in [51], proposes a new motion model for the cubemap projection format that performs the motion estimation and compensation in the spherical coordinates system.

The deformable motion model in [72] proposes a new motion estimation method for ERP content. The method adapts the classical exhaustive block-matching algorithm in a way that it considers the object deformation in the spherical domain.

Rotational motion model, which is proposed in [88], performs the motion com-

pensation stage directly on the 3D spherical domain. To this end, the authors make use of a radial pattern search over the sphere. In their later work [91], they added new features to the rotational model such as motion vector refinement, enabling bi-prediction and multiple reference prediction methods for further improving the performance.

In [89], a method is studied where the motion compensation process is applied in the spherical domain for compensating the camera motion. The method assumes that the camera motion is available through the external devices such as accelerometer and gyroscope. In a later work [90], they proposed an extension of their method to modulate the motion vectors within a block for capturing the pixel-wise motion of samples in the block.

The described methods mainly perform the motion estimation and/or compensation stages in the spherical domain rather than the 2D coordinate. Thus, they are able to catch the non-linear motion of the projected content in a more accurate manner compared to the existing methods in the 2D codecs. However, applying these methods is costly in terms of number of computation (and hence processing times) as well as memory access and related buffering. Moreover, the adaptation of the motion estimation and compensation of existing codecs to be used with these methods requires significant changes in different blocks of the standard codecs. These issues make the deployment of the described methods in real-world applications impractical.

2.3 Omnidirectional Video Streaming

360° omnidirectional content is typically displayed on a Head-Mounted Display (HMD) device with limited field-of-views (FOVs). Current HMD technologies typically use FOVs of around 110°×90° [55, 56]. Thus, streaming the whole 360° scene at the highest resolution, quality, and frame rate, puts significant burden on the network bandwidth. The ideal streaming scenario for such content is to transmit only the content in the viewing orientation that the user is currently watching at each time instance (i.e. viewport area). On the other hand, in real world streaming cases, this would not be a feasible approach due to several factors, such as segment-based delivery and delays in end-to-end transmission especially in the case of fast head movement.

In practical streaming scenarios, the portion of the 360° video that is not in the viewing orientation of the user (i.e. non-viewport) is also transmitted to the user at a lower quality and/or resolution. In case of user's head movement or viewing orientation change, the lower quality non-viewport area is displayed for a short period of time until the next higher resolution and higher quality viewport is decoded, rendered and displayed. Another approach for resolving high bitrate issues of streaming high-quality 360° video is by making use of remote-rendering techniques [47, 87]. In these methods, the rendering is off-loaded from the user device to the cloud edge server. Based on the user's viewing orientation the corresponding FOV is rendered and displayed to the user. This thesis focuses only on the former described method where the rendering is done on the end user device and not over the edge server.

To implement the viewport-adaptive streaming scheme, two distinct streaming techniques have been proposed: projection-based viewport-adaptive streaming [5, 15, 23, 24, 44, 48, 49, 93] and tile-based viewport-adaptive streaming methods [3, 21, 27, 106]. Sections 2.3.1 and 2.3.2 provide more details regarding these two methods.

2.3.1 Projection-based Viewport-adaptive Streaming

In the projection-based viewport-adaptive streaming methods, unequal re-sampling approach is used in different parts of the 360° projection plane. The main concept of these methods is to assign a higher sampling density to the viewport area and a lower sampling density to the remaining parts of the content. In order to cover different parts of the 360° scene in a high resolution viewport, multiple versions of the content is generated in a way that each of them consists of different viewport region. Then, these versions of the projected content is encoded and stored in the server. Based on the user's viewing orientation, the corresponding version of the content is selected and transmitted to the user. In the following, some examples of this method are reviewed.

The pyramid projection format is one of the early approaches of projection-based VAS category that was developed by Facebook [49]. In this method, illustrated in Figure 2.6, the 360° content is projected onto a pyramid format in such a way that the base of the pyramid comprises the viewport and the remaining non-viewport region are projected into the other sides of the pyramid. As a result, the viewport (base of the pyramid) includes a higher sampling density than the non-viewport parts in

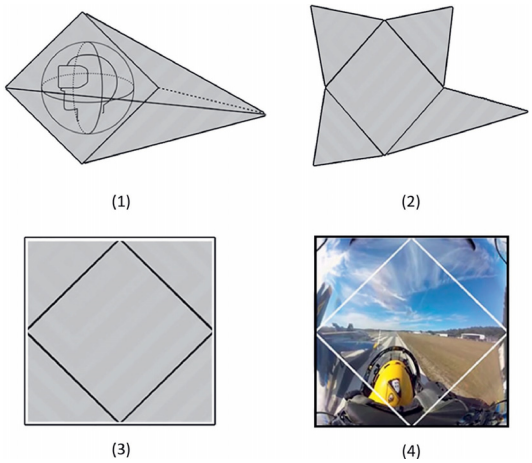


Figure 2.6 An illustration of the pyramid projection [49]

other pyramid sides. For using the current 2D video codecs, the base and other sides of the pyramid are unfolded into a rectangular plane.

The pyramid-based projection approach is further developed by other researchers in order to improve the coding efficiency. For example in [44], a different arrangement for the non-viewport sides of the pyramid is considered. The authors used rotated versions of the non-viewport sides in the frame packing stage in order to avoid the high frequency information that sharp edges of the pyramid creates in the coding stage. Furthermore, the authors rotated the base of the pyramid in such a way that it has vertical and horizontal edges in the projection plane rather than diagonal ones. The truncated square pyramid (TSP) is another variation of this projection format that is proposed in [5, 44], where the apex of the pyramid is clipped and the arrangement of the of pyramid faces are changed in the rectangular grid.

The cubemap offset projection is another projection type that is introduced by researchers at Facebook for streaming VR video [48]. This method projects the 360° video to the cubemap format with a camera offset toward the back of the cube projection. By doing so, more pixels are allocated to the viewport area than other parts of the scene. This results in decreasing the overall resolution while preserving the high resolution in the viewport.

The asymmetric circular (ASC) projection is proposed in [93], where the authors use equal area projection in the viewport area to have a high sampling density and a decreasing sampling density in non-viewport area. The equal area projection [75,

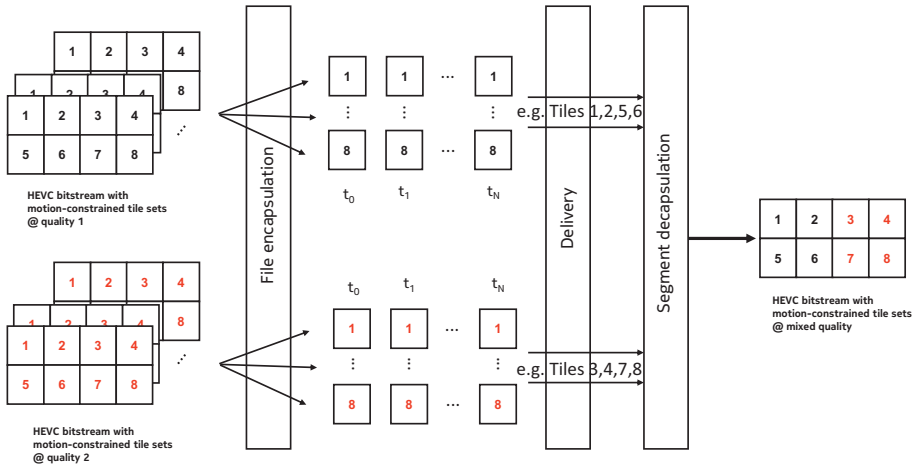


Figure 2.7 Single-layer viewport-adaptive streaming with MCTS method [32]

95] used for viewport follows the same ratio in the 2D projection plane as its corresponding spherical domain. The non-viewport region in ASC method is gradually down-sampled. In other words, the sampling density decreases as one moves further away from the center of the viewport.

The projection-based VAS methods typically require to encode and store a high number of versions of the same content in the server. In recent works [14, 15], there is an effort to reduce the number of coded and stored versions. For example, the work in [15] uses a certain number of Quality Emphasis Centers (QECs) where each QEC corresponds to a 360° stream with certain viewport area coverage. To this end, the authors use cubemap projections in which the front face of CMP is located at the center of a QEC and other faces were encoded at a lower bitrate compared to the viewport/front face. As a result, 5 to 7 versions of the content are sufficient for viewport-adaptive streaming practices. However, due to the concerns regarding the storage requirements of projection-based VAS schemes, some studies suggest to use tile-based methods instead [50, 104].

2.3.2 Tile-based Viewport-adaptive Streaming

Tile-based viewport-adaptive streaming methods have been widely investigated for delivering omnidirectional video in recent years [18, 27, 37, 64]. The Motion-constraint Tile Set (MCTS) technique, that was first introduced in [33], is utilized for VAS prac-

tices. According to MCTS technique, the picture can be split into multiple tiles. Each tile is encoded in such a way that all the predictions (i.e. the spatial and temporal) along with the loop filtering operations are restricted to the tile boundaries. Consequently, because of these restrictions, each tile can be decoded independently. Independent decoding benefit of MCTS makes it very useful tool for delivering 360° video in VAS applications.

Various studies conducted over the past years have shown the advantages of tile-based approaches for streaming omnidirectional content. For example, the method in [3] makes use of tiling technique and adapting the quality of each region of the content based on the viewing likelihood for streaming the omnidirectional video.

A tile-based method was studied in [19, 21] for streaming the 360° video in such a way that the generated bitstream is decodable using a single hardware decoder. Furthermore, the authors proposed a Generated Reference Picture (GRP) technique to reduce the bitrate in region-of-interest (ROI) switching points. Later, they extended the GRP concept to be used in the extensions of the HEVC standard. They developed Multi-view GRP (MGRP) [73] for multi-view extension of HEVC and Multi-layer GRP (ML-GRP) [20] for the scalable extension.

The work in [106] divides the 360° video into multiple sets of tiles. Then, the tiled video is encoded in two versions 1) high-resolution and 2) low-resolution. For streaming, a set of tiles, corresponding to the viewport area, is selected from the high-resolution version of the encoded video. The remaining parts of the 360° video, i.e. the non-viewport area, is selected from the low-resolution version of the content. The selected high-resolution tiles along with the low-resolution tiles are transmitted to the viewer. Figure 2.7, illustrates the described MCTS-based VAS of [106] with 4×2 tiling arrangement. Similar approach as [106] is considered for streaming stereoscopic 360° video in [108]. The proposed method interleaves the two 360° views in a single video content instead of coding them in a multi-layer or stereoscopic fashion. The inter-view prediction functionality in Intra Random Access Point (IRAP) pictures is used for reducing the bitrate in viewport switching points in the second view content. Another tile-based VAS method is studied in [58] which is a similar approach followed in [**P4**] where the scalable HEVC is used for streaming the 360° video. The method encodes the low-quality CMP content in the base-layer (BL) and the high-quality version of the content in the enhancement-layer (EL). The entire BL content along with the portion of EL, that is corresponding to the viewing ori-

entation, are transmitted to the user. Moreover, they configure the BL to use longer IRAP intervals in order to avoid frequent intra-coded pictures in low-quality video. The works in [105, 107] propose mixed-resolution packing schemes for viewport-adaptive streaming of omnidirectional video. The authors proposed two distinct packing methods for 6K and 8K ERP content in such a way that these packing methods enable 6K and 8K effective viewport resolutions in the content, whereas they use lower resolutions in non-viewport areas. Furthermore, the viewport and non-viewport areas are encoded using MCTS technique.

The Omnidirectional MediA Format (OMAF) [13, 34, 66] is a newly developed standard for 360° images and video by the Moving Picture Experts Group (MPEG) [12]. The first version of this standard, referred to as OMAF v1, was finalized in October 2017 and several compatible implementations for that is already publicly available [59, 63, 65] . The second version of OMAF [16], OMAF v2, is completed during writing this thesis in October 2020. OMAF v2 provides new features such as overlays, multiple viewpoints as well as improvements for viewport-dependent delivery of 360° content in addition to the existing features of version 1 of the standard. There are already some implementations available which support the new features of OMAF v2 standard [38, 77]. The OMAF v1 standard specifies the file and delivery formats for omnidirectional content. This standard makes use of other existing MPEG standards and specifies new extensions to the ISO Base Media File Format (ISOBMFF) [40] and the Dynamic Adaptive Streaming over HTTP (DASH) [76]. Readers interested in more details regarding OMAF standard may refer to e.g. [34, 66].

OMAF v1 enables adaptive streaming of VR content under certain restrictions. For example the VAS scheme must be compliant with the following:

- Frequent viewport switching
- Single-layer decoding constraint
- Decoding with single decoder instance
- 4K decoding constraint

The above-mentioned VAS schemes are able to reduce the streaming bitrate of omnidirectional video significantly. Among these, the multi-layer approaches [20, 58, 73] are not aligned with the single-layer decoding constraint of OMAF. On the

other hand, even though the single-layer MCTS-based methods [106, 108] are compliant with this standard, these methods still require frequent IRAP pictures for enabling seamless viewport switching operations. Consequently, high bitrates of intra-coded pictures make these schemes sub-optimal for 360° video streaming practices. This thesis investigates and proposes a novel VAS scheme in order to tackle the mentioned issues while reducing the streaming bitrate of omnidirectional video compared to state-of-the-art single-layer tile based streaming.

3 ENCODING SOLUTIONS FOR OMNIDIRECTIONAL VIDEO

This chapter proposes methods for improving the compression efficiency of the omnidirectional video format over High Efficiency Video Coding (HEVC) and Versatile Video Coding (VVC) standards. We particularly focus on motion estimation and compensation problems in equirectangular projection (ERP) and provide solutions for improving the compression performance for this 360° video format. The reason for focusing on ERP is that content deformation in this projection type is more severe than in the cubemap projection (CMP), hence, the existing video coding standards are not optimised to efficiently model such non-linear behavior of the motion in 360° video formats. This chapter studies three methods for improving the inter prediction in ERP which target to address the first research question in this thesis concerning the development of more efficient coding tools to improve the compression performance of current codecs in order to deal with immersive virtual reality video.

The chapter is organized as following. First, the performance evaluation methodology for 360° content is described in Section 3.1. In Section 3.2, an ERP-specific motion vector scaling method is proposed for improving the motion vector prediction process. An adaptive motion vector prediction approach is described in Section 3.3 where a linear regression scheme is used for predicting the motion information of a coding block based on the motion information of its neighboring blocks. The linear regression-based method is extended in Section 3.4 in order to provide the motion information of a block in 4×4 and 8×8 sub-block levels in order to achieve an enhanced granularity of motion information.

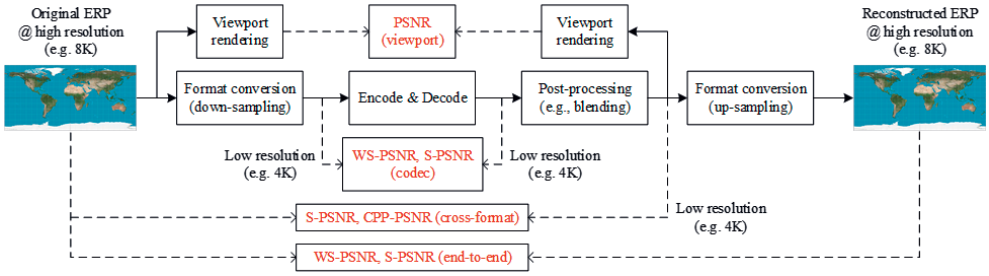


Figure 3.1 360° video common test procedure [31]

3.1 Compression Performance Evaluation Methodology

For evaluating the compression performance of 360° images and video, the Joint Video Experts Team (JVET) has developed a testing procedure which provides a performance evaluation testbed for such content [4, 31]. The overall procedure for JVET 360° common test conditions is shown in Figure 3.1. Furthermore, JVET community has developed a toolbox known as 360Lib for the testing procedure, where all the components shown in the figure are integrated in the software [43, 97, 99].

According to the testing methodology, the high-fidelity ERP video (i.e. 8K resolution) is considered to be the original format, whereas the coding domain format or resolution may differ from that of the source. For example, if the coding domain is considered to be ERP, then the source high-fidelity ERP is down-sampled to a lower resolution (i.e. 4K) in the coding domain and then used as input to the codec. Finally, at the decoder side, the decoded low-resolution version is up-sampled to the original high-fidelity ERP, as illustrated in Figure 3.1. Similarly, in case the coding domain is not ERP, then the source high-fidelity ERP is converted to a lower resolution coding domain format e.g. 4K resolution CMP and then it is used in the encoding chain. The decoded coding domain format (e.g. CMP) is then up-sampled to the 8K resolution ERP format. The reason for such resolution conversion is to remove the bias from the ERP format when comparing the performance of different projection formats.

For the quality measurements of the above procedure, several objective metrics have been developed and included in the 360Lib software [43]. These metrics are

Table 3.1 Quality metrics supported in 360Lib software [98]

Quality metric	Description of the metric
ERP-PSNR	Conventional PSNR calculation with equal weight for all samples.
WS-PSNR	In PSNR calculation, the distortion at each sample is weighted based on its latitude position on the sphere.
S-PSNR-NN	Calculates the PSNR based on a set of points uniformly sampled on the sphere. Nearest neighbor (NN) rounding is used for finding the sample value at the corresponding position on the 2D plane.
S-PSNR-I	Calculates the PSNR based on a set of points uniformly sampled on the sphere. Bicubic interpolation is used for finding the sample value at the corresponding position on the 2D plane.
CPP-PSNR	Calculates the PSNR of points by projecting them into Crasters Parabolic Projection (CPP) domain.

ERP-PSNR, Weighted to Spherically uniform PSNR (WS-PSNR) [81], Spherical PSNR (S-PSNR) [101] and Crasters Parabolic Projection PSNR (CPP-PSNR) [103]. Table 3.1 provides a summary of these quality metrics that are integrated in the 360Lib software.

The compression performance of the studied methods in this chapter is analyzed by using the well-known Bjontegaard Delta Bitrate (BD-Rate) criterion [6]. According to BD-Rate method, the negative values illustrate how much the bitrate is decreased over a set of different quality levels while positive values show how much the bitrate is increased for the tested quality levels. Finally, for conducting the experiments, the main profile random access (RA) configuration of the JVET common test condition [7, 8, 31] is used. The quantization parameters (QPs) of 22, 27, 32 and 37 are used in the simulations.

Finally, for conducting the experiments, the JVET 360° test sequences [4, 31, 68] were used. These sequences are in 8K ERP resolutions with different characteristics of motion e.g., high-motion, low-motion, stationary behavior.

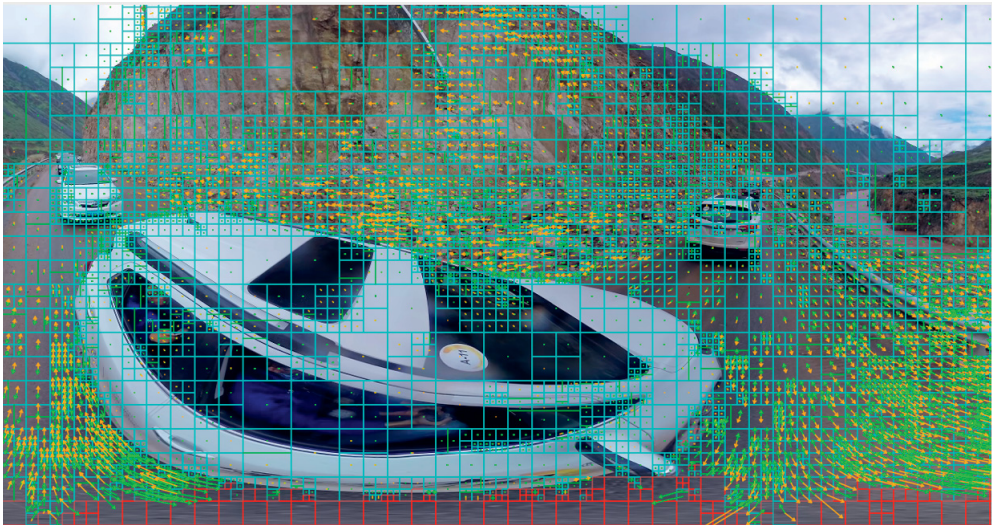


Figure 3.2 An example of motion vector behavior in ERP [**P1**]

3.2 Geometry-based Motion Vector Scaling

This section provides a summary of the motion vector scaling method that is proposed in [**P1**]. The method proposes a novel motion vector scaling scheme for improving the motion vector prediction performance of 360° ERP content when using the existing video compression standards (e.g. HEVC and VVC).

The layout of this section is as follows. The motion vector scaling method is described in Section 3.2.1. The simulation results of the proposed method is provided in 3.2.2.

3.2.1 Algorithm Description

In the existing video coding standards, a motion vector prediction process is used for coding the motion information of the block. The Advanced Motion Vector Prediction (AMVP) and Merge tools in the HEVC [79] and VVC [9] standards are two examples of such schemes. In AMVP and Merge modes, a list of MVP candidates are generated based on the motion information of the neighboring blocks. The motion vectors of the current block are predicted based on the generated MVP list and the index of the MVP candidate from the list is signaled into the bitstream. In case of

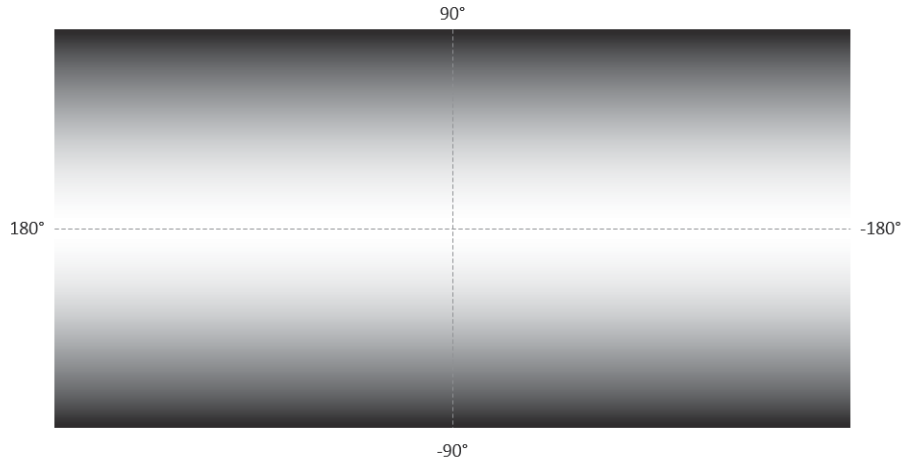


Figure 3.3 Sampling weight map in ERP [81]

AMVP, the motion vector difference (MVD) between the MV of current block and the best performing prediction candidate from the list is signalled into the bitstream as well.

As mentioned before, the 360° ERP format includes content deformation and stretching caused by over-sampling characteristics of the projection plane. Such sampling characteristics create a non-linear motion behavior through the video sequence. Thus, the motion vectors of a block in the frame may vary substantially in terms of magnitude and direction, compared to its neighboring blocks. Figure 3.2 illustrates an example of a behaviour in *DrivingInCountry* video from JVET sequences [4, 31]. As can be observed, the content includes object deformation, particularly in areas near the south pole where the motion is higher. Consequently, the motion vectors required for motion estimation are large and their directions change when approaching the polar regions. Furthermore, when the MVs are significantly large, the inter prediction method fails to perform well. In such cases, those blocks are coded in intra prediction mode that typically require a higher bitrate than conventional inter prediction modes. Such scenarios can be seen in the bottom blocks of Figure 3.2, where the red boundary blocks are coded in intra prediction mode.

In order to alleviate the described issue, a motion vector scaling method is introduced in [**P1**]. According to this approach, the neighboring motion vectors that are used as predictors for motion information of the current block are scaled based

on their location in the ERP plane. The proposed scheme, scales the neighboring MVs up or down in order to make them suitable for the current block. The scaling method is motivated by the weighted quality metric (i.e. WS-PSNR [81]) for 360° content, where for PSNR calculation, the decoded pixels are weighted based on their latitude position in the spherical domain. The weighting considers the fact that in the spherical domain, the sampling density decreases latitude-wise toward the poles. Hence, samples closer to polar areas assume smaller weights than the ones closer to the equator areas in the projection plane. Such behavior is shown in Figure 3.3. This weight distribution can be represented as a cosine-like function where the weight value has its maximum magnitude (i.e. equal to 1) at the equator regions and its minimum value (i.e. close to zero) at the polar areas. A similar approach is used in this work for scaling the neighboring motion vectors based on their latitude-wise location in ERP plane. However, since the motion vectors are calculated in a block level not pixel level, weight derivation is used for obtaining one weight value for each block, relative to the center location of that block. To this end, the weight derivation function is defined as:

$$W[x,y] = cos(\frac{y - \frac{h}{2} + 0.5}{h} \times \pi) \qquad Where: \begin{cases} 0 \leq x < w \\ 0 \leq y < h \end{cases} \tag{3.1}$$

In this equation, W is the calculated weight value for the block at location (x, y) where x and y are the horizontal and vertical components of the center location of the block, respectively; h and w are the height and width of ERP image. For scaling the motion vectors, a parameter called Scaling Factor (SF) is defined as:

$$SF = W_{max} - (W_C - W_N) \tag{3.2}$$

where W_C and W_N are the weight values for the current and the neighbor block, respectively, which are calculated based on Eq. (3.1). W_{max} represents the maximum weight value which is, based on the cosine weight function, equal to 1. Finally, the scaling factor is applied to the neighboring block's motion vector, which includes

both horizontal and vertical components of the motion information.

$$\vec{MV}_{scaled} = round(\vec{MV} \times SF) \tag{3.3}$$

Using the above procedure, the neighboring motion vectors that are used for predicting the motion information of the current block are scaled up or down based on the block's latitude location in ERP.

The scaling method in this work is applied to prediction candidates of both AMVP and Merge coding tools. The scaled MVs provide a more uniform motion vector predictors for the current block, hence, they reduce the bitrate and lead to more efficient coding.

3.2.2 Experimental Results

The described MV scaling algorithm is implemented over VTM-1.0 test model [85] of VVC standard [9]. The experiments are conducted according to the common test conditions with Random Access (RA) configuration [8]. The testset is divided into two categories as follows:

1. High motion: sequences include global motion (e.g. camera motion) and/or high object motion particularly in the polar areas.
2. Low motion: sequences include low motion, or motion only at the equator (i.e. linear motion)

Table 3.2 shows the performance of the motion vector scaling method using WS-PSNR metric. The results for high motion category show that the proposed scaling method is able to provide high BD-Rate improvements compared to the anchor. The average gain of 1.0% was achieved for this category. Particularly, when the content has motion in the polar areas, the proposed method provides the highest gain. This can be observed in *ChairliftRide*, *DrivingInCountry* and *Glacier* sequences where a bitrate reduction of around 2.0% was achieved.

On the other hand, the proposed method did not provide any gains or losses for the low motion category, where the average BD-Rate impact was 0.0%. This was an expected performance behavior since the MV scaling method was designed in a way that improves the MV prediction when a non-linear motion behavior is present in the content.

Table 3.2 BD-Rate (%) performance of the motion vector scaling method

Category	Sequence	Y	U	V
High Motion	ChairliftRide	-2.1%	-2.0%	-1.8%
	Skateboard	-0.4%	-0.4%	-0.1%
	Balboa	-0.7%	-0.4%	-0.8%
	BranCastle2	-0.5%	-0.3%	-0.4%
	Landing2	-0.4%	-0.1%	-0.2%
	DrivingInCountry	-1.7%	-1.7%	-1.5%
	Bicyclist	-0.7%	-0.3%	-0.4%
	Glacier	-2.0%	-1.5%	-1.7%
	Building	-0.6%	-0.3%	-0.5%
Low Motion	Gaslamp	0.0%	0.0%	0.0%
	Trolley	0.0%	0.0%	0.0%
	KiteFlite	0.0%	0.0%	0.0%
	Harbor	0.0%	0.0%	0.1%
	Broadway	0.0%	0.0%	0.1%
	AerialCity	-0.1%	0.0%	0.0%
	DrivingInCity	0.0%	0.2%	0.0%
	Paramotor	-0.1%	-0.1%	-0.2%
High Motion Average		-1.0%	-0.8%	-0.8%
Low Motion Average		0.0%	0.0%	0.0%
Overall		**-0.6%**	**-0.4%**	**-0.4%**

Table 3.3 Average runtimes (%) of the MV scaling method compared to VTM-1.0 [**P1**]

Category	High Motion	Low Motion
Encoding Runtime	99.1%	100.6%
Decoding Runtime	101.5%	101.3%

Table 3.3 shows the average encoding and decoding runtimes of the proposed method in both categories of testsets compared to reference VTM-1.0. As can be observed from the table, the proposed method does not have any impact on the encoding and decoding runtimes, as the applied scaling process is local and simple. Thus, the complexity overhead of the whole procedure is negligible.

3.3 Motion Vector Prediction with Linear Regression Model

This section introduces an adaptive method for predicting the motion vectors of a block based on its neighboring motion information [**P2**]. A regression-based approach is followed to predict the motion vectors by modeling the motion behavior of the coding block in an adaptive way according to the motion behavior of the neighbors and their locations. The proposed MV prediction scheme is described in Section 3.3.1 and the results of the experiments are presented in Section 3.3.2.

3.3.1 Algorithm Description

The motion vector scaling scheme presented in Section 3.2 seeks uniform motion vector predictors for the current block based on the latitude-wise location of the block in ERP format. Here, we propose an alternative method that is not restricted by the geometry of the projection plane. Unlike motion vector scaling, we aim to provide proper MV predictors for the block with an adaptive approach by modeling the motion vectors of the current block based on the motion information and (x, y) locations of the spatial neighboring blocks with a linear model. For modeling purposes, the linear regression with the Mean Square Error (MSE) minimization is used.

The 6-parameter motion model of Eq. (3.4) is used for adaptive MV prediction:

$$\begin{bmatrix} MV_x \\ MV_y \end{bmatrix} = \begin{bmatrix} a_{\mathrm{x0}} & a_{\mathrm{x1}} & a_{\mathrm{x2}} \\ a_{\mathrm{y0}} & a_{\mathrm{y1}} & a_{\mathrm{y2}} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.4}$$

where MV_x and MV_y are the horizontal and vertical MVs of a block, respectively. The adaptive prediction model calculates the motion vectors according to the center

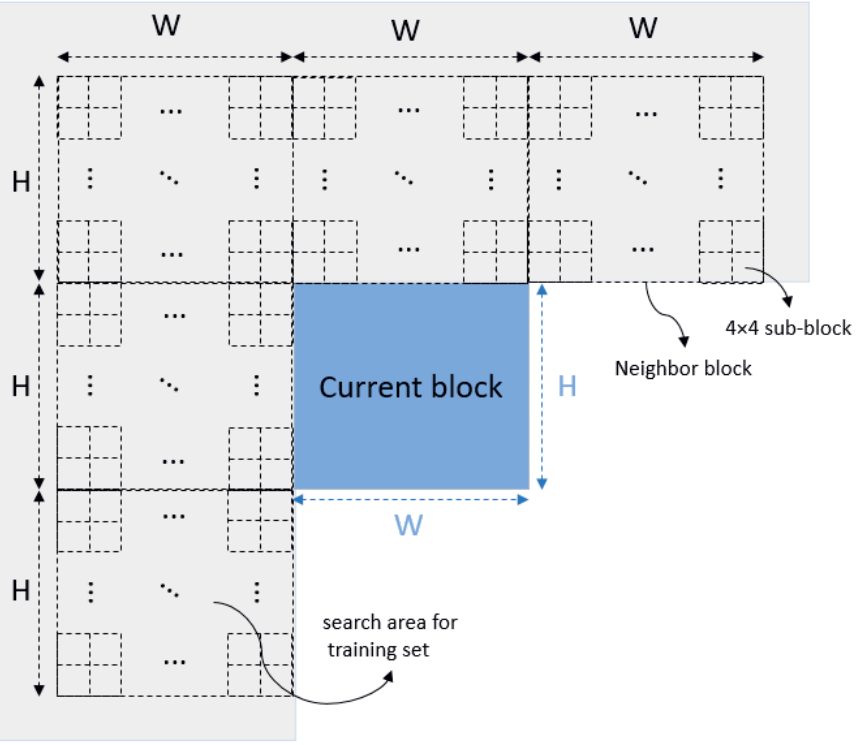


Figure 3.4 Neighboring sub-block MVs that are used for training the motion model [P2]

locations of each block, hence, x and y in above formula represent the horizontal and vertical center locations of the block, respectively. Finally, the remaining 6 parameters in Eq. (3.4), namely a_{x0}, ..., a_{y2}, referred to as model parameters hereafter, are calculated based on the motion information of the neighboring blocks with linear regression method. For model parameter calculation, motion information is collected from the spatially neighboring blocks. Figure 3.4 shows the search range used for the neighboring motion information collection. The search range depends on the size of the current block. As can be seen from the figure, the search area is limited to twice of corresponding height or width of the block, depending on the search direction, e.g., the search range from the top of the block is limited to twice of the block's height.

The motion vectors in many video coding standards such as HEVC and VVC are stored in 4×4 sub-block accuracy. Thus, in the training data collection process from the neighbors, the MVs are collected in 4×4 sub-blocks and their corresponding (x,

y) center locations. Using the neighboring MVs in 4×4 sub-blocks is beneficial since it can take into account the size of the neighboring blocks in the parameter estimation. In other words, the number of 4×4 MVs from a block that are used in the training is a proportion to the size of that block.

The collected neighboring 4×4 MVs and their locations are then used as inputs to the linear regression process. In this process, it can be assumed that the motion model in top and/or left neighbors is also defined as in Eq. (3.4) and the parameters of the model can be solved by minimizing the MSE. Finally, the motion vector of the block is calculated based on the estimated model parameters, motion model of Eq. (3.4) as well as the center location of the current block.

The calculated adaptive motion vector is included in the AMVP and/or Merge coding tools as an additional MV predictor. Moreover, the new predictor is prioritized in AMVP and Merge lists to the existing ones, meaning that it is inserted to the list as the first candidate.

3.3.2 Experimental Results

The proposed adaptive motion vector prediction method is implemented over the HEVC reference software HM-16.16 [35]. The experiments were conducted based on the common test condition (CTC) with RA configuration [7].

Table 3.4 shows the performance of the method when the adaptive MV candidate is added to AMVP, Merge or to both lists. As can be observed from the results, the proposed adaptive MV prediction technique improves the prediction performance when it is added to either or both lists where the bitrate reductions of around 1.0% on average is achieved. The performance is slightly higher when the new predictor is added to both AMVP and Merge lists. Another observation is that the proposed method provides higher gains in sequences with higher motion than those with stationary behavior. For example, the bitrate reductions in *DrivingInCountry*, *ChairliftRide*, *Bicyclist* and *Glacier* are significantly high due to the high motion particularly in the polar areas. On the other hand, the proposed method does not bring any significant improvements to somewhat stationary content.

Table 3.4 BD-Rate (%) results of the adaptive MV prediction method for luma component [**P2**]

Sequence	AMVP	Merge	Both
Trolley	0.01%	-0.01%	0.04%
AerialCity	-0.07%	-0.10%	-0.02%
DrivingInCity	0.05%	-0.17%	0.14%
DrivingInCountry	-2.04%	-1.00%	-2.24%
ChairliftRide	-1.41%	-1.03%	-1.80%
Skateboard	-0.43%	-0.43%	-0.42%
Balboa	-0.86%	-1.63%	-1.80%
BranCastle	-1.11%	-1.52%	-1.86%
Landing	-1.60%	-1.28%	-1.87%
Broadway	-0.53%	-1.68%	-1.67%
Bicyclist	-1.30%	-0.50%	-1.10%
Glacier	-2.40%	-1.00%	-2.20%
Paramotor	0.00%	-0.20%	-0.10%
Overall	**-0.90%**	**-0.81%**	**-1.15%**

3.4 Regression-based Motion Vector Field

The adaptive motion vector prediction of Section 3.3 derives a motion vector for the entire block. However, a single MV may not be able to represent the motion of the whole block especially if the block size is large. In this case, we propose an alternative solution that follows a similar approach as in adaptive MV prediction, except that the MVs of the block are derived at sub-block levels with linear regression, instead of block level. Hence, the proposed method is referred to as Regression-based Motion Vector Field (RMVF) and appeared in [**P3**] and JVET contributions [25] and [26]. Moreover, the model parameter derivation process is simplified in order to achieve lower complexities.

3.4.1 Algorithm Description

As mentioned above, a single MV is not an efficient approach for predicting the motion of a block. Particularly, in cases when the characteristics of the motion is

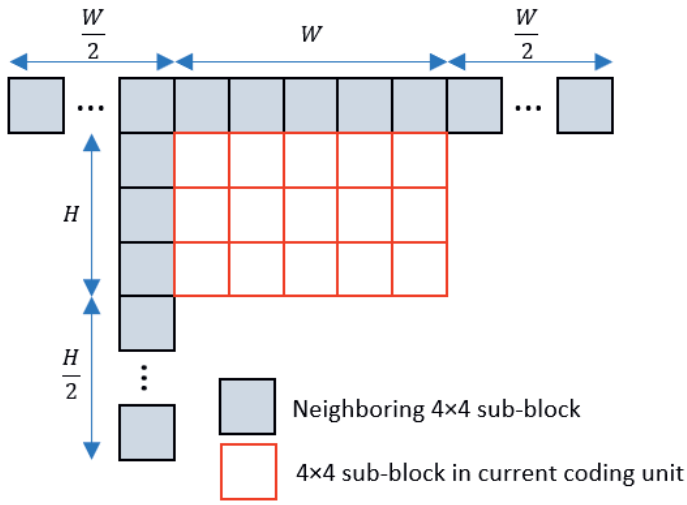


Figure 3.5 Neighboring sub-block MVs that are used for training the RMVF model [**P3**]

not linear, conventional motion estimation and compensation approaches achieve sub-optimal results.

Such non-linear behavior in the motion may be due to camera rotation and zooming, content deformations caused by the capturing device, or sampling of the projection plane in case of 360° content such as ERP. In such scenarios, the adaptive MV prediction of Section 3.3 does not provide the optimal results, especially in the presence of advanced sub-block based coding tools such as Affine Motion Compensation (AMC) and Alternative Temporal Motion Vector Prediction (ATMVP) which are used in VVC standard [10]. These tools provide finer motion granularity and hence are able to capture the non-linear motion behavior in a more efficient way than conventional methods. Thus, in this section we investigate and propose an adaptive MV prediction scheme at sub-block level fashion.

The Regression-based Motion Vector Field (RMVF) method uses a similar motion model as the adaptive MV prediction approach. Let us assume that the 6-parameter motion model of Eq. (3.5) is used for this purpose.

$$\begin{bmatrix} MV_x \\ MV_y \end{bmatrix} = \begin{bmatrix} a_{x0} & a_{x1} & a_{x2} \\ a_{y0} & a_{y1} & a_{y2} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.5}$$

where MV_x and MV_y are the horizontal and vertical components of each sub-block, respectively, and x and y are the center locations of the corresponding sub-block. The remaining 6 parameters in the equation are the motion model parameter (referred to as RMVF-parameters hereafter) are calculated based on the neighboring motion information and their locations in a similar way to the adaptive MV prediction method.

To obtain the RMVF-parameters, we proceed in a similar way to the previous method and use the neighboring 4×4 sub-block MVs and locations. However, the number of collected neighboring sub-blocks is reduced significantly. Figure 3.5 illustrates the neighboring sub-block MV range that is considered for RMVF-parameter calculation. As shown in the figure, only one row and one column of the neighboring 4×4 sub-blocks are used. Finally, the sub-block MVs for the current block are derived according to the motion model in Eq. (3.5) using the RMVF-parameters and the central sub-block locations inside the block.

The proposed RMVF method is implemented as a separate Merge mode in the VTM-2.0 test model [86] of VVC standard [10]. Two configurations for sub-block sizes are implemented and tested as specified below:

— Motion vectors prediction and motion compensation in 4×4 sub-blocks,

— Motion vectors prediction and motion compensation in 8×8 sub-blocks.

For the motion compensation (MC) part in a 4×4 sub-block, the default sub-block motion compensation of VTM-2.0, which operates in 4×4 sizes is used. In the second case, MC in a 8×8 sub-block size, the sub-block motion compensation function of VTM-2.0 is modified in such a way that for RMVF blocks it uses 8×8 sub-block sizes for MC. Apart from these modifications, a coding unit (CU) level RMVF flag in Merge mode is encoded and signaled into the bitstream in order to indicate the use of RMVF mode for that block.

Table 3.5 BD-Rate (%) results of RMVF method over VTM-2.0 [**P3**]

Class	Sequence	4 × 4 sub-blocks			8 × 8 sub-blocks		
		Y	U	V	Y	U	V
Class S1	Skateboard	-0.59%	-1.30%	-1.34%	-0.47%	-1.12%	-1.18%
	ChairliftRide	-1.52%	-1.68%	-1.52%	-1.28%	-1.45%	-1.40%
	KiteFlite360	-0.10%	-0.14%	-0.24%	-0.08%	-0.13%	-0.27%
	Harbor360	-0.03%	-0.08%	-0.06%	-0.04%	-0.10%	-0.16%
	Trolley	-0.06%	-0.05%	-0.04%	-0.05%	-0.10%	-0.05%
	Gaslamp	-0.01%	0.02%	0.00%	0.01%	0.02%	0.02%
Class S2	Balboa	-1.58%	-1.76%	-1.46%	-1.30%	-1.33%	-1.17%
	Broadway	-1.37%	-1.26%	-1.37%	-1.19%	-1.14%	-1.10%
	Landing2	-0.54%	-0.56%	-0.85%	-0.36%	-0.48%	-0.63%
	BranCastle2	-0.70%	-0.47%	-0.48%	-0.58%	-0.52%	-0.58%
Overall Class S1		-0.39%	-0.54%	-0.53%	-0.32%	-0.48%	-0.51%
Overall Class S2		-1.05%	-1.01%	-1.04%	-0.86%	-0.87%	-0.87%
Overall		**-0.65%**	**-0.73%**	**-0.74%**	**-0.53%**	**-0.63%**	**-0.65%**
Encoding Time		107%			103%		
Decoding Time		105%			102%		

3.4.2 Experimental Results

As mentioned in Section 3.4.1, the proposed RMVF method is implemented on top of VTM-2.0 test model [86] of VVC standard [10]. In the experiments, the common test condition for 360° content [31] with RA configuration is used.

Table 3.5 presents the performance of the RMVF method in 4 × 4 and 8 × 8 sub-block sizes. As can be observed from the table, the RMVF method outperforms VTM-2.0 on average by 0.65% and 0.53% with 4 × 4 and 8 × 8 sub-block sizes, respectively.

Another observation from the results is that, the RMVF method performs better when the 4 × 4 sub-block sizes are used. However, the encoder and decoder runtimes are also higher compared to the 8 × 8 sub-block MC case.

In both sub-block sizes, the RMVF method provides better bitrate reductions when the content includes non-linear and/or global motion, especially in polar regions of ERP content. Such performance behavior can be observed in *ChairliftRide*,

Balboa and *Broadway* sequences, where the RMVF method was able to provide more than 1.0% bitrate reduction in both 4×4 and 8×8 sub-block sizes.

In terms of complexity analysis, the proposed RMVF method does not impose significant overheads to the codec. As it can be observed from Table 3.5, RMVF introduces 7% and 3% encoding runtime increases for 4×4 and 8×8 sub-block sizes, respectively. Moreover, the proposed method increases the decoding runtimes by 5% and 2% for 4×4 and 8×8 sub-blocks, respectively. Considering the overall compression efficiency of this method, the runtime increases are in reasonable ranges. Particularly for the 8×8 sub-block size case the compression efficiency versus complexity trade-off is really interesting.

Among the proposed methods in this chapter, the RMVF method is considered a more versatile solution for dealing with the non-linear motion of 360° video. Compared to the MV scaling method of Section 3.2, RMVF is not limited to the geometry of the projected content and is able to estimate and compensate such motion regardless of the projection format. Furthermore, results in [**P3**] demonstrate that RMVF method is also beneficial in conventional 2D content with non-linear motion. When compared to the method in Section 3.3, RMVF provides motion estimation in finer granularity, thus, is able to compensate the non-linear motion in a more efficient way than full block approach.

4 STREAMING SOLUTIONS FOR OMNIDIRECTIONAL VIDEO

This chapter studies novel solutions for improving 360° virtual reality content streaming. The chapter summarizes the works that have been conducted in [**P4**] and [**P5**].

Unlike traditional 2D video that has limited FOV, omnidirectional content covers the 360° FOV around the capturing device in order to be used in immersive VR applications. In traditional streaming scenarios, the whole 360° video is encoded into a single bitstream and transmitted to the receiver/user. At the user side, the full 360° FOV omnidirectional video is decoded. The portion of the 360° video that is in the viewing orientation of the user, i.e. around 110°×90° [55, 56], is rendered and displayed. Since, only a limited FOV of the content is displayed at each time instance, transmitting the whole content in the highest resolution and quality requires high bitrate and consequently consumes high network bandwidth.

In this chapter, three tile-based viewport-adaptive streaming (VAS) methods are proposed for improving the streaming performance of omnidirectional video. Section 4.1 describes the performance evaluation methodology, used in this chapter for assessing the performance of VAS schemes. In Section 4.2, two methods are proposed for improving the streaming performance of 360° video compared to the state-of-the-art tile-based scheme. Finally, a novel Shared Coded Picture (SCP) method is proposed in Section 4.3 that improves the streaming bitrate of omnidirectional video in viewport switching points. The proposed method is fully compliant with the OMAF standard.

4.1 Streaming Performance Evaluation Methodology

The testing methodology that is proposed in MPEG [45, 46] is used in this chapter for evaluating the quality of viewing experience for the described streaming meth-

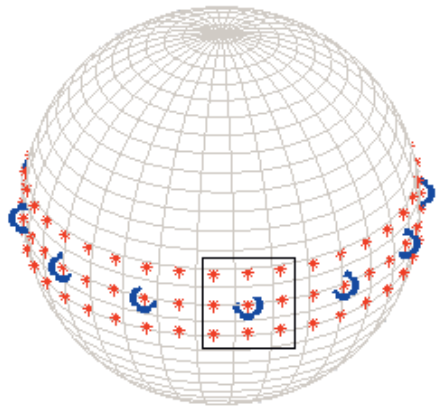


Figure 4.1 Illustrates the location of QAVs in the quality evaluation methodology [**P5**]

ods. This evaluation framework uses a pre-defined set of viewing orientations, called Quality Assessment Views (QAVs), over the high-quality version of the 360° stream for measuring the objective quality. For this purpose, a set of viewport-based streams of high-quality 360° content is assumed in the equator area, where each viewport-based stream covers 110°×110° FOV. Overall, 12 viewport-based streams are considered in the equator with inter-distance of 30°. In order to measure the viewing quality, 24 uniformly distributed QAVs are defined along the equator. Moreover, 12 QAVs are defined in ±15° latitudes. Thus, in total, 72 QAVs are defined. Figure 4.1 provides an illustration of the selected QAVs. In this figure, the center of viewport streams are shown with blue marks and the center of QAVs with red marks.

For the selected QAVs, a viewport of 90°×90° FOV is rendered with a rectilinear projection. The viewport PSNR is measured in the rendered 90°×90° FOV. The final quality of the stream is calculated as the average of 72 rendered viewport PSNRs. For the bitrate performance evaluation, the Bjontegaard Delta Rate (BD-Rate) method [6] is used.

4.2 Tile-based Viewport-adaptive Streaming

This section provides a summary of the viewport-adaptive streaming methods that have been studied in [**P4**] and the MPEG contributions in [28, 29].

In Section 2.3, it was explained that tile-based VAS schemes provide high stream-

ing bitrate reductions compared to transmitting the whole 360° high resolution and quality content. However, as it was also mentioned, the state-of-the-art tile-based VAS (i.e. MCTS-based) method uses intra-coded pictures in the switching points, i.e. Intra Random Access Points (IRAPs), in order to have seamless viewport switching operations. However, IRAP pictures consume high bitrates compared to inter-coded pictures. We therefore aim to reduce the large bitrate of viewport switching points, while preserving its frequent viewport switching capability.

4.2.1 SHVC Region-of-Interest VAS

In order to reduce the bitrate of IRAP pictures in the viewport switching points, the multi-layer SHVC with region-of-interest (ROI) coding approach is used for viewport-adaptive streaming. Two coding layers, one base layer (BL) and one enhancement layer (EL), have been considered. The base layer consists of the lower quality version or non-viewport area of the 360° content, and the higher quality version or the viewport is coded in the enhancement-layer of the SHVC.

In SHVC-ROI method, the viewport content in EL is coded using the motion constraint tile set (MCTS) technique. In order to resolve the issue of frequent IRAPs in high-quality content, the inter-layer prediction (ILP) feature of SHVC is used. To do this, the EL tiles in IRAP pictures are predicted from the co-located regions in the BL. By utilizing such prediction, the switching point pictures are coded as P pictures, thus requiring significantly lower bitrates compared to I-coded IRAPs. Furthermore, the ILP feature from the BL content may be used in non-IRAP pictures as well for further reducing the bitrate of high-quality video.

Using the ILP feature brings the requirements of streaming the entire BL video. Consequently, the portion of the content that is used as non-viewport area must be transmitted along with the portion which is used for ILP operation. Thus, there is no need to use tiling in BL. Moreover, for reducing the bitrate of IRAPs in BL, longer IRAP intervals are considered compared to high-quality EL content. The SHVC-ROI process for EL and BL is illustrated in Figure 4.2.

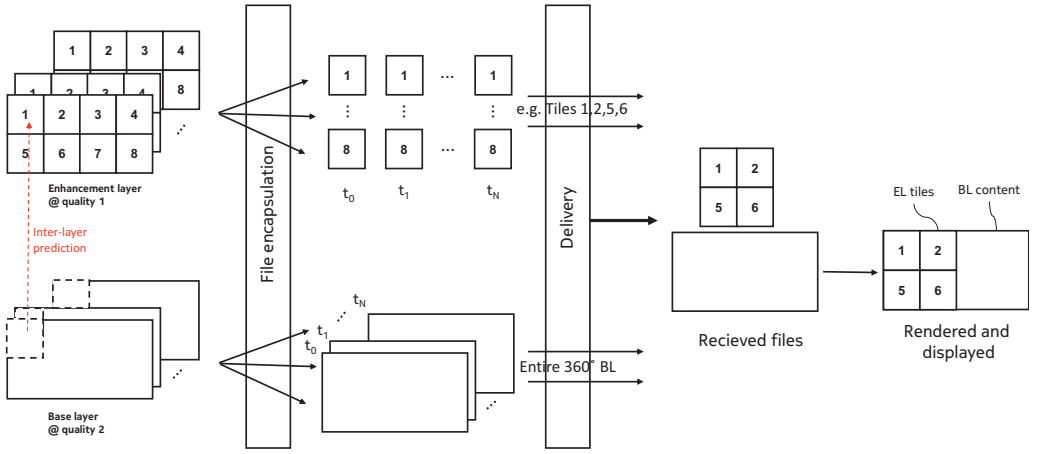


Figure 4.2 Tile-based viewport-adaptive streaming with multi-layer SHVC method [**P5**]

4.2.2 Simulcast HEVC VAS

The SHVC-ROI method presented in Section 4.2.1 resolves the frequent IRAP requirement for viewport switching in both high- and low-quality content. However, this scheme requires to use the scalable decoder of HEVC which is not widely supported in hardware implementations. This section describes a method, called Simulcast HEVC, for resolving the frequent IRAP issue in the low-quality 360° without the use of SHVC codec.

According to the Simulcast HEVC method, illustrated in Figure 4.3, the high-quality version of the content is coded the same as MCTS-based technique with the conventional IRAP intervals for switching operations. However, for reducing the bitrate of the lower quality version of the 360° video, the same approach as the base-layer of SHVC-ROI method is used. In other words, the low-quality 360° video is coded conventionally and without using tiling technique and longer IRAP intervals are considered compared to high-quality version. This method can be considered as a combination of MCTS-based (in viewport area) and SHVC-ROI (in non-viewport area) approaches.

The proposed method reduces streaming bitrate for lower quality version of the content significantly by using longer IRAP intervals in the encoded content. The Simulcast method also removes the necessity of tiling in low-quality content, thus, benefits from not having tiling compression drawbacks in such content. Further-

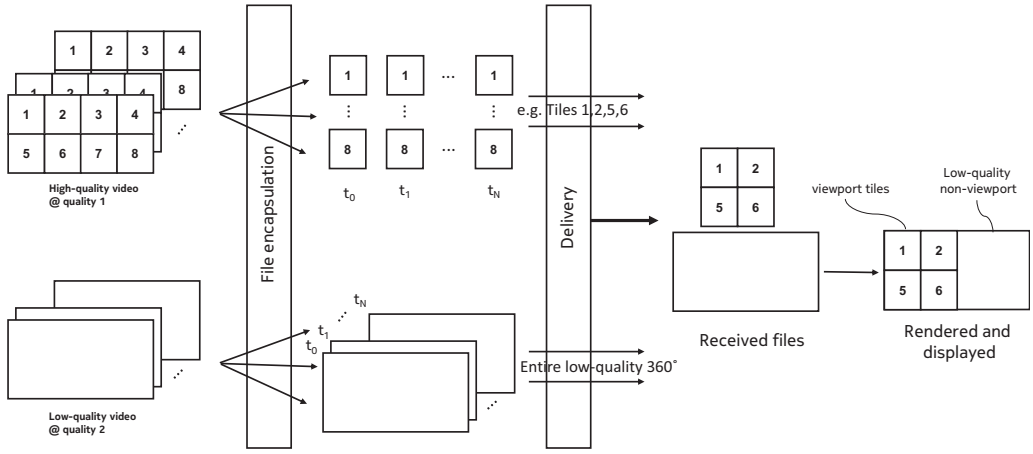


Figure 4.3 Viewport-adaptive streaming with Simulcast HEVC method

Table 4.1 Average streaming BD-Rate (%) comparison of the methods in various tile grids [**P4**]

Method	Tiling grid			
	4×2	6×3	12×4	12×8
MCTS-based	-19.1%	-29.3%	-25.6%	-27.8%
SHVC-ROI	-34.4%	-47.0%	-51.2%	-53.0%
Simulcast HEVC	-19.5%	-36.9%	-40.6%	-45.8%

more, the transmitted bitstreams are decodable using single-layer HEVC decoders. However, this method requires frequent IRAPs in higher quality version of the content which consume majority of the streaming bitrate compared to low-quality versions. Another disadvantage of this method is that it requires transmitting the entire 360° low-quality video to the user which is not optimal in terms of decoding complexity considerations.

4.2.3 Experimental Results

The HM-16.15 [35] test model of the HEVC standard [36] was used for conducting the simulations in reference, MCTS-based and Simulcast HEVC methods. For the SHVC-ROI approach, the SHM-12.2 reference software [71] of scalable extension of HEVC was used.

Table 4.1 shows the average streaming performance of the MCTS-based, SHVC-

ROI and Simulcast HEVC methods with different tile grids compared to the conventional 360° video streaming case. As can be seen from the results, the SHVC-ROI scheme provides high bitrate reductions in all tiling cases, particularly in 12×4 and 12×8 tiling grids where bitrate reductions of more than 50% is achieved. This is mainly due to using infrequent IRAP pictures in switching points of both low- and high-quality content. On the other hand, the Simulcast HEVC avoids using IRAPs only in low-quality version of the 360° stream. Hence, it provides better performance than MCTS-based method and worse performance than SHVC-ROI scheme. Moreover, both SHVC-ROI and Simulcast methods avoid using tiles in low-quality 360° video, thus, these methods benefit from not having compression penalty due to tiling in the content. Compression loss analysis of different tiling grids are discussed in the next section.

4.3 Shared Coded Picture Technique for Tile-based Viewport-adaptive Streaming

The viewport-adaptive streaming methods, presented in Section 4.2, reduce the switching points bitrates using infrequent IRAP pictures in both high- and low-quality streams (in SHVC-ROI method) or only in low-quality bitstream (in Simulcast HEVC method). However, as mentioned, the SHVC-ROI method requires scalable codecs which are not widely supported in hardware decoders. The second method, i.e. Simulcast HEVC, uses single-layer HEVC decoders but resolves the frequent IRAP pictures only in low-quality version of the 360° video and does not take advantage of the similarity between low- and high-quality streams.

This section proposes a novel Shared Coded Picture (SCP) technique that allows infrequent IRAP pictures in both high- and low-quality video content, while the generated bitstreams are decodable with single-layer HEVC decoders. Furthermore, the proposed SCP-based scheme is fully compliant with the Omnidirectional MediA Format (OMAF) standard [42]. According to OMAF v1, the VAS methods must follow the requirements and constraints stated below:

- Frequent viewport switching capability,
- Single-layer decoding constraint,
- Decoding with single decoder instance,

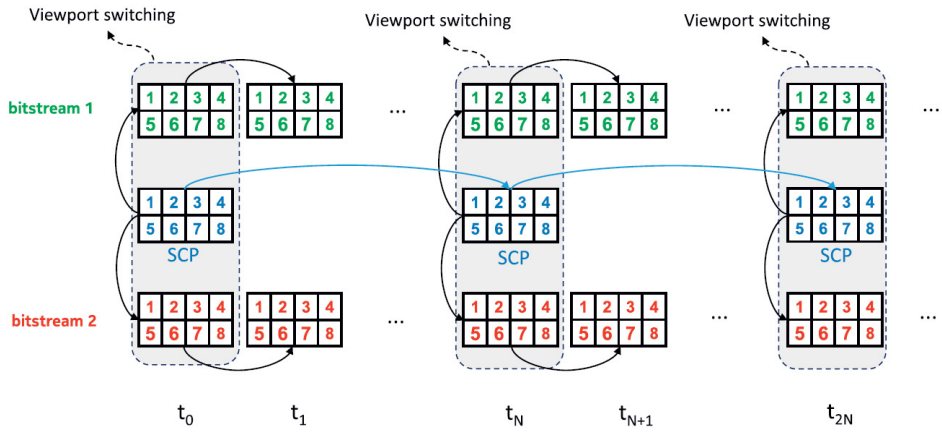


Figure 4.4 Illustrates the SCP concept in viewport switching points [**P5**]

- 4K decoding constraint.

4.3.1 Algorithm Description

The core concept of the proposed SCP-based scheme is illustrated in Figure 4.4. The main idea, as can be seen in the figure, is to have certain pictures in the low- and high-quality bitstreams that are identical to each other. These identical pictures are marked as Shared Coded Pictures (SCPs) in the figure. Since SCP pictures are coded in a way that are identical in all versions of the bitstreams, they can be used as viewport switching points from one bitstream to another. Moreover, SCP pictures are configured in such a way that every SCP is inter-coded from the previously coded SCP, thus they require significantly lower bitrates than intra-coded switching points.

In order to enable such coding scheme, the video sequence is pre-processed in order to duplicate certain pictures in the sequence. These duplicate pictures are used as viewport switching operations and as a result, the sequence contains one extra picture per switching point. For distinguishing these pictures, the first picture of the repeated pictures is referred to as Shared Coded Picture (SCP) and the second picture is called Duplicated Picture (DP).

Figure 4.5 demonstrates the designed prediction hierarchy for enabling the SCP-based method. SCPs are marked with blue colors in the prediction hierarchy of both streams in the figure. As can be seen, SCPs are predicted from the previously

coded SCP in the prediction configuration whereas the DPs are predicted based on their corresponding SCP. The remaining pictures in the interval of SCPs follow the normal prediction hierarchy. Due to the repetition of content into SCPs and DPs, these pictures have the same content. In order to avoid displaying the same content twice to the user, SCP pictures are marked as non-output pictures by setting the *pic_output_flag* syntax element (SE) of these pictures to zero. Consequently, these pictures are marked as non-output pictures in the file encapsulation, as specified in [39].

The selection of viewport and non-viewport tiles in the SCP-based method is done similarly as the MCTS-based method, where a set of high-quality tiles (corresponding to viewport area) is selected from the high-quality bitstream and the remaining non-viewport area is selected from the low-quality tiles in the second bitstream. The streaming process of the proposed SCP-based technique is illustrated in Figure 4.6. As shown in the figure, the viewport switching from one viewing orientation to another takes place in the SCP that are inter-coded pictures. As a consequence, the bitrate of these switching point pictures decreases significantly and this results in streaming bitrate improvements compared to MCTS-based methods with frequent IRAPs. Furthermore, the proposed SCP-based VAS method achieves the following advantages:

- Enables frequent viewport switching without the need for IRAP pictures,
- Compliant with 4K decoding requirement of OMAF,
- Compliant with single-layer decoding requirement of OMAF,
- Better streaming performance compared to MCTS-based method.

4.3.2 Experimental Results

The HM-16.7 test model [35] of the HEVC standard [36] was used for conducting the experiments in reference, MCTS-based, Simulcast HEVC and SCP-based methods. For the SHVC-ROI approach, the SHM-12.2 reference software [71] of the scalable extension of HEVC was used. The simulations are performed based on the common test condition (CTC) with Random Access (RA) configuration [7].

In the experiments, two versions of the SHVC-ROI method are tested:

- Full Inter-layer Prediction (FILP): where the inter-layer prediction feature is

Table 4.2 Average streaming BD-Rate (%) comparison of the methods in various tile grids [**P5**]

Method	Tiling grid				
	4×2	6×3	8×4	12×4	12×8
MCTS-based	-21.9%	-32.4%	-26.6%	-33.0%	-31.6%
FILP SHVC-ROI	-32.6%	-45.4%	-40.9%	-50.5%	-53.4%
CILP SHVC-ROI	-30.9%	-44.2%	-39.6%	-49.5%	-52.7%
Simulcast HEVC	-18.1%	-35.0%	-28.5%	-40.8%	-44.3%
SCP-based	**-33.5%**	**-45.0%**	**-38.9%**	**-46.2%**	**-45.2%**

enabled in all frames.

- Constraint Inter-layer Prediction (CILP): where the inter-layer prediction feature is enabled only in viewport switching points.

The average streaming performance of the proposed methods under different tiling setups is presented in Table 4.2. As can be observed from the results, the SCP-based and SHVC-ROI schemes provide the highest bitrate reductions among the described VAS approaches.

The performance of the SCP-based and SHVC-ROI methods are very close in coarse tiling configurations (4×2 and 6×3 tilings), whereas, the SHVC-ROI methods provide better bitrate reductions in finer tiling configurations (12×4 and 12×8 tilings). The reason for this performance behavior is that the SCP-based method uses MCTS technique in both high- and low-quality content, whereas, the SHVC-ROI method utilizes this technique only in high-quality version of the video. Thus, the tiling penalty increase is higher in the finer tiling granularity compared to the scalable coding approach. The impact of having motion constraint tiles to the compression performance of the entire 360° ERP video is shown in Table 4.3. As seen, tiling overhead corresponds to the number tiles that are used in the content. In coarser tiling grids (e.g. 4×2 and 6×3), the compression penalty is in the range of 3% to 6%, however, in the finer tiling scenarios (e.g. 12×4 and 12×8) compression losses are in the range of 9% to 19% which are significant.

Table 4.4 provides an analysis of decoder-side complexities of described VAS schemes in terms of required number of pixels to be decoded relative to the number of pixels in the full 360° ERP video. Among them the MCTS-based method requires 100% of the pixels to be decoded as the selected tiles from low- and high-quality bitstreams are

Table 4.3 Average tiling overheads in terms of BD-Rate (%) for different tile grids [**P3**]

Category	Tiling grid				
	4×2	6×3	8×4	12×4	12×8
High-quality content	3.5%	4.2%	7.1%	9.4%	12.8%
Low-quality content	4.4%	6.0%	9.8%	13.5%	19.1%

Table 4.4 Decoder-side complexities in terms on number of decoded pixels [**P5**]

Method	Content	Tiling grid				
		4×2	6×3	8×4	12×4	12×8
MCTS-based	High-quality	59.9%	47.7%	41.7%	38.7%	32.4%
	Low-quality	40.1%	52.3%	58.3%	61.3%	67.6%
	Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**
SHVC-ROI	High-quality	59.9%	47.7%	41.7%	38.7%	32.4%
	Low-quality	100.0%	100.0%	100.0%	100.0%	100.0%
	Total	**159.9%**	**147.7%**	**141.7%**	**138.7%**	**132.4%**
Simulcast	High-quality	59.9%	47.7%	41.7%	38.7%	32.4%
	Low-quality	100.0%	100.0%	100.0%	100.0%	100.0%
	Total	**159.9%**	**147.7%**	**141.7%**	**138.7%**	**132.4%**
SCP-based	High-quality	59.9%	47.7%	41.7%	38.7%	32.4%
	Low-quality	44.3%	56.5%	62.5%	65.5%	71.8%
	Total	**104.2%**	**104.2%**	**104.2%**	**104.2%**	**104.2%**

complementary to each other. Similarly, in SCP-based method the selected tiles from different quality bitstreams are complementary, however, since this method requires an additional SCP picture at each viewport switching interval the overall number of decoded pixels are slightly higher than MCTS-based method. In the case of SHVC-ROI and Simulcast VAS schemes, the entire 360° video at low-quality along with the high-quality viewport tiles are transmitted to the user. Hence, these methods have significantly higher decoding complexities in terms of number of pixels needed to be decoded.

The goal of this chapter was to improve the streaming performance of omnidirectional video compared to the state-of-the-art MCTS-based method. The three methods that are proposed for this purpose provide significant streaming bitrate re-

ductions, however, the proposed SCP-based method is considered to be the most optimal choice when considering the bitrate reduction versus complexity trade-off. The SCP-based technique is able to reduce the streaming bitrate on average by 10% to 15%, depending on the choice of tiling granularity, compared to conventional MCTS-based method.

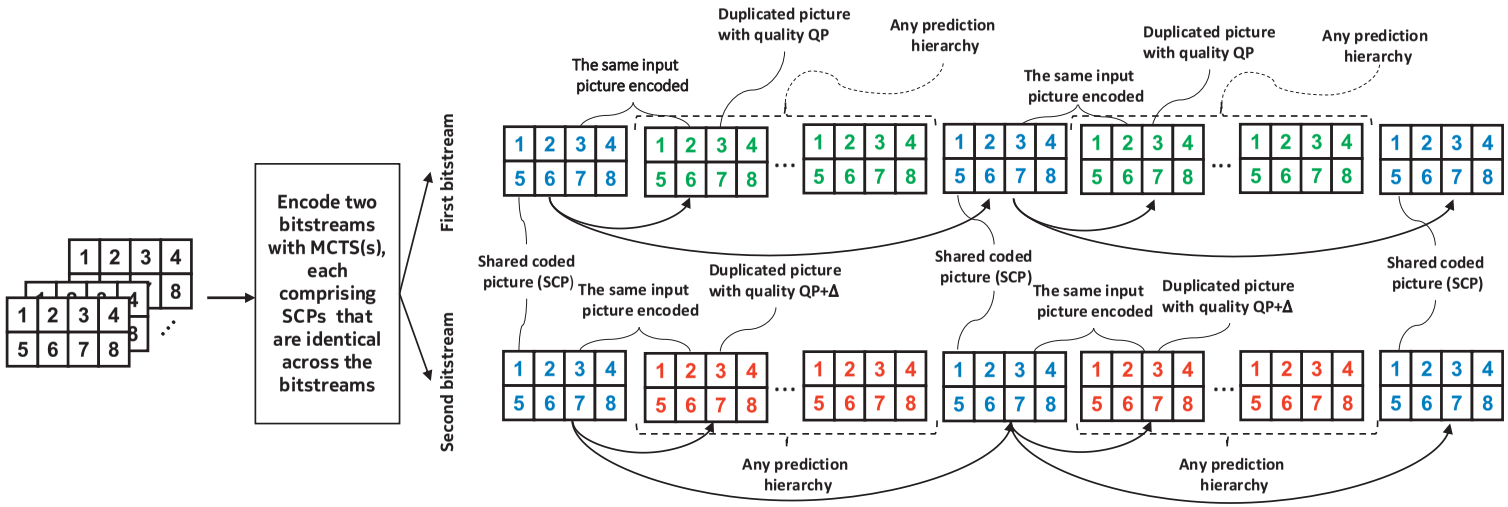


Figure 4.5 Illustration of the SCP-based encoding with MCTS [P5]

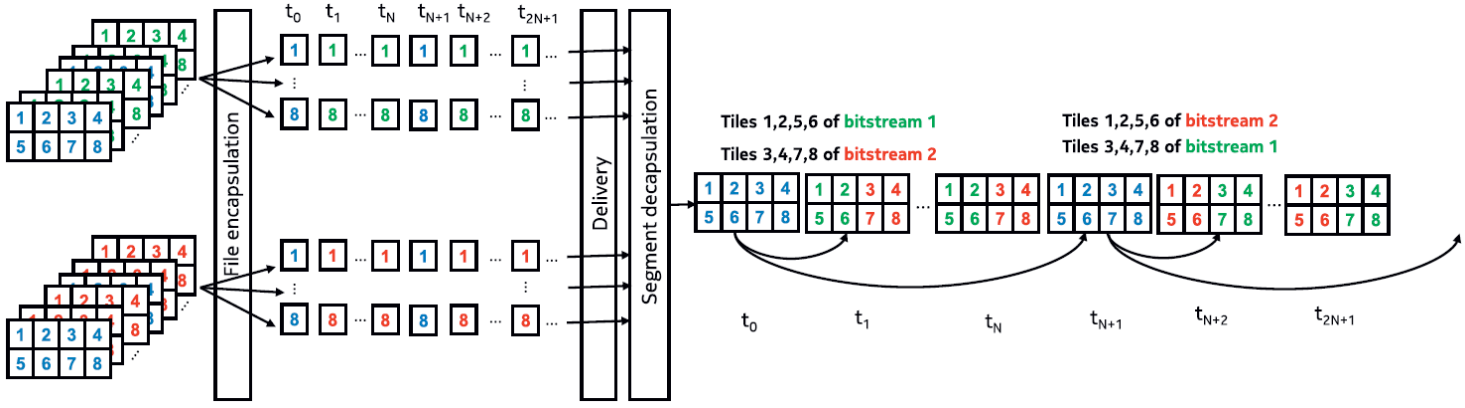


Figure 4.6 Illustration of the SCP-based streaming with MCTS [P5]

5 CONCLUSION

Virtual reality (VR) applications are becoming mainstream nowadays. The recent technological breakthroughs enabled applications such as gaming, entertainment, streaming live events (e.g., sports and concerts), health, education. 3-DoF VR technology makes use of omnidirectional content for creating immersion in the virtual environment. Omnidirectional content is characterized by 360° and 180° field-of-views (FOVs) in the horizontal and vertical directions, respectively. In order to create a proper immersive experience via VR, apart from the 360° FOV requirement, the content must be presented to the user in high resolution, quality and frame rate (90 Hz). Such combined requirements bring significant challenges for the current content handling and content communication.

This dissertation presents solutions for improving the compression and streaming performances of omnidirectional video. The proposed solutions are divided into two categories which are presented in Chapters 3 and 4.

In the first category, presented in Chapter 3, we studied several methods for improving the compression efficiency of 360° video. The methods targeted motion estimation and compensation challenges and proposed three approaches for increasing the inter-prediction efficiency in the projected omnidirectional video. We proposed a novel motion vector scaling method that is applied to the motion information of the neighboring blocks based on the geometry of the projection plane. The scaled motion information provides uniform predictors for the motion vector of the current block, resulting in an improved motion vector prediction and lower bitrates. The experiments illustrated that the proposed method was able to reduce the bitrate on average around 1.0%, for the sequences with high motion, compared to the reference HEVC test model. The second contribution consists of an adaptive motion vector prediction method, where a 6-parameter motion model is proposed for estimating the motion information. The introduced motion model uses neighboring motion information for deriving the parameters of the model. The model is able

to efficiently predict the motion vectors of the block. An extension of the adaptive motion vector prediction is also presented, where the 6-parameter motion model is used for deriving the motion vectors of a block at a finer granularity (at the level of 4×4 and 8×8 sub-block accuracy). The proposed sub-block motion vector prediction method is able to estimate and compensate the non-linear motion behavior of the 360° video in a more efficient way. The experiments show that the sub-block RMVF method is able to decrease the average bitrate by more than 0.50% compared to the state-of-the-art VVC test model. The results also illustrate that the proposed method does not impose significant complexities to the existing coding chain. Recall the first research question, can we develop more efficient coding tools to improve the compression performance of current codecs (HEVC and VVC) in order to deal with immersive virtual reality video, we can see that the contributions presented in this thesis have addressed the question and proposed efficient solutions which are able to reduce the bitrates by up to 1.0%, which is a significant reduction in video compression.

In the second category we investigated and proposed novel methods for improving the tile-based viewport-adaptive streaming (VAS), where the main goal is to reduce the high bitrate of the viewport switching points. To this end, three methods are proposed for enabling seamless viewport switching operations without the need for frequent Intra Random Access Point (IRAP) pictures. First, we propose a multi-layer coding scheme which makes use of the scalable extension of the HEVC standard in such a way that the viewport and non-viewport content are encoded in different layers of the codec. The inter-layer prediction (ILP) functionality of this codec is used for resolving the frequent IRAPs in viewport area. Using the ILP requires to stream the whole 360° base-layer video, thus, it is not necessary to use frequent IRAPs for this content. Therefore, longer IRAP intervals are considered for the non-viewport area. An extension of this method, called Simulcast HEVC, uses longer IRAP intervals in non-viewport area. The viewport content is coded using the tile-based approach with conventional IRAP period. This method resolves the frequent intra-coded pictures in switching points of non-viewport area while using single-layer codecs. Finally, a novel Shared Coded Picture (SCP) method is proposed to facilitate the viewport switching operations via SCP pictures that are shared in different quality versions of the content and coded identically. These SCPs are coded based on the previously coded SCPs, thus they require significantly lower bitrates

than the conventional intra-coded viewport switching pictures. Consequently, the proposed SCP-based VAS enables the seamless viewport switching operation with infrequent IRAPs in a way that the generated bitstream is decodable with standard single-layer codecs. Recall the second research question: how we can stream a high quality immersive video to the viewer, the latter contributions of the thesis have addressed this question by exploring tile-based viewport-adaptive streaming, Simulcast HEVC and Shared Coded Pictures. The proposed methods achieved substantial results, exceeding bitarate reductions of 20% in the cases of SHVC-ROI scheme and 15% in the SCP-based method compared to the state-of-the-art tile-based streaming.

To conclude, the proposed methods provide substantial improvements in compression and streaming omnidirectional video. In compression category, the proposed RMVF algorithm is able to predict the motion vectors of the block at sub-block level accuracy, hence, it is capable of modeling the complex non-linear motion behavior which is caused by the deformations of the underlying content. Although this approach has been shown to be appropriate, it presents some inherent limitations. In particular, estimating the parameters of the motion model is still considered to be complex to some extent since it uses all the available sub-block motion vectors from the immediate neighboring blocks. As future work, the complexity reduction of this algorithm can be investigated for example by determining the optimal number of motion vectors needed for training and the optimal locations for selecting neighboring motion vectors. Furthermore, for the geometry-based motion vector scaling, the proposed method only studied the scaling of MVs in ERP format. This scheme can be extended to other projection formats, such as cubemap projection.

In omnidirectional video streaming, the SCP-based technique was able to remove the need for frequent IRAP pictures in viewport switching points using Shared Coded Pictures. However, the SCP method has a number of limitations, including the fact that it is applicable to only VAS schemes with quality adaptation approach. Therefore, it is worthy to further investigate how this method can additionally be used in resolution adaptation schemes. Additional potential aspects for further study may include investigating the impact of different SCP intervals and its relation to motion-to-high-quality delay. Finally, further studies may target the tiling granularity and the quality selection for the viewport and non-viewport tiles in the case of user's head motion.

REFERENCES

[1] *Advanced Video Coding*. Document Rec. ITU-T H.264, ISO/IEC 14496- 10 AVC, May 2003.

[2] S.-N. Akula, A. Singh, A. Dsouza and R.-N. Gadde. AHG8: Efficient Frame Packing for Icosahedral Projection. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, document JVET-E0029* (Jan. 2017).

[3] P. R. Alface, J.-F. Macq and N. Verzijp. Interactive omnidirectional video delivery: A bandwidth-effective approach. *Bell Labs Technical Journal* 16.4 (Mar. 2012), 135–147. DOI: 10.1002/bltj.20538.

[4] E. Alshina, J. Boyce, A. Abbas and Y. Ye. JVET common test conditions and evaluation procedures for 360 degree video. *document JVET-G1030* (July 2017).

[5] G. V. der Auwera, M. Coban, Hendry and M. Karczewicz. Truncated square pyramid projection (TSP) for 360 video. *document JVET-D0071* (Oct. 2016).

[6] G. Bjøntegaard. Calculation of average PSNR differences between RD-curves (VCEG-M33). *VCEG Meeting (ITU-T SG16 Q. 6)*. 2001.

[7] F. Bossen. Common test conditions and software reference configurations. *document ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Joint Collaborative Team on Video Coding (JCT-VC), document JCTVC-L1100* (2013).

[8] F. Bossen, J. Boyce, X. Li, V. Seregin and K. Sühring. JVET common test conditions and software reference configurations for SDR video. *JVET document, JVET-K1010* (July 2018).

[9] B. Bross. Versatile Video Coding (VVC) Draft 1. *MPEG Joint Video Exploration Team* document JVET-J1001 (Apr. 2018).

[10] B. Bross. Versatile Video Coding (VVC) Draft 2. *MPEG Joint Video Exploration Team* document JVET-K1001-v7 (July 2018).

[11] M.-L. Champel, R. Koenen, G. Lafruit and M. Budagavi. Working Draft 0.4 of TR: Technical Report on Architectures for Immersive Media. *120th MPEG meeting of ISO/IEC JTC1/WG11, Document N17264*. Oct. 2017.

[12] L. Chiariglione. Moving picture experts group (MPEG). *Scholarpedia* 4.2 (2009), 6600.

[13] B. Choi, Y.-K. Wang, M. M. Hannuksela, Y. Lim and A. Murtaza. Information technology–coded representation of immersive media (MPEG-I)–part 2: Omnidirectional media format. *ISO/IEC* (2017), 23090–2.

[14] X. Corbillon, A. Devlic, G. Simon and J. Chakareski. Optimal set of 360-degree videos for viewport-adaptive streaming. *Proceedings of the 25th ACM international conference on Multimedia*. Oct. 2017, 943–951. DOI: https://doi.org/10.1145/3123266.3123372.

[15] X. Corbillon, G. Simon, A. Devlic and J. Chakareski. Viewport-adaptive navigable 360-degree video delivery. *IEEE international conference on communications (ICC)*. May 2017, 1–7. DOI: 10.1109/ICC.2017.7996611.

[16] S. Deshpande, Y.-K. Wang and M. M. Hannuksela. Text of ISO/IEC FDIS 23090-2 2nd edition OMAF. *ISO/IEC JTC1 SC29 WG3 document N00072*. Dec. 2020.

[17] *Facebook 360 Virtual reality*. [Accessed November 2020]. URL: https://facebook360.fb.com/.

[18] Y. S. de la Fuente, G. Bhullar, R. Skupin, C. Hellge and T. Schierl. Delay impact on MPEG OMAF's tile-based viewport-dependent 360° video streaming. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.1 (Feb. 2019), 18–28. DOI: 10.1109/JETCAS.2019.2899516.

[19] Y. S. de la Fuente, R. Skupin and T. Schierl. Compressed domain video processing for tile based panoramic streaming using HEVC. *IEEE International Conference on Image Processing (ICIP)*. Sept. 2015, 2244–2248. DOI: 10.1109/ICIP.2015.7351200.

[20] Y. S. de la Fuente, R. Skupin and T. Schierl. Compressed domain video processing for tile based panoramic streaming using SHVC. *Proceedings of the 3rd International Workshop on Immersive Media Experiences*. Oct. 2015, 13–18. DOI: https://doi.org/10.1145/2814347.2814353.

[21] Y. S. de la Fuente, R. Skupin and T. Schierl. Video processing for panoramic streaming using HEVC and its scalable extensions. *Multimedia Tools and Applications* 76.4 (Dec. 2017), 5631–5659.

[22] M. Ghanbari. *Standard codecs: Image compression to advanced video coding*. 49. Iet, 2003.

[23] R. Ghaznavi-Youvalari, A. Aminlou and M. M. Hannuksela. Analysis of regional down-sampling methods for coding of omnidirectional video. *Picture Coding Symposium (PCS)*. IEEE. Dec. 2016, 1–5. DOI: 10.1109/PCS.2016.7906403.

[24] R. Ghaznavi-Youvalari, A. Aminlou, M. M. Hannuksela and M. Gabbouj. Efficient coding of 360-degree pseudo-cylindrical panoramic video for virtual reality applications. *IEEE International Symposium on Multimedia (ISM)*. Dec. 2016, 525–528. DOI: 10.1109/ISM.2016.0115.

[25] R. Ghaznavi-Youvalari, A. Aminlou and J. Lainema. CE4-related: Merge mode with regression based motion vector field (RMVF). *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-L0171* (Oct. 2018).

[26] R. Ghaznavi-Youvalari, A. Aminlou and J. Lainema. CE2: Merge mode with regression-based motion vector field (test 2.3.3). *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-M0302* (Jan. 2019).

[27] R. Ghaznavi-Youvalari, M. M. Hannuksela, A. Aminlou and M. Gabbouj. Viewport-dependent delivery schemes for stereoscopic panoramic video. *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*. IEEE. June 2017, 1–4. DOI: 10.1109/3DTV.2017.8280404.

[28] R. Ghaznavi-Youvalari, A. Zare, A. Aminlou and M. M. Hannuksela. OMAF VDC CE: Test results on viewport dependent delivery schemes. *MPEG meeting, document m40411* (Mar. 2017).

[29] R. Ghaznavi-Youvalari, A. Zare, A. Aminlou and M. M. Hannuksela. OMAF: HEVC MCTS-based encoding and SHVC-ROI scalability for viewport dependent streaming. *ISO/IEC JTC1/SC29/WG11 MPEG meeting, document m39898* (Jan. 2017).

[30] *GoPro Max 360 camera*. [Accessed November 2020]. URL: https://projectgo.pro/gopro-max-manual/.

[31] F. Hanhart, J. Boyce, K. Choi and J. Lin. JVET common test conditions and evaluation procedures for 360 degree video. *document JVET-L1012* (Oct. 2018).

[32] M. M. Hannuksela. OMAF: viewport dependent video coding schemes. *ISO/IEC JTC1/SC29/WG11 (MPEG) document M39864* (Jan. 2017).

[33] M. M. Hannuksela, Y.-K. Wang and M. Gabbouj. Isolated regions in video coding. *IEEE Transactions on Multimedia* 6.2 (Mar. 2004), 259–267. DOI: 10.1109/TMM.2003.822784.

[34] M. M. Hannuksela, Y.-K. Wang and A. Hourunranta. An overview of the OMAF standard for 360 video. *Data Compression Conference (DCC)*. IEEE. June 2019, 418–427. DOI: 10.1109/DCC.2019.00050.

[35] *High Efficiency Video Coding (HEVC) reference software HM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute*. [Accessed December 2020]. URL: https://hevc.hhi.fraunhofer.de/..

[36] *High Efficiency Video Coding, Version 1*. Document Rec. ITU-T H.265, ISO/IEC 23008-2, Jan. 2013.

[37] J. V. der Hooft, M. Vega, S. Petrangeli, T. Wauters and F. Turck. Tile-based adaptive streaming for virtual reality video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.4 (Dec. 2019), 1–24. DOI: https://doi.org/10.1145/3362101.

[38] *How ClearVR drives and leverages standards*. [Accessed November 2020]. URL: https://www.tiledmedia.com/index.php/standards/.

[39] *Information technology – Coding of audio-visual objects – Part 15: Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format*. 4th ed. Feb. 2017.

[40] *Information technology — Coding of audio-visual objects — Part 12: ISO base media file format*. ISO/IEC 14496-12:2015, Dec. 2015.

[41] *ISO/IEC 23009-1, "Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats".*

[42] *ISO/IEC 23090-2, Information technology – Coded representation of immersive media (MPEG-I), Part 2: Omnidirectional Media Format (OMAF).*

[43] *JVET 360Lib software*. [Accessed December 2020]. URL: https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/tags/.

[44] K. Kammachi-Sreedhar, A. Aminlou, M. M. Hannuksela and M. Gabbouj. Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications. *IEEE International Symposium on Multimedia (ISM)*. Dec. 2016, 583–586. DOI: 10.1109/ISM.2016.0126.

[45] K. Kammachi-Sreedhar, A. Aminlou, A. Zare and M. M. Hannuksela. Testing methodology for viewport-dependent encoding and streaming. *MPEG meeting, document m39081*. Oct. 2016.

[46] K. Kammachi-Sreedhar, A. Zare, A. Aminlou and M. M. Hannuksela. Testing methodology for viewport-dependent encoding and streaming. *ITU-T Joint Video Exploration Team (JVET), document JVET-D0079* (Oct. 2016).

[47] J. Kravec. Remote rendering for VR and mobile devices with efficient illumination streaming. *Czech Technical University*. 2020. URL: https://cescg.org/cescg_submission/remote-rendering-for-vr-and-mobile-devices-with-efficient-illumination-streaming/.

[48] E. Kuzyakov. *End-to-end optimizations for dynamic streaming*. Feb. 2017. URL: https://code.facebook.com/posts/637561796428084/.

[49] E. Kuzyakov and D. Pio. *Next-generation video encoding techniques for 360 video and VR*. [Accessed December 2020]. Jan. 2016. URL: https://code.facebook.com/posts/1126354007399553/.

[50] S. Lederer. Today's and future challenges with new forms of content like 360° AR and VR. *invited talk in MPEG workshop Global Media Technology Standards for an Immersive Age*. [Accessed December 2020]. 2017. URL:

https://mpeg.chiariglione.org/sites/default/files/events/06_Lederer.pdf.

[51] L. Li, Z. Li, M. Budagavi and H. Li. Projection based advanced motion model for cubic mapping for 360-degree video. *IEEE International Conference on Image Processing (ICIP)*. IEEE. Sept. 2017, 1427–1431. DOI: 10.1109/ICIP.2017.8296517.

[52] H.-C. Lin, C.-C. Huang, C.-Y. Li, Y.-H. Lee, J.-L. Lin and S.-K. Chang. AHG8: An improvement on the compact OHP layout. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, document JVET-E0056* (Jan. 2017).

[53] H.-C. Lin, C.-Y. Li, J.-L. Lin, S.-K. Chang and C.-C. Ju. AHG8: An efficient compact layout for Octahedron format. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0142* (Oct. 2016).

[54] D. Marpe, T. Wiegand and G. J. Sullivan. The H.264/MPEG4 advanced video coding standard and its applications. *IEEE communications magazine* 44.8 (Aug. 2006), 134–143. DOI: 10.1109/MCOM.2006.1678121.

[55] W. Mason. *VR HMD Roundup: Technical Specs*. [Accessed December 2020]. URL: http://uploadvr.com/vr-hmd-specs/.

[56] A. Mehrfard, J. Fotouhi, G. Taylor, T. Forster, N. Navab and B. Fuerst. A comparative analysis of virtual reality head mounted display systems. *arXiv preprint arXiv:1912.02913* (Dec. 2019).

[57] MPEG Strategic Standardisation Roadmap. *115th MPEG meeting of ISO/IEC JTC1/WG11, Document N16316*. June 2016.

[58] A. T. Nasrabadi, A. Mahzari, J. Beshay and R. Prakash. Adaptive 360-degree video streaming using scalable video coding. *Proceedings of the 25th ACM international conference on Multimedia*. Oct. 2017, 1689–1697. DOI: https://doi.org/10.1145/3123266.3123414.

[59] *Nokia's OMAF implementation*. [Accessed December 2020]. URL: https://github.com/nokiatech/omaf.

[60] *Oculus Quest 2 Virtual Reality Headset*. [Accessed November 2020]. URL: https://www.oculus.com/quest-2/.

[61] *Oculus Quest Virtual Reality Headset*. [Accessed November 2020]. URL: https://www.oculus.com/quest/.

[62] *Oculus Rift Virtual Reality Headset*. [Accessed November 2020]. URL: https://www.oculus.com/rift/.

[63] *Open visual cloud immersive video samples*. [Accessed December 2020]. URL: https://github.com/OpenVisualCloud/Immersive-Video-Sample.

[64] G. Papaioannou and I. Koutsopoulos. Tile-based caching optimization for 360 videos. *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. July 2019, 171–180. DOI: https://doi.org/10.1145/3323679.3326515.

[65] D. Podborski, J. Son, G. S. Bhullar, R. Skupin, Y. Sanchez, C. Hellge and T. Schierl. HTML5 MSE playback of MPEG 360 VR tiled streaming: JavaScript implementation of MPEG-OMAF viewport-dependent video profile with HEVC tiles. *Proc. of ACM Multimedia Systems Conference*. June 2019. URL: https://github.com/fraunhoferhhi/omaf.js.

[66] D. Podborski, E. Thomas, M. M. Hannuksela, S. Oh, T. Stockhammer and S. Pham. Virtual reality and DASH. *International Broadcasting Convention (IBC)*. 2017.

[67] *Ricoh Theta V camera*. [Accessed November 2020]. URL: https://theta360.com/en/about/theta/v.html.

[68] J. Ridge, M. M. Hannuksela, E. Aksu, J. Lainema and A. Aminlou. Nokia test sequences for virtual reality video coding. *ITU-T Joint Video Exploration Team (JVET), document JVET-C0064*. June 2016.

[69] *Samsung Gear 360 camera*. [Accessed November 2020]. URL: https://www.samsung.com/global/galaxy/gear-360/.

[70] *Samsung Gear VR Head Mounted Display*. [Accessed November 2020]. URL: https://www.samsung.com/global/galaxy/gear-vr/.

[71] *Scalable Extensions of the High Efficiency Video Coding (SHVC) reference software SHM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute,* [Accessed December 2020]. URL: https://hevc.hhi.fraunhofer.de/shvc..

[72] F. D. Simone, P. Frossard, N. Birkbeck and B. Adsumilli. Deformable block-based motion estimation in omnidirectional image sequences. *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. Oct. 2017, 1–6. DOI: 10.1109/MMSP.2017.8122254.

[73] R. Skupin, Y. S. de la Fuente and T. Schierl. Compressed domain processing for stereoscopic tile based panorama streaming using MV-HEVC. *IEEE 5th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. Sept. 2015, 160–164. DOI: 10.1109/ICCE-Berlin.2015.7391222.

[74] R. Skupin, Y. Sanchez, Y.-K. Wang, M. M. Hannuksela, J. Boyce and M. Wien. Standardization status of 360 degree video coding and delivery. *IEEE Visual Communications and Image Processing (VCIP)*. IEEE. 2017, 1–4. DOI: 10.1109/VCIP.2017.8305083.

[75] J. Snyder. An equal-area map projection for polyhedral globes. *Cartographica: The International Journal for Geographic Information and Geovisualization* 29.1 (1992), 10–21.

[76] I. Sodagar. The MPEG-DASH standard for multimedia streaming over the internet. *IEEE multimedia* 18.4 (Apr. 2011), 62–67. DOI: 10.1109/MMUL.2011.71.

[77] K. K. Sreedhar, I. D. D. Curcio, A. Hourunranta and M. Lepistö. Immersive media experience with MPEG OMAF multi-viewpoints and overlays. *Proc. of ACM Multimedia Systems Conference*. May 2020. URL: https://www.youtube.com/watch?v=WcucAw3HNVE.

[78] J. A. Steers. *An introduction to the study of map projections*. London: University of London Press, 1959.

[79] 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* 22.12 (Sept. 2012), 1649–1668. DOI: 10.1109/TCSVT.2012.2221191.

[80] G. J. Sullivan, P. Topiwala and A. Luthra. The H.264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions. *Applications of Digital Image Processing XXVII*. Vol. 5558. International Society for Optics and Photonics. Nov. 2004, 454–474. DOI: https://doi.org/10.1117/12.564457.

[81] Y. Sun, A. Lu and L. Yu. AHG8: WS-PSNR for 360 video objective quality evaluation. *document JVET-D0040*. Oct. 2016.

[82] V. Sze, M. Budagavi and G. J. Sullivan. High efficiency video coding (HEVC). Vol. 39. Springer, 2014, 49–90.

[83] I. Tosic, I. Bogdanova, P. Frossard and P. Vandergheynst. Multiresolution motion estimation for omnidirectional images. *13th European Signal Processing Conference*. IEEE. Sept. 2005, 1–4.

[84] *Versatile Video Coding*. Document Recommendation ITU-T H.266, Aug. 2020. URL: https://www.itu.int/rec/T-REC-H.266-202008-I/en.

[85] *Versatile Video Coding (VVC) reference software VTM-1.0. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute*. [Accessed December 2020]. URL: https://jvet.hhi.fraunhofer.de/.

[86] *Versatile Video Coding (VVC) reference software VTM-2.0. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute*. [Accessed December 2020]. URL: https://jvet.hhi.fraunhofer.de/.

[87] M. Viitanen, J. Vanne, T. Hämäläinen and A. Kulmala. Low latency edge rendering scheme for interactive 360 degree virtual reality gaming. *IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE. 2018, 1557–1560. DOI: 10.1109/ICDCS.2018.00168.

[88] B. Vishwanath, T. Nanjundaswamy and K. Rose. Rotational motion model for temporal prediction in 360 video coding. *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. Oct. 2017, 1–6. DOI: 10.1109/MMSP.2017.8122231.

[89] B. Vishwanath, T. Nanjundaswamy and K. Rose. Motion compensated prediction for translational camera motion in spherical video coding. *IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. Aug. 2018, 1–4. DOI: 10.1109/MMSP.2018.8547066.

[90] B. Vishwanath and K. Rose. Spherical video coding with motion vector modulation to account for camera motion. *IEEE Visual Communications and Image Processing (VCIP)*. IEEE. Dec. 2019, 1–4. DOI: 10.1109/VCIP47243.2019.8966083.

[91] B. Vishwanath, K. Rose, Y. He and Y. Ye. Rotational motion compensated prediction in HEVC based omnidirectional video coding. *Picture Coding Symposium (PCS)*. IEEE. June 2018, 323–327. DOI: 10.1109/PCS.2018.8456296.

[92] *Vive Virtual Reality Headsets*. [Accessed November 2020]. URL: https://www.vive.com/eu/.

[93] Y. Wang, R. Wang, Z. Wang and W. Gao. Asymmetric circular projection for dynamic virtual reality video stream switching. *IEEE International Conference on Image Processing (ICIP)*. Sept. 2017, 2726–2730. DOI: 10.1109/ICIP.2017.8296778.

[94] Y.-K. Wang, M. M. Hannuksela, B. Choi, A. Murtaza and Y. Lim. Revised text of ISO/IEC FDIS 23090-2 Omnidirectional Media Format. *MPEG output document N17563* (Apr. 2018).

[95] S. Whittemore-Boggs. A new equal-area projection for world maps. *The Geographical Journal* 73.3 (1929), 241–245.

[96] M. Wien, J. Boyce, T. Stockhammer and W.-H. Peng. Standardization status of immersive video coding. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9.1 (2019), 5–17. DOI: 10.1109/JETCAS.2019.2898948.

[97] Y. Ye, E. Alshima and J. Boyce. JVET-E1003: Algorithm descriptions of projection format conversion and video quality metrics in 360Lib. *Joint Video Exploration Team (JVET) of ITU-T SG* 16 (Jan. 2017).

[98] Y. Ye, E. Alshina and J. Boyce. Algorithm descriptions of projection format conversion and video quality metrics in 360Lib (Version 5). *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-H1004* (Oct. 2017).

[99] Y. Ye, E. Alshina and J. Boyce. Algorithm descriptions of projection format conversion and video quality metrics in 360Lib, Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 6th Meeting: doc. *document JVET-F1003* 31 (Mar. 2017).

[100] *YouTube Virtual Reality*. [Accessed November 2020]. URL: https://vr.youtube.com/.

[101] M. Yu, H. Lakshman and B. Girod. A framework to evaluate omnidirectional video coding schemes. *IEEE International Symposium on Mixed and Augmented Reality*. Oct. 2015, 31–36.

[102] V. Zakharchenko, E. Alshina, K.-P. Choi, A. Singh and A. Dsouza. AhG8: Icosahedral projection for 360-degree video content. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, document JVET-D0028* (Oct. 2016).

[103] V. Zakharchenko, K. Choi and J. Park. Quality metric for spherical panoramic video. *Optics and Photonics for Information Processing X*. Vol. 9970. International Society for Optics and Photonics. Sept. 2016, 99700C.

[104] A. Zare, A. Aminlou and M. M. Hannuksela. Virtual reality content streaming: Viewport-dependent projection and tile-based techniques. *IEEE International Conference on Image Processing (ICIP)*. Sept. 2017, 1432–1436. DOI: 10.1109/ICIP.2017.8296518.

[105] A. Zare, A. Aminlou and M. M. Hannuksela. 6K effective resolution with 4K HEVC decoding capability for OMAF-compliant 360 video streaming. *Proceedings of the 23rd Packet Video Workshop*. 2018, 72–77. DOI: https://doi.org/10.1145/3210424.3210425.

[106] A. Zare, A. Aminlou, M. M. Hannuksela and M. Gabbouj. HEVC-compliant tile-based streaming of panoramic video for virtual reality applications. *Proceedings of the 24th ACM international conference on Multimedia*. Oct. 2016, 601–605. DOI: https://doi.org/10.1145/2964284.2967292.

[107] A. Zare, M. Homayouni, A. Aminlou, M. M. Hannuksela and M. Gabbouj. 6K and 8K Effective Resolution with 4K HEVC Decoding Capability for 360 Video Streaming. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15.2s (2019), 1–22. DOI: https://doi.org/10.1145/3335053.

[108] A. Zare, K. Sreedhar-Kammachi, V. K. M. Vadakital, A. Aminlou, M. M. Hannuksela and M. Gabbouj. HEVC-compliant viewport-adaptive streaming of stereoscopic panoramic video. *Picture Coding Symposium (PCS)*. IEEE. Dec. 2016, 1–5. DOI: 10.1109/PCS.2016.7906401.

[109] C. Zhang, Y. Lu, J. Li and Z. Wen. AhG8: segmented sphere projection (SSP) for 360-degree video content. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, document JVET-D0030* (Oct. 2016).

[110] C. Zhang, Y. Lu, J. Li and Z. Wen. AHG8: Segmented Sphere Projection for 360-degree video. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, document JVET-E0025* (Jan. 2017).

[111] M. Zhou. AHG8: A study on compression efficiency of cube projection. *Document JVET-D0022, Chengdu, CN* (Oct. 2016).

PUBLICATIONS

PUBLICATION I

Geometry-based motion vector scaling for omnidirectional video coding

R. Ghaznavi-Youvalari and A. Aminlou

*IEEE International Symposium on Multimedia (ISM)*2018, 127–130
DOI: 10.1109/ISM.2018.00030

Geometry-based Motion Vector Scaling for Omnidirectional Video Coding

Ramin Ghaznavi-Youvalari, Alireza Aminlou

Nokia Technologies, Tampere, Finland
Email: firstname.lastname@nokia.com

***Abstract*—Virtual reality (VR) applications make use of 360° omnidirectional video content for creating immersive experience to the user. In order to utilize current 2D video compression standards, such content must be projected onto a 2D image plane. However, the projection from spherical to 2D domain introduces deformations in the projected content due to the different sampling characteristics of the 2D plane. Such deformations are not favorable for the motion models of the current video coding standards. Consequently, omnidirectional video is not efficiently compressible with current codecs. In this work, a geometry-based motion vector scaling method is proposed in order to compress the motion information of omnidirectional content efficiently. The proposed method applies a scaling technique, based on the location in the 360° video, to the motion information of the neighboring blocks in order to provide a uniform motion behavior in a certain part of the content. The uniform motion behavior provides optimal candidates for efficiently predicting the motion vectors of the current block. The conducted experiments illustrated that the proposed method provides up to 2.2% bitrate reduction and on average around 1% bitrate reduction for the content with high motion characteristics in the VTM test model of Versatile Video Coding (H.266/VVC) standard.**

***Index Terms*—Video coding, H.266/VVC, 360° video, motion vector**

I. Introduction

Omnidirectional spherical content cover 360° field-of-view (FOV), hence are widely used in virtual reality (VR) applications. In order to use the modern video coding standards such as High Efficiency Video Coding (H.265/HEVC) [1] or Versatile Video Coding (H.266/VVC) [2] for compressing such video, these spherical content are projected onto a two-dimensional (2D) image plane.

Equirectangular projection (ERP) format is among the commonly used formats for omnidirectional content. However, the resulted 2D projection suffers from deformations which caused by different sampling properties in different parts of the ERP image plane. These deformations are more severe in the polar areas. Consequently, current video coding standards are not capable of efficiently modeling such motion behavior in the motion estimation and compensation processes. As a result, the magnitude and direction of the motion vectors in a certain area can vary a lot. Figure 1 demonstrates an example of such motion behavior in different regions of the ERP video. As can be seen, the magnitude and direction of the motion vectors of blocks in a certain area are changing. This is more severe particularly in the blocks that are closer to the polar areas. Due to this non-uniform motion behavior, the motion vector difference (MVD) between blocks would be large and this leads to increase in bitrate of such content.

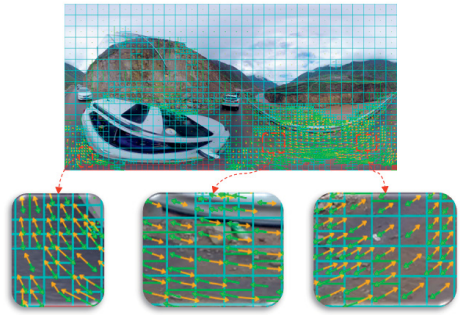


Fig. 1: Region-wise motion behavior in 360° ERP

In video coding algorithms, the motion information of the spatially neighboring blocks are used as predictors for coding the motion information of the current block. This process is applied in Advanced Motion Vector Prediction (AMVP) and Merge coding tools [1], [2]. In this work, an adaptive motion vector scaling method is proposed in order to provide a uniform motion behavior for efficient motion vector coding purposes. The proposed method calculates and applies a scaling factor for the motion vectors of the spatial neighbors based on the location of the current and neighbor block(s) in the 360° ERP video. The scaled motion vectors replace the existing predictor candidates in the AMVP and Merge coding tools.

The conducted experiments illustrated that the proposed motion vector scaling method provides up to 2.2% bitrate reduction and on average by around 1% bitrate reduction in the sequences with high motion. Moreover, there is no bitrate increase observed in the stationary sequences when the proposed method is applied.

The remainder of the paper is organized as follows. Section II reviews the related work for coding the omnidirectional video. The proposed geometry-based motion vector scaling method is described in Section III. Section IV discusses the experimental results. Finally, Section V provides the conclusion of the work.

II. Related Work

Recent studies address the issues of the 2D projection deformations by investigating projection-specific tools for omnidirectional content.

A multiresolution motion estimation method is studied in [3], where the motion estimation for the omnidirectional content is considered in spherical domain. In [4], a translational motion model is proposed for cubemap projection formats which applies the motion estimation operation in the 3D spherical domain by projecting the current and reference blocks to the sphere. A method studied in [5] that considers the motion estimation and block matching processes in the spherical coordinates by using 8×8 block sizes. Such block size limitation is used for reducing the complexity overheads of the motion estimation.

Rotational motion model is studied in [6], [7] where the motion estimation and compensation operations are done in the 3D spherical domain. Moreover, the authors proposed a radial pattern search for this operation in the spherical domain. Another motion model is introduced in [8], in which apart from the block-level motion vectors, pixel-wise motion vectors are also calculated in motion compensation process. This has been done by projection to spherical coordinates and depth calculation.

The above-mentioned methods improve the coding efficiency of the omnidirectional video by considering the near-real behavior of the motion by using the motion estimation and/or compensation operations in the spherical domain. However, since these operations are applied block-wise, the encoder and decoder complexities increase significantly. Thus, makes the deployment of such methods impractical in real-world scenarios. Furthermore, adapting the whole process of motion estimation and compensation of the current standards with the new approach requires significant changes in the standard chain of codecs.

III. Proposed Geometry-based Motion Vector Scaling Algorithm

As mentioned in Section I, the equirectangular projection of 360° video results in region-wise oversamplings and deformations in the projection plane. Such characteristics in the content introduce issues in the compression process using the current video coding standards.

An example of this issue is illustrated in Figure 2. As can be seen, in blocks near the polar areas of the ERP video, the motion vector behavior changes a lot. For example in the northern areas (Figure 2a), the motion vector of current block is smaller than the above blocks which are closer to northern pole and include higher deformations. Moreover, the motion vectors of the below blocks in this area may have smaller motion vectors due to less deformations compared to the current block. Similar phenomenon can be observed for the blocks in southern polar areas (Figure 2b) in which the neighboring blocks motion vector behavior varies compared to the current block depending on the location in latitude of the 360° video. This motion vector variation is more severe if the content has high motion. Such behavior of motion vectors result in larger motion vector difference (MVD) between the current block and neighbor blocks (that are used as motion vector predictors) and consequently, bitrate increase for MVDs.

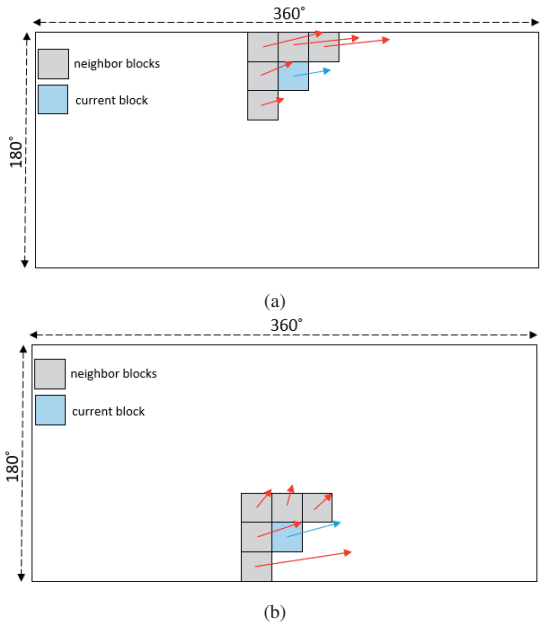


Fig. 2: Motion behavior of ERP video for blocks in the areas of a) north pole, b) south pole

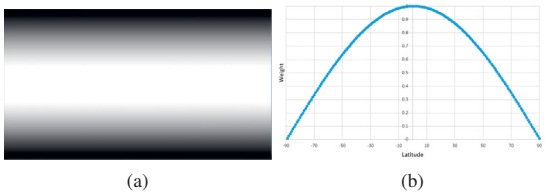


Fig. 3: Weight map in ERP

A. Geometry-based Motion Vector Scaling

In this section, a method is proposed for scaling the neighboring motion vectors in order to make them suitable for predicting the motion vectors of the current block. This method considers the sampling density behavior in different locations of the ERP image. The sampling density distribution of the ERP content is shown in Figure 3a. Such characteristics is used for weighted quality assessment (WS-PSNR) of 360° content [9]. According to WS-PSNR scheme, for each pixel in the ERP domain, a weight depending on the latitude coordinate is assigned in a way that the samples in the polar areas have lower weights than the samples in equatorial areas. The weight

derivation function is as below:

$$W[x,y] = cos(\frac{y - \frac{h}{2} + 0.5}{H} \times \pi) \quad Where: \begin{cases} 0 \leq x < w \\ 0 \leq y < h \end{cases} \tag{1}$$

In this equation, *W* is the calculated weight of the [x, y] pixel location in ERP and h and w are the height and width of the projection plane, respectively. The derived weights are illustrated in Figure 3b. Based on this representation, the equatorial areas include higher weights than the polar areas.

This formula can simulate the motion vector magnitude behaviors in different parts of the 360° video, in which the blocks that are closer to the polar areas have large motion vectors compared to their neighbors (as also shown in Figure 1). Therefore, this work takes advantage of weight map of ERP in order to calculate a scaling factor (SF) for scaling the neighboring motion vectors. In order to use such approach in block-level, the weights are derived according to the center location of each block by using (1) rather than the pixel-level weights. The reason being that the motion vectors are calculated in relation to the center points of blocks. The scaling factor is calculated as below:

$$SF = MaxWeight - (W_C - W_N) \tag{2}$$

In this formula, W_C and W_N are the calculated weights for current and neighbor blocks, respectively. *MaxWeight* represents the maximum weight in the ERP according to WS-PSNR [9] that is equal to 1. The calculated scaling factor is then applied to the horizontal and vertical components of the neighboring motion vectors with using equation (3)

$$\vec{MV}_{scaled} = round(\vec{MV} \times SF) \tag{3}$$

By using the described method, the motion vectors of the neighbor blocks are scaled up or down in a way that align with the central position of the current block. This results a more uniform motion behavior in a certain area of the 360° ERP video and smaller required MVD between the motion vector of current block and its neighbor. The scaled motion vectors are used as substitutes of the default unscaled candidates in AMVP and/or Merge lists.

The advantages of using the proposed geometry-based motion vector scaling method can be summarized as below:

- Efficient prediction of motion vectors particularly in the sequences with high motion.
- No compression loss for the content with stationary behavior (as the results in Section IV illustrate).
- Negligible complexity overhead.
- No extra signalling in the bitstream since the scaling process is applied in both encoder and decoder.
- Unlike the methods that are explained in Section II, the required changes in the video coding algorithm are quite simple and local.

IV. Experiments

A. Experimental Condition

The proposed motion vector scaling technique is implemented on top of the Versatile Video Coding (H.266/VVC) standard [2] test model VTM version 1.0 [10].

In order to evaluate the performance of the proposed method, 17 omnidirectional video clips in ERP format from JVET 360° test material [11] is used. The testset include video sequences in 4K, 6K, and 8K resolutions, each consisting 10 seconds i.e., equal to 300 and 600 frames for 30 fps and 60 fps sequences, respectively. The test materials are divided into two categories. The first category consists of sequences with high motion (e.g., camera/scene motion) and the second category contains content with lower motion.

The experiments conducted based on the proposed processing chain of Joint Video Experts Team (JVET) common test condition for 360° video [11]. According to the proposed processing chain, the high fidelity ERP sequences were downsampled to lower resolution versions (i.e., the coding resolution) prior to encoding process. The quality of the decoded content were assessed with different objective quality metrics proposed in the processing chain. Out of those, we used high fidelity ERP-PSNR, CPP-PSNR and WS-PSNR for quality evaluations. All the pre- and post-processing operations (e.g., resamplings, quality assessments, etc.) in the discussed processing chain are done by using the 360Lib software provided by JVET community [12]. Main profile Random Access (RA) configuration of common test condition [13] is used for encoding the test data. Moreover, the quantization parameters (QPs) are set to 22, 27, 32 and 37.

The performances were analyzed by using the well-known *Bjøntegaard* Delta Bitrate (BDBR) criterion [14] for luma and chroma pictures. The negative values are the indications of how much the bitrate is decreased in the same peak signal-to-noise ratio (PSNR). Similarly, the positive values show the bitrate increment for the same quality level.

B. Analysis of the Results

Table I presents the results of the proposed motion vector scaling method when it is applied in both AMVP and Merge coding tools compared to the reference in different quality metrics. As can be seen from the table, the proposed method provided consistent bitrate reduction for the high motion category on average by 1% for luma and 0.8% for chroma components.

In the low motion category of sequences, the proposed method does not have any impact in bitrate reduction or increase. This is an expected performance behavior, since the proposed method is designed for reducing the bitrate of the sequences that the motion behavior is in a way that the conventional prediction methods are not capable of modeling it. Furthermore, as the results illustrate, similar rate-distortion performances were observed in all the objective quality metrics that are used for assessing the quality of the compressed omnidirectional video.

TABLE I: BD-Rate (%) performance of the proposed method for sequences with high and low motion

Category	Sequence	ERP PSNR			CPP-PSNR			WS-PSNR		
		Y	U	V	Y	U	V	Y	U	V
High Motion	ChairliftRide	-2.2	-2.0	-1.9	-2.1	-1.9	-1.8	-2.1	-2.0	-1.8
	Skateboard	-0.5	-0.4	0.0	-0.4	-0.4	-0.1	-0.4	-0.4	-0.1
	Balboa	-0.7	-0.3	-0.6	-0.7	-0.4	-0.8	-0.7	-0.4	-0.8
	BranCastle2	-0.6	-0.3	-0.4	-0.5	-0.3	-0.4	-0.5	-0.3	-0.4
	Landing2	-0.4	-0.1	-0.1	-0.4	-0.1	-0.2	-0.4	-0.1	-0.2
	DrivingInCountry	-1.7	-1.7	-1.6	-1.7	-1.7	-1.5	-1.7	-1.7	-1.5
	Bicyclist	-0.7	-0.4	-0.4	-0.7	-0.3	-0.4	-0.7	-0.3	-0.4
	Glacier	-2.1	-1.4	-1.5	-2.0	-1.4	-1.7	-2.0	-1.5	-1.7
	Building	-0.6	-0.4	-0.5	-0.6	-0.3	-0.5	-0.6	-0.3	-0.5
Low Motion	Gaslamp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Trolley	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	KiteFlite	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Harbor	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1
	Broadway	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1
	AerialCity	-0.1	0.0	0.0	-0.1	0.0	0.0	-0.1	0.0	0.0
	DrivingInCity	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.2	0.0
	Paramotor	-0.1	-0.1	-0.2	-0.1	-0.1	-0.2	-0.1	-0.1	-0.2
	High Motion Average	**-1.0**	**-0.8**	**-0.8**	**-1.0**	**-0.8**	**-0.8**	**-1.0**	**-0.8**	**-0.8**
	Low Motion Average	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
	Overall	**-0.6**	**-0.4**	**-0.4**	**-0.6**	**-0.4**	**-0.4**	**-0.6**	**-0.4**	**-0.4**

TABLE II: Average encoding and decoding complexities (%) of the proposed method compared to reference

Category	High motion	Low motion
Encoding complexity	99.1%	100.6%
Decoding complexity	101.5%	101.3%

General observation from the results is that, the geometry-based motion vector scaling method provides better bitrate reduction for the sequences that have higher global motion particularly in the polar areas of the 360° scene. For example, *ChairliftRide*, *Glacier* and *DrivingInCountry* sequences have higher global motion and as a result the proposed method performs better (around 2% bitrate reduction) in these sequences compared to other content.

Table II illustrates the average complexities in terms of encoding and decoding runtimes when the proposed method is applied compared to the reference. As it can be observed, the proposed method has negligible impact in the encoder/decoder complexity overhead for both categories of testsets.

V. Conclusion

In this work, a geometry-based motion vector scaling technique was proposed for handling the issue of large and non-uniform motion behavior of the equirectangular projected 360° content. The proposed method, based on the location of the block in the 360° image plane, derives a scaling factor between the current and neighbor coding blocks and applies a scaling method to the neighboring motion vectors that are used as predictors. The proposed motion vector scaling technique provided consistent gain in the sequences with higher motion by up to 2.2% bitrate reduction in the best cases and on average by 1% and 0.8% for luma and chroma components, respectively. Moreover, the proposed method had no impact in the rate-distortion performance of the stationary sequences. From the complexity point of view, this method had negligible impact in the encoding and decoding runtimes in all test sequences.

References

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, *et al.*, "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.

[2] B. Bross, "Versatile Video Coding (VVC) draft 1," *MPEG Joint Video Exploration Team*, vol. JVET-J1001-v21, Apr. 2018.

[3] I. Tosic, I. Bogdanova, P. Frossard, and P. Vandergheynst, "Multiresolution motion estimation for omnidirectional images," in *Signal Processing Conference, 2005 13th European*, pp. 1–4, IEEE, 2005.

[4] L. Li, Z. Li, M. Budagavi, and H. Li, "Projection based advanced motion model for cubic mapping for 360-degree video," *arXiv preprint arXiv:1702.06277*, 2017.

[5] F. De Simone, P. Frossard, N. Birkbeck, and B. Adsumilli, "Deformable block-based motion estimation in omnidirectional image sequences," in *19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, 2017.

[6] B. Vishwanath, T. Nanjundaswamy, and K. Rose, "Rotational motion model for temporal prediction in 360 video coding," in *19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, 2017.

[7] B. Vishwanath, , K. Rose, Y. He, and Y. Ye, "Rotational motion compensated prediction in HEVC based omnidirectional video coding," in *Picture Coding Symposium (PCS)*, IEEE, 2018.

[8] Y. Wang, L. Li, D. Liu, F. Wu, and W. Gao, "A new motion model for panoramic video coding," in *International Conference on Image Processing (ICIP)*, pp. 1407–1411, IEEE, 2017.

[9] S. Yule, A. Lu, and Y. Lu, "WS-PSNR for 360 video objective quality evaluation," *MPEG Joint Video Exploration Team*, vol. 116, 2016.

[10] "Versatile Video Coding (VVC) reference software VTM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute." https://jvet.hhi.fraunhofer.de/, July 2018.

[11] E. Alshina, J. Boyce, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360 degree video," *JVETG1030, m41362, Aug*, 2017.

[12] Y. Ye, E. Alshina, and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib," *Joint Video Exploration Team of ITU-T SG*, vol. 16, 2017.

[13] J. Boyce, K. Suehring, X. Li, and V. Seregin, "Common test conditions and software reference configurations," in *Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11*, 2018.

[14] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," *VCEG-M33*, 2001.

PUBLICATION
II

Adaptive motion vector prediction for omnidirectional video
R. Ghaznavi-Youvalari and A. Aminlou

*IEEE Visual Communications and Image Processing (VCIP)*2018, 1–4
DOI: 10.1109/VCIP.2018.8698614

Adaptive Motion Vector Prediction for Omnidirectional Video

Ramin Ghaznavi-Youvalari, Alireza Aminlou

Nokia Technologies, Tampere, Finland
Email: firstname.lastname@nokia.com

***Abstract*—Omnidirectional video is widely used in virtual reality applications in order to create the immersive experience to the user. Such content is projected onto a 2D image plane in order to make it suitable for compression purposes by using current standard codecs. However, the resulted projected video contains deformations mainly due to the oversampling of the projection plane. These deformations are not favorable for the motion models that are used in the recent video compression standards. Hence, omnidirectional video is not efficiently compressible with the current codecs. In this work, an adaptive motion vector prediction method is proposed for efficiently coding the motion information of such content. The proposed method adaptively models the motion vectors of the coding block based on the motion information of the neighboring blocks and calculates a more optimal motion vector predictor for coding the motion information. The experimented results showed that the proposed motion vector prediction method provides up to 2.2% bitrate reduction in the content with high motion and on average 1.1% bitrate reduction for the tested sequences.**

***Index Terms*—Video coding, HEVC, 360° video, Motion model**

I. Introduction

Virtual reality (VR) applications make use of 360° spherical content in order to create the immersive experience to the users. For compression considerations, the spherical content is projected onto a two-dimensional (2D) image plane in order to make it suitable for the current standard compression algorithms (e.g., High Efficiency Video Coding (HEVC)) which operate only on 2D representations of the content.

One of the most popular projection formats for the omnidirectional content is equirectangular projection format (ERP) because of its ease of use and wide support in software development environments. However, the resulted projected 2D image suffers from content deformations caused by oversampling. The deformation is larger particularly in the polar areas of the ERP image compared to the equator areas. As a consequence of such deformations, the current compression algorithms are not capable of efficiently modeling this kind of behavior in the motion estimation and compensation processes. Consequently, the magnitude and direction of the motion vector of the current block could change a lot compared to the neighbor blocks.

Figure 1 illustrates an example of such motion behavior in different regions of the ERP video. As can be seen, the magnitude and direction of the motion vectors of blocks in a certain area are changing. This is more severe particularly in the blocks that are closer to the polar areas. As a result, the motion vector difference between the current block and neighbor block would be large and this leads to increase in bitrate of such content.

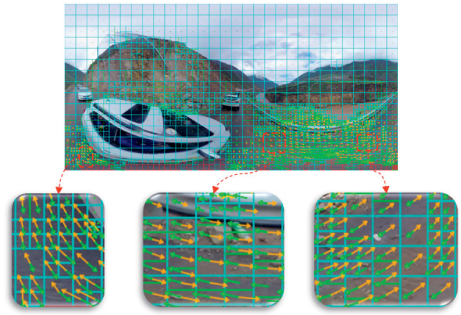


Fig. 1: Region-wise motion behavior in ERP

Recent studies address this issue by investigating projection-specific tools for omnidirectional content. A multiresolution motion estimation method is studied in [1], where the motion estimation for the omnidirectional content is considered in spherical domain. In [2], a translational motion model is proposed for cubemap projection formats which applies the motion estimation operation in the 3D spherical domain by projecting the current and reference blocks to the sphere. De Simone et. al. in [3] studied a method that considers the motion estimation and block matching processes in the spherical coordinates by using 8×8 block sizes only for reducing the complexity overheads.

Rotational motion model is studied in [4] where the motion estimation and compensation operations are done in the 3D spherical domain. Moreover, the authors proposed a radial pattern search for this operation in the spherical domain. Another motion model is introduced in [5], in which apart from the block-level motion vectors, pixel-wise motion vectors are also calculated in motion compensation process. This has been done by projection to spherical coordinates and depth calculation.

The above-mentioned methods improve the coding efficiency of the omnidirectional video by considering the near-

real behavior of the motion by using the motion estimation and/or compensation operations in the spherical domain. However, since these operations are applied block-wise, the encoder and decoder complexities increase significantly. Thus, makes the deployment of such methods impractical in real-world scenarios. Furthermore, adapting the whole process of motion estimation and compensation of the current standards with the new approach requires significant changes in the standard chain of the codec.

In this work, an adaptive method is proposed for modeling the motion vectors of each block, based on the motion information of the spatially neighboring blocks. The proposed method uses linear regression algorithm for extracting the parameters of the adaptive motion model. A new motion vector predictor (MVP) is calculated with the proposed adaptive motion model and included in the motion vector prediction candidate lists as an additional candidate in the advanced motion vector prediction (AMVP) and Merge coding tools. The conducted experiments illustrated that the proposed adaptive motion vector prediction method provides up to 2.2% bitrate reduction in the sequences with high motion. Moreover, there is no bitrate increase in the stationary sequences when the proposed method is applied.

The remainder of the paper is organized as follows. The proposed adaptive motion vector prediction is described in Section II. Section III discusses the experimental results. Finally, Section IV provides the conclusion of the work.

II. Proposed Adaptive Motion Vector Prediction

The conventional way of coding the motion vectors using the spatial/temporal candidates is not efficient in the content with below characteristics:

- The video include object/scene deformations due to image format and different sampling in various regions when representing 360° video on 2D representation formats e.g., ERP and cubemap projections.
- The video content has zooming and/or rotation caused by object and/or camera movement.
- Deformations in the video caused by the characteristics of the capturing device e.g., fisheye lenses.

We noted that, in such cases, the magnitude and direction of motion vectors change gradually by the location of the block, while the conventional motion vector prediction methods are based on the assumption that the motions of the neighboring block(s) are very close or identical to the current block. Accordingly, the motion vector prediction tools such as AMVP and Merge that are using motion vector information of the neighbor blocks for coding the motion vectors of the current block, are not able to predict the motion information efficiently. Such issues result in very large motion vector difference (MVD) and consequently requires more bits for coding it. Moreover, for the cases of very high object motion, intra prediction is usually selected by the encoder instead of inter prediction which results in higher bitrate.

The main idea of this paper is to locally model the motion vector of a block based on its location in a region of a frame. The center point of each block is considered as the location of the block. The parameters of the model for each block is calculated based on the motion vector and location of the neighboring blocks. Then, the motion vector of the current block is estimated using the extracted model and the location of the current block. The calculated adaptive motion vector is added to AMVP and/or Merge lists as an additional motion vector prediction candidate which can be selected by the encoder.

A. Adaptive Motion Model

In this work, an Adaptive Motion Model (AMM) method is considered for modeling the motion of the ERP video. It is assumed that the motion of each block in a video frame is modeled with a function that relates the motion vector of a block to its location (e.g., x and y coordinates of the center of the block) and the motion behavior of spatially neighboring blocks. For that, linear cross component model (LCCM) of equation (1) is used for the AMM method.

$$\begin{bmatrix} MV_{\mathrm{x}} \\ MV_{\mathrm{y}} \end{bmatrix} = \begin{bmatrix} a_{\mathrm{xx}} & a_{\mathrm{xy}} \\ a_{\mathrm{yx}} & a_{\mathrm{yy}} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} b_{\mathrm{x}} \\ b_{\mathrm{y}} \end{bmatrix} \tag{1}$$

In this equation, $\mathrm{MV_x}$ and $\mathrm{MV_y}$ are the horizontal and vertical components of motion vector of a block, respectively. X and Y parameters represent the center location of the current block. The remaining parameters (i.e., $\mathrm{a_{xx}}$, $\mathrm{a_{xy}}$, $\mathrm{a_{yx}}$, $\mathrm{a_{yy}}$, $\mathrm{b_x}$, $\mathrm{b_y}$) are calculated based on the motion information of the neighboring blocks that are used as training dataset for modeling the motion behavior of the current block. This is done based on a mean square error (MSE) minimization with the collected training dataset.

As this function is used for modeling the motion only in a small region of the video, linear model should efficiently work for this purpose.

B. Parameter Extraction for AMM

The adaptive motion model in (1) consists of 6 parameters that are calculated for each block using the motion information of the neighboring blocks. For this purpose, motion vectors and center locations of the neighboring blocks are collected and a linear regression method with MSE is applied to calculate the parameters. The size of the neighboring block is also considered in training process. For example, the information of the larger blocks has more influence on the model.

In the HEVC standard, motion information is stored in 4×4 sub-block accuracy. For the training process, the motion vectors and their corresponding locations are collected in 4×4 sub-block level. Figure 2 illustrates the training dataset collection process from neighboring blocks. In this case, motion vector of a block, when it is larger than 4×4, is considered several times in the training process. This is a welcomed property as the influence of the motion vector of each block in model extraction becomes proportional to the size of that block. For the sake of keeping the motion model behavior locally, the motion search area for training dataset is limited to twice of the size of coding block from each

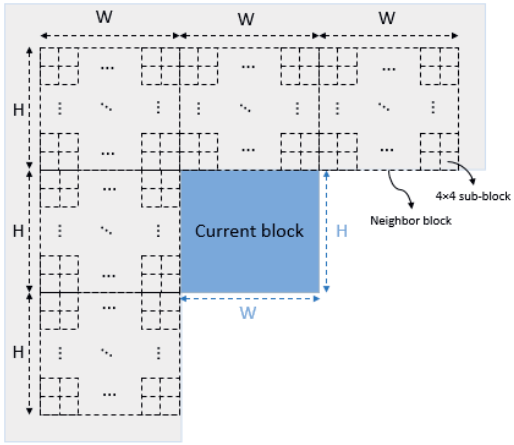


Fig. 2: Illustration of the neighboring motion information in 4×4 sub-block level that are considered for adaptive motion model training

neighboring side. The blocks in above, left, above-left, above-right and bottom-left are considered for extracting the training set.

C. Neighboring Motion Vector Scaling

In the HEVC standard, each inter-coded frame has several reference frames. Thus, each block may be predicted from different reference frames. Because of the motion in temporal domain, a block may have different motion vectors for different reference frames. This issue should be carefully considered in training and motion vector prediction processes as discussed below.

In the case of AMVP coding tool, motion vector of a block is predicted for a given reference frame with specific Picture Order Count (POC). If the neighboring blocks are predicted from different reference frame(s) than the reference frame of the current block, their motion information is scaled according to the POC number. For the Merge coding tool, in the training dataset collection process from neighboring blocks, the priority is given to the motion information that have the same POC number as the motion information of the current block. In the case that none of the neighboring blocks have the same POC as the the current block, their motion vectors are scaled according to the POC number of the motion information of the current block and are used as training dataset for calculating the adaptive motion vector prediction candidate.

The whole process of the adaptive motion vector prediction method can be summarized as below for each block:

1) Collect the MV and location of the neighboring blocks (i.e., the training dataset).
2) Calculate the parameters of the adaptive motion model using the training dataset information.
3) Calculate the adaptive motion vector predictor with using equation (1) with the parameters of previous step and the center location of the current block.
4) Apply redundancy check for the calculated adaptive motion vector. The new MV candidate is discarded if it is equal to one of the available MV candidates in the corresponding list of the coding tool (i.e., AMVP, Merge).
5) Add and prioritize the adaptive motion vector candidate by inserting it as the first prediction candidate in the AMVP and/or Merge list.

D. Advantages of the Proposed Method

The advantages of the described AMM method can be summarized as follows:

- Unlike the methods that are explained in Section I, the required changes in the video coding algorithm using the AMM method are quite local. This method only needs adding a new MV candidate to the AMVP and/or Merge coding lists, and no changes are required in bitstream syntax.
- The AMM model is simple, generic and locally adaptive. It can model different changes in motion or deformation in objects in different ways. For example it can support zooming in and out, rotation, and object deformation for example in 360° video formats.

III. Experiments

A. Experimental Conditions

The HEVC reference software (HM) version 16.16 [6] was used for implementing the proposed method. In order to evaluate the performance of the proposed method, 13 omnidirectional video clips in ERP format from Joint Video Experts Team (JVET) 360° test material [6] is used. The test material consists of video clips in 4K and 8K resolution, each consisting 10 seconds i.e., equal to 300 and 600 frames for 30 fps and 60 fps sequences, respectively

The experiments conducted based on the proposed processing chain of JVET common test condition for 360° video [7]. According to the proposed processing chain, the high fidelity ERP video is downsampled to lower resolution version (i.e., the coding resolution) prior to encoding process. The quality of the decoded video is assessed by three of the objective quality metrics proposed in the processing chain, i.e., S-PSNR-I, CPP-PSNR, and WS-PSNR. All the pre- and post-processing operations (e.g., resamplings, quality assessments, etc.) in the discussed processing chain are done by using the 360Lib software provided by JVET community [8]. Main profile Random Access (RA) configuration of common test condition [9] is used for encoding the test data. Moreover, the quantization parameters (QPs) are set to 22, 27, 32 and 37.

The performances were analyzed by using the well-known *Bjøntegaard* Delta Bitrate (BDBR) criterion [10] for luma pictures. The negative values are the indications of how much the bitrate is decreased in the same peak signal-to-noise

TABLE I: BD-Rate(%) results of the proposed method in AMVP, Merge, and overall (AMVP + Merge) coding tools under different quality metrics

Sequence	AMVP Coding Tool			Merge Coding Tool			Overall (AMVP + Merge)		
	S-PSNR-I	CPP-PSNR	WS-PSNR	S-PSNR-I	CPP-PSNR	WS-PSNR	S-PSNR-I	CPP-PSNR	WS-PSNR
Trolley	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
AerialCity	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	-0.1%	0.0%	0.0%	0.0%
DrivingInCity	0.1%	0.1%	0.1%	-0.2%	-0.2%	-0.2%	0.1%	0.1%	0.1%
DrivingInCountry	-2.0%	-2.0%	-2.0%	-1.0%	-1.0%	-1.0%	-2.2%	-2.2%	-2.2%
ChairliftRide	-1.4%	-1.4%	-1.4%	-1.0%	-1.0%	-1.0%	-1.8%	-1.8%	-1.8%
skateboard_in_lot	-0.4%	-0.4%	-0.4%	-0.5%	-0.4%	-0.4%	-0.4%	-0.4%	-0.4%
Balboa	-0.9%	-0.9%	-0.9%	-1.7%	-1.6%	-1.6%	-1.8%	-1.8%	-1.8%
BranCastle	-1.1%	-1.1%	-1.1%	-1.5%	-1.5%	-1.5%	-1.9%	-1.9%	-1.9%
Landing	-1.6%	-1.6%	-1.6%	-1.3%	-1.3%	-1.3%	-1.8%	-1.9%	-1.9%
Broadway	-0.5%	-0.5%	-0.5%	-1.7%	-1.7%	-1.7%	-1.7%	-1.7%	-1.7%
Bicyclist	-1.3%	-1.3%	-1.3%	-0.5%	-0.5%	-0.5%	-1.1%	-1.1%	-1.1%
Glacier	-2.4%	-2.4%	-2.4%	-1.0%	-1.0%	-1.0%	-2.2%	-2.2%	-2.2%
Paramotor	0.0%	0.0%	0.0%	-0.2%	-0.2%	-0.2%	-0.1%	-0.1%	-0.1%
Average	**-0.9%**	**-0.9%**	**-0.9%**	**-0.8%**	**-0.8%**	**-0.8%**	**-1.1%**	**-1.1%**	**-1.1%**

ratio (PSNR). Similarly, the positive values show the bitrate increment for the same quality level.

B. Analysis of the Results

Table I demonstrates the BD-Rate performance of the proposed adaptive motion vector prediction method of Section II in the case that it has been applied to AMVP, Merge and both AMVP and Merge coding tools.

As can be observed from the table, the proposed method provides consistent gain when it is applied to different coding tools, regardless of the utilized quality metrics. The adaptive motion vector prediction tool brings almost similar bitrate saving when it is applied to both tools separately, which is on average 0.9% for the AMVP tool and 0.8% in the case of Merge tool. The performance is slightly higher when the proposed method is integrated in both tools, on average around 1.1% gain, as it is illustrated in the table.

Furthermore, the performance is higher in the sequences that have global motion, whereas in the stationary sequences the proposed method does not have any impact. For example, *DrivingInCountry*, *ChairliftRide* and *Glacier* sequences include higher motion and consequently the proposed method have more impact on them with around 2% bitrate saving. This performance difference illustrates that the adaptive motion vector prediction is able to model the unusual motion in the polar areas efficiently. On the other hand, in the *Trolley* and *AerialCity* sequences, the additional new motion vector does not have any impact in the rate-distortion performance. This is an expected performance behavior since the model is designed to handle the motion behavior caused by the projection deformations of the ERP video and no significant loss when the content is stationary.

Moreover, the proposed adaptive motion vector prediction method uses limited amount of dataset for the training process, hence the complexity impact is not huge when compared to the previous motion estimation and compensation methods in [1]–[5]. In the conducted experiments, on average around 3% to 5% and 10% to 12% computational overheads were observed for the encoder and decoder, respectively.

IV. Conclusion

In this work, we proposed an adaptive motion vector prediction method for efficiently coding the motion information of the equirectangular projected video. The proposed method models the motion of the coding block based on the motion information and locations of the already coded neighboring blocks. A linear regression approach is used for modeling the motion vectors. The conducted experiments with the adaptive model demonstrated gains of up to 2.2% for the sequences with high motion and no compression loss in the stationary sequences. Moreover, the average bitrate reduction over the tested sequences was observed to be around 1.1%.

References

[1] I. Tosic, I. Bogdanova, P. Frossard, and P. Vandergheynst, "Multiresolution motion estimation for omnidirectional images," in *Signal Processing Conference, 2005 13th European*, pp. 1–4, IEEE, 2005.

[2] L. Li, Z. Li, M. Budagavi, and H. Li, "Projection based advanced motion model for cubic mapping for 360-degree video," *arXiv preprint arXiv:1702.06277*, 2017.

[3] F. De Simone, P. Frossard, N. Birkbeck, and B. Adsumilli, "Deformable block-based motion estimation in omnidirectional image sequences," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*, pp. 1–6, IEEE, 2017.

[4] B. Vishwanath, T. Nanjundaswamy, and K. Rose, "Rotational motion model for temporal prediction in 360 video coding," in *Multimedia Signal Processing (MMSP), 2017 IEEE 19th International Workshop on*, pp. 1–6, IEEE, 2017.

[5] Y. Wang, L. Li, D. Liu, F. Wu, and W. Gao, "A new motion model for panoramic video coding," in *Image Processing (ICIP), 2017 IEEE International Conference on*, pp. 1407–1411, IEEE, 2017.

[6] "High Efficiency Video Coding (HEVC) reference software HM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute." https://hevc.hhi.fraunhofer.de/., April 2018.

[7] E. Alshina, J. Boyce, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360 degree video," *JVETG1030, m41362, Aug*, 2017.

[8] Y. Ye, E. Alshina, and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib," *Joint Video Exploration Team of ITU-T SG*, vol. 16, 2017.

[9] F. Bossen, "Common test conditions and software reference configurations," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 5th meeting, Jan. 2011*, 2011.

[10] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," *VCEG-M33*, 2001.

PUBLICATION III

Regression-based motion vector field for video coding

R. Ghaznavi-Youvalari, A. Aminlou and J. Lainema

IEEE Transactions on Circuits and Systems for Video Technology (2019)
DOI: 10.1109/TCSVT.2019.2942086

Regression-based Motion Vector Field for Video Coding

Ramin Ghaznavi-Youvalari, *Member, IEEE,* Alireza Aminlou, and Jani Lainema

***Abstract*—In this paper, we study a method for compensating the non-translational motion behavior in video coding. The proposed method models the motion field of a prediction block based on the motion information of the neighboring blocks by using a linear regression approach. In order to provide a finer granularity of motion vectors the Regression-based Motion Vector Field (RMVF) method derives the motion field in 4×4 sub-block accuracy. Such approach generates a smooth and more realistic motion vector field inside the prediction block. The motion field generated with RMVF is then used as a new merge mode along with other merge modes in VTM-2.0 test model of the Versatile Video Coding (H.266/VVC) standard. The conducted experiments with JVET CTC sequences illustrate that the proposed RMVF method provides 0.77%, 0.19% and 0.41% bitrate reductions with random access (RA), low delay B (LDB) and low delay P (LDP) configurations, respectively. Furthermore, this method provides on average 0.65% bitrate saving for the 360° sequences in equirectangular projection format (ERP) with RA configuration.**

***Index Terms*—Video coding, motion model, affine motion, 360° video, linear regression, VVC.**

I. Introduction

The legacy video compression standards, such as Advanced Video Coding (H.264/AVC) [1] and High Efficiency Video Coding (H.265/HEVC) [2], use only translational motion compensation (TMC) for inter prediction purposes. Even though TMC is able to compensate the majority of the motion between frames, it is not capable of modeling more complicated motion behaviors such as zooming, rotation and shearing efficiently. Hence, more sophisticated methods are needed to be considered for modeling the non-translational motions for the next generations of video coding standards. Recently, the standardization work of a new video codec known as Versatile Video Coding (H.266/VVC) [3] has been started. The VVC standard is under development by Joint Video Experts Team (JVET) which is a collaborative team formed by ISO/IEC MPEG and ITU-T Study Group 16's VCEG. The aim of this work is to study new methods for further improving the motion estimation and compensation aspects of VVC.

Non-translational motion behavior in video content can occur for example due to camera zooming, rotating or moving, or it can be because of objects' motion in the content. The characteristics of capturing device itself may cause such properties in the resulted video content. For example, fisheye lenses that are used for capturing wide field-of-view (FOV) content cause certain distortions and non-linearities of motion in the content [4]. Another type of non-translational motion exists in 360° video content that are used for virtual reality (VR) applications. In such cases, the 360° FOV of spherical content is projected onto a two-dimensional (2D) image plane for example using equirectangular projection format (ERP) [5]. This 3D to 2D projection process creates content distortions due to the different sampling characteristics of the projection plane, particularly in the areas near the poles of the 360° content.

Affine Motion Compensation (AMC) method has been studied for more than thirty years for non-translational motions such as zooming, rotation, shearing, etc. Due to significant amount of computational complexity in both encoder and decoder, the AMC method was not considered as a feasible approach in previous standards. Over the past years, significant efforts have been made for simplifying AMC [6]–[9] in order to reduce the complexity while preserving the compression benefits. Finally, a simplified sub-block based AMC was adopted in the VVC standard [3] that operates on 4×4 sub-block accuracy and uses 4- and 6-parameter motion models. However, the simplified AMC is not capable of efficiently modeling all the non-translational motions. For example, AMC may not be an ideal approach for modeling the motion distortions in fisheye and 360° content.

For handling non-linearity of motion in fisheye content, an elastic motion model is used in [10]. That method uses 2D discrete cosine basis functions for modeling the motion. In another work [11], the motion estimation scheme is altered by applying an equisolid re-projection of each block in fisheye content with using fisheye to perspective projection.

Efficient motion modeling for 360° video have been widely studied in recent years. A rotational motion model is used in [12] that considers the motion estimation and compensation in 3D spherical domain. In [13], motion estimation and block matching processes are applied in spherical coordinates. The method in [14] generates uniform and efficient predictors for AMVP and Merge coding tools by scaling the neighboring motion vectors based on the projection geometry and the location of each block in the 360° scene.

The above methods provide improvements compared to the conventional translational motion compensation. However, these methods are sub-optimal for handling all the non-translational motions in general. In order to provide an efficient motion model for non-translational scenarios, we studied a method in [15] in which the motion of a block was modeled by using a linear regression approach in HEVC environment. The proposed method calculates a motion vector for an entire block based on the motion behavior of the neighboring blocks.

R. Ghaznavi-Youvalari, A. Aminlou, and J. Lainema are with Nokia Technologies, 33100 Tampere, Finland (e-mail: ramin.ghaznavi-youvalari@nokia.com; alireza.aminlou@nokia.com; jani.lainema@nokia.com)

The calculated motion vector is used as a new motion vector predictor candidate in Advance Motion Vector Prediction (AMVP) and Merge coding tools. However, this is not an efficient approach when it is used in VVC codec that consists of more advanced tools such as AMC.

In this paper, the regression-based motion model is used for deriving the motion field of a prediction block in 4×4 sub-block accuracy and hence referred to as Regression-based Motion Vector Field (RMVF). The proposed method relates the motion vector of each 4×4 sub-block of a larger block to its spatially neighboring 4×4 sub-block motion vectors and locations. In other words, the RMVF method models the motion of a block based on its neighboring motion vectors and locations of those in sub-block level. Furthermore, the RMVF method is used as a standalone merge mode in VVC among other merge tools such as affine merge. The experimented results illustrate that for JVET common test condition (CTC) sequences [16], the RMVF tool provides on average 0.77%, 0.19% and 0.41% bitrate reductions with random access (RA), low delay B (LDB) and low delay P (LDP) configurations, respectively. Moreover, this tool provides 0.65% bitrate improvement for the 360° CTC testset [17] with RA configuration.

The remainder of the paper is organized as follows. The proposed RMVF method is described in Section II. The performance of the RMVF method is studied in Section III. Finally, Section IV provides the conclusion of the work.

II. Proposed Regression-based Motion Vector Field

As mentioned in Section I, the regression-based motion vector prediction method that was studied in [15] may not always be an efficient approach for deriving the motion information of a block due to the fact that a single motion vector may not be able to represent the motion behavior of an entire block especially for large blocks. This is an important issue since the draft VVC standard [3] includes tools which provide the motion vector field of a block in sub-block level accuracy, such as 4×4 block based affine motion compensation and also alternative temporal motion vector prediction (ATMVP).

In order to generate a finer granularity of motion vector field, the Regression-based Motion Vector Field (RMVF) method is studied in this section. The aim of the RMVF method is to generate accurate sub-block motion field inside a block, based on the motion behavior of spatially neighboring blocks, so that it is able to estimate motion information in case of linear or non-linear motion changes in the neighborhood blocks such as rotation, zooming, geometry distortion or combination of these properties.

For this purpose, the RMVF method attempts to model the motion of a prediction block based on a 6-parameter motion model of equation (1).

$$\begin{bmatrix} MV_x \\ MV_y \end{bmatrix} = \begin{bmatrix} a_{\mathrm{x0}} & a_{\mathrm{x1}} & a_{\mathrm{x2}} \\ a_{\mathrm{y0}} & a_{\mathrm{y1}} & a_{\mathrm{y2}} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

In this motion model, MV_x and MV_y represent the horizontal and vertical motion vectors of each 4×4 sub-block inside the larger block, respectively, and x and y are the horizontal and vertical center locations of each 4×4 sub-block. Moreover, the remaining parameters (i.e., referred to as RMVF-parameters hereafter) of the motion model are the parameters that model the motion behavior in the neighborhood of that block. These parameters are calculated based on the motion information of the neighboring blocks by using a linear regression with mean square error (MSE) minimization approach.

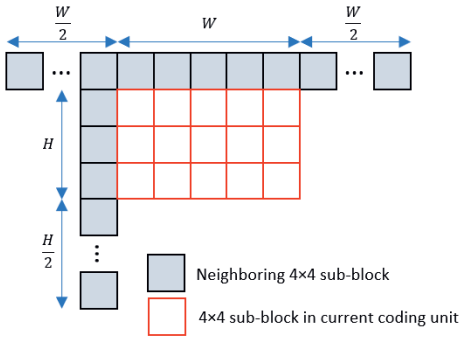


Fig. 1: Illustrates the neighboring 4×4 sub-blocks that are used for RMVF parameter derivation

In order to derive the RMVF-parameters of equation (1), certain amount of motion information are collected from the neighboring blocks. Figure 1 illustrates the utilized motion information for RMVF-parameter derivation. As can be seen from the figure, one column and row of the available neighboring motion vectors in 4×4 sub-blocks are collected along with their central (x, y) locations. The usage of available motion vectors from top-left, top-right and bottom-left corner areas are limited to half of the corresponding height or width of the current block. The reason for such limitations is that, since the RMVF attempts to model the motion locally for each block, using motion information from longer distances to the current block will result in inaccurate parameter derivation and consequently decreases the motion estimation and compensation performance.

The available 4×4 motion vectors and their center locations are used as an input to the linear regression operation for calculating the RMVF-parameters. For this purpose, it can be assumed that the motion model is defined as equation (1).

Parameters of model (1) can be estimated using MSE minimization method, for the N available neighboring MVs, where MSE is calculated as the average of the squared difference between the estimated values by model (1) and the actual values of the neighboring blocks MVs as defined in (2), for each horizontal and vertical component separately:

$$MSE_{\mathrm{c}} = \frac{1}{N} \sum_{k=0}^{N-1} (MV_{\mathrm{c}} - MV'_{\mathrm{c}})^2 \tag{2}$$

In (2), MV is the actual motion vector of the neighboring blocks, and MV′ is the estimated motion vector of the neigh-

boring block using the model in (1), and c indicates the horizontal or vertical component of motion vector.

In order to minimize the MSE, model parameters can be found using the following equations (3)-(4). In these equations, MV_c^k is the k-th neighboring motion vector, and ℓ is the center location (x, y) of the corresponding neighboring sub-block MV.

$$sumMV\ell_{c,d} = \sum_{k=0}^{N-1} (MV_c^k \times \ell_d^k) \tag{3}$$

$$sum\ell\ell_{i,j} = \sum_{k=0}^{N-1} (\ell_i^k \times \ell_j^k) \tag{4}$$

$$Where: \begin{cases} c \in \{x, y\} \\ d, i, j \in \{0, 1, 2\} \\ k \in \{0, 1, ..., N-1\} \\ \ell_0 : center\ location\ x\ of\ 4 \times 4\ block \\ \ell_1 : center\ location\ y\ of\ 4 \times 4\ block \\ \ell_2 = 1 \end{cases} \tag{5}$$

Using the above equations, 3×3 matrices A and B can be defined as below:

$$A_{i,j} = sum\ell\ell_{i,j} \tag{6}$$

$$B_{i,j}^{c,d} = \begin{cases} sumMV\ell_{c,d} & ; \ j = d \\ sum\ell\ell_{i,j} & ; \ Otherwise \end{cases} \tag{7}$$

The RMVF-parameters can then be derived by calculating and dividing the determinants of A and B matrices as below:

$$a_{c,d} = \frac{\det(B^{c,d})}{\det(A)} \tag{8}$$

The determinants of equation (8) are calculated according to equation (9) for A and B, separately:

$$\begin{aligned} \det(M) = (&M_{0,0} \times M_{1,1} \times M_{2,2} \\ &+ M_{1,0} \times M_{2,1} \times M_{0,2} \\ &+ M_{2,0} \times M_{0,1} \times M_{1,2}) \\ &- (M_{0,0} \times M_{2,1} \times M_{1,2} \\ &+ M_{1,0} \times M_{0,1} \times M_{2,2} \\ &+ M_{2,0} \times M_{1,1} \times M_{0,2}) \end{aligned} \tag{9}$$

Finally, the sub-block motion vectors of the current block can be calculated based on the derived RMVF-parameters in (8) and the motion model of (1) relative to the (x, y) location of each sub-block inside the current block.

The motion vector field generated with RMVF is then used as a separate merge mode in VTM-2.0 test model of VVC along with other merge modes (e.g. affine merge, regular merge, etc.). For motion compensation, RMVF mode uses the 4×4 sub-block motion compensation function of VTM-2.0 test model. Furthermore, in merge mode, a coding unit (CU) level RMVF flag is signalled into the bitstream in order to indicate the usage of RMVF mode for the CU.

III. Experimental Results

The proposed RMVF method is implemented on top of VTM-2.0 test model [18] of draft VVC standard [3]. The performance of the RMVF method was evaluated by using two distinct categories of testsets as below:

- JVET testset [16] that consists of 26 video sequences with different characteristics and resolutions.
- JVET 360° testset [17] that consists of 10 sequences in equirectangular projection format (ERP) with 8K resolutions.

Both testsets consist of video sequences with 10 second durations. The JVET CTC sequences were evaluated based on the common test condition [16] with RA, LDB and LDP configurations. The evaluation procedure of JVET 360° CTC [17], [19] was used for conducting the 360° tests with RA configuration, in which the high-fidelity ERP sequences were downsampled to a lower resolution versions (i.e., 4K coding resolution) prior to encoding process. The quality of the decoded 360° testset was evaluated by using the WS-PSNR method [17] of the 360° CTC process. The performances in the experiments were evaluated based on Bjontegaard Delta Bitrate (BDBR) criterion [20] for Luminance and Chrominance components.

Table I illustrates the performance of the proposed RMVF method for the JVET testset. As can be observed from the table, the RMVF method provides on average of all classes 0.77%, 0.70%, 0.90% bitrate reduction in Y, U and V components, respectively with RA configuration. The gains for LDB configuration are 0.19%, 0.19% and 0.29% for Y, U and V components, respectively. In the LDP configuration, the RMVF technique reduces the bitrate on average by 0.41%, 0.10% and 0.07% in different components.

The impact of RMVF method in 360° testset is shown in Table II. In this testset, the proposed method with RA configuration has reduced the bitrate on average by 0.65%, 0.73% and 0.74% in luma and chroma components, respectively.

In JVET testset, the RMVF method provides better performances in Class A1, A2 and B which contain sequences with camera motion, rotation and zooming. Results indicate that the RMVF method is capable of modeling such motion behavior properly. Especially high performance behavior can be seen in *Tango2*, *FoodMarket4*, *CatRobot1*, *DaylightRoad2* and *Cactus* sequences where the RMVF method was able to reduce the bitrate by more than 1%. On the other hand, the RMVF method does not bring high bitrate reduction when the content has low non-linear motion behavior. This can be observed in some sequences in Class D and Class F (which consists of screen content mainly). In these cases, the proposed method was not performing efficiently, particularly in LDB and LDP configurations where the RMVF method introduces bitrate loss due to signalling overhead of the CU level RMVF flag.

Similarly, in case of 360° testset, the RMVF method is capable of modeling the motion more efficiently in the sequences with non-linear motion due to the sampling properties of the ERP format, camera motion, etc. For example, such motion behavior exist in *ChairliftRide*, *Balboa* and *Broadway* se-

quences, where RMVF tool performs significantly better than VTM-2.0 test model. In the case of content with stationary and/or linear motion in this category, the RMVF tool does not provide significant improvements. For example in *Harbor360*, *Trolley* and *Gaslamp* sequences, the bitrate improvements are insignificant by using the RMVF method.

Another observation from Table I is that the performance of RMVF method is higher in RA configuration compared to LDB and LDP configurations. The reason for such performance behavior is that, in LDB and LDP cases the temporal distance of reference picture(s) is very close to the current picture, thus most of the motion in these cases are translational and TMC can compensate such motion efficiently. Furthermore, LDP configuration uses only uni-prediction for motion compensation, whereas bi-prediction is used in LDB case. Thus, considering the temporal distance and bi-prediction, the TMC of LDB configuration is capable of modeling the motion in a more efficient way than LDP. Consequently, the RMVF provides better performance in LDP configuration than LDB one.

The results of Tables I and II indicate that the affine motion compensation of VTM-2.0 is not capable of modeling and compensating all the non-translational motions. On the other hand, the RMVF scheme was able to provide further improvements to the cases where AMC fails to succeed.

In terms of complexity, the RMVF method does not impose significant encoding and decoding runtimes to the codec. As can be seen from Table I, the encoding overheads for JVET CTC testset were 6%, 9% and 10% and the decoding overheads were 5%, 3% and 3%, for RA, LDB and LDP configurations, respectively. Furthermore, as shown in Table II, the encoding and decoding runtime increases in 360° category were on average 7% and 5% for RA configuration, respectively.

Table III presents the coding statistics of RMVF, affine merge and normal merge modes in terms of average percentage of usage of each tool in the merge mode for JVET testset with RA configuration. As can be observed from the table, the normal merge tool has been selected most of the time compared to the other merge tools. The reason being that the majority of the motion in the sequences are translational motion, hence normal merge tool is capable of compensating such behavior. Moreover, the usage percentage of RMVF tool has been observed to be significantly higher compared to affine merge mode. The reason for such higher usage would be that the RMVF method uses all the available neighboring sub-block MVs for deriving the motion model, whereas the affine merge mode makes use of only neighboring corner MVs for this purpose. Hence, the RMVF tool can provide a more accurate motion model. Another observation is that the usage percentage of RMVF method is higher in the sequences with higher and/or non-translational motion. This is an expected behavior since the tool is designed in a way that can compensate non-translational motion more efficiently.

IV. Conclusion

This paper proposed a Regression-based Motion Vector Field (RMVF) method for compensating the non-translational motion in an efficient way. To that end, the RMVF tool uses a 6-parameter motion model for generating the motion field of a prediction block in 4×4 sub-block level. The RMVF-parameters were estimated locally for each block based on the neighboring motion vectors and their locations. The proposed method was able to model the motion behavior of blocks efficiently, especially in cases of zooming, rotation, shearing and geometry distortions. The conducted experiments illustrated that the RMVF method reduced the bitrate of the JVET CTC sequences on average by 0.77%, 0.19% and 0.41% for luma component in RA, LDB and LDP configurations, respectively. Furthermore, the proposed method was able to decrease the bitrate for 360° testset by 0.65% on average for luma component in RA configuration.

References

[1] *Advanced Video Coding. Document Rec. ITU-T H.264, ISO/IEC 14496-10 AVC*, May 2003.

[2] *High Efficiency Video Coding, Version 1. Document Rec. ITU-T H.265, ISO/IEC 23008-2*, Jan. 2013.

[3] B. Bross, "Versatile Video Coding (VVC) draft 2," MPEG Joint Video Exploration Team, document: JVET-K1001-v7, Jul. 2018.

[4] J. Wei, C.-F. Li, S.-M. Hu, R. R. Martin, and C.-L. Tai, "Fisheye video correction," IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 10, pp. 1771–1783, 2012.

[5] J. P. Snyder, "Flattening the earth: two thousand years of map projections," University of Chicago Press, 1997.

[6] K. Zhang, Y.-W. Chen, L. Zhang, W.-J. Chien, and M. Karczewicz, "An improved framework of affine motion compensation in video coding," IEEE Transactions on Image Processing, vol. 28, no. 3, pp. 1456–1469, 2019.

[7] L. Li, H. Li, D. Liu, Z. Li, H. Yang, S. Lin, H. Chen, and F. Wu, "An efficient four-parameter affine motion model for video coding," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 8, pp. 1934–1948, 2018.

[8] T. Wiegand, E. Steinbach, and B. Girod, "Affine multipicture motion-compensated prediction," IEEE transactions on circuits and systems for video technology, vol. 15, no. 2, pp. 197–209, 2005.

[9] X. Li, J. R. Jackson, A. K. Katsaggelos, and R. M. Merserau, "Multiple global affine motion model for H.264 video coding with low bit rate," in Image and Video Communications and Processing 2005, vol. 5685, pp. 185–195, International Society for Optics and Photonics, 2005.

[10] A. Ahmmed, M. M. Hannuksela, and M. Gabbouj, "Fisheye video coding using elastic motion compensated reference frames," in 2016 IEEE International Conference on Image Processing (ICIP), pp. 2027–2031, 2016.

[11] A. Eichenseer, M. Bätz, and A. Kaup, "Motion estimation for fisheye video with an application to temporal resolution enhancement," IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[12] B. Vishwanath, K. Rose, Y. He, and Y. Ye, "Rotational motion compensated prediction in hevc based omnidirectional video coding," in 2018 Picture Coding Symposium (PCS), pp. 323–327, 2018.

[13] F. De Simone, P. Frossard, N. Birkbeck, and B. Adsumilli, "Deformable block-based motion estimation in omnidirectional image sequences," in 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6, 2017.

[14] R. Ghaznavi-Youvalari and A. Aminlou, "Geometry-based motion vector scaling for omnidirectional video coding," in IEEE International Symposium on Multimedia (ISM), pp. 127–130, Dec. 2018.

[15] R. Ghaznavi-Youvalari and A. Aminlou, "Adaptive motion vector prediction for omnidirectional video," in 2018 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4, 2019.

[16] F. Bosson, J. Boyce, K. Suehring, X. Li, and V. Seregin, "JVET common test conditions and software reference configurations for SDR video," in Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Document: JVET-K1010-v2, July 2018.

[17] P. Hanhart, J. Boyce, and K. Choi, "JVET common test conditions and evaluation procedures for 360° video," Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 - Document JVET-K1012-v1, July 2018.

TABLE I: BD-Rate(%) results of RMVF tool for JVET CTC category over VTM-2.0

Class	Sequence	Random Access			Lowdelay B			Lowdelay P		
		Y	U	V	Y	U	V	Y	U	V
Class A1 (4K)	Tango2	-1.30%	-0.83%	-1.57%	-	-	-	-	-	-
	FoodMarket4	-1.58%	-1.32%	-1.53%	-	-	-	-	-	-
	Campfire	-0.09%	-0.05%	-0.22%	-	-	-	-	-	-
Class A2 (4K)	CatRobot1	-1.42%	-1.50%	-1.82%	-	-	-	-	-	-
	DaylightRoad2	-1.41%	-1.46%	-1.62%	-	-	-	-	-	-
	ParkRunning3	-0.52%	-0.63%	-0.65%	-	-	-	-	-	-
Class B (1080p)	MarketPlace	-0.79%	-0.51%	-1.04%	-0.52%	-0.47%	-0.20%	-0.55%	-0.25%	-0.19%
	RitualDance	-0.54%	-0.56%	-0.50%	-0.41%	-0.61%	-0.39%	-0.49%	-0.17%	-0.30%
	Cactus	-1.03%	-0.89%	-1.02%	-0.26%	-0.15%	-0.40%	-0.47%	-0.17%	-0.13%
	BasketballDrive	-0.65%	-0.75%	-0.68%	-0.47%	-0.48%	-0.59%	-0.57%	-0.41%	-0.58%
	BQTerrace	-0.55%	-0.53%	-0.47%	0.11%	0.36%	0.29%	-0.50%	1.82%	1.01%
Class C (WVGA)	BasketballDrill	-0.30%	-0.45%	-0.55%	-0.08%	-0.31%	-0.61%	-0.28%	-0.59%	0.44%
	BQMall	-0.58%	-0.17%	-0.98%	-0.08%	0.27%	-0.29%	-0.14%	-0.37%	-0.52%
	PartyScene	-0.58%	-0.44%	-0.76%	-0.01%	0.04%	-0.08%	-0.19%	0.11%	0.09%
	RaceHorses	-0.24%	-0.35%	-0.11%	-0.31%	-0.29%	-0.30%	-0.55%	-0.79%	-0.92%
Class D (WQVGA)	BasketballPass	-0.38%	-1.08%	-0.89%	-0.24%	-0.10%	0.15%	-0.31%	-0.24%	-0.76%
	BQSquare	-1.41%	-1.08%	-1.25%	0.19%	-0.39%	-1.24%	0.06%	1.89%	-1.31%
	BlowingBubbles	-0.69%	-0.29%	-0.58%	0.10%	0.86%	-0.03%	-0.05%	0.68%	0.04%
	RaceHorses	-0.26%	-0.56%	-0.81%	-0.33%	-0.45%	-0.51%	-0.52%	-0.47%	-0.09%
Class E (720p)	FourPeople	-0.44%	-0.41%	-0.52%	0.39%	-0.15%	-0.33%	-0.17%	-0.25%	-0.21%
	Johnny	-0.86%	-0.82%	-0.85%	-0.46%	-0.63%	-0.24%	-0.31%	0.36%	0.70%
	KristenAndSara	-0.70%	-0.67%	-0.59%	-0.21%	0.21%	-0.31%	-0.74%	-0.53%	-0.18%
Class F	BasketballDrillText	-0.21%	-0.39%	-0.36%	-0.06%	0.17%	0.26%	-0.33%	0.12%	-0.03%
	ArenaOfValor	-0.11%	-0.28%	-0.28%	0.06%	0.31%	0.13%	0.02%	-0.29%	-0.05%
	SlideEditing	-0.02%	-0.06%	-0.06%	-0.18%	0.27%	0.43%	0.84%	0.82%	0.87%
	SlideShow	0.01%	0.00%	0.01%	-0.45%	-0.31%	0.01%	-0.49%	-1.74%	2.56%
Overall Class A1		-0.99%	-0.73%	-1.11%	-	-	-	-	-	-
Overall Class A2		-1.12%	-1.20%	-1.37%	-	-	-	-	-	-
Overall Class B		-0.71%	-0.65%	-0.74%	-0.31%	-0.27%	-0.26%	-0.52%	0.17%	-0.04%
Overall Class C		-0.43%	-0.35%	-0.60%	-0.12%	-0.07%	-0.32%	-0.29%	-0.41%	-0.23%
Overall Class E		-0.67%	-0.64%	-0.65%	-0.09%	-0.19%	-0.29%	-0.41%	-0.14%	0.10%
Overall		**-0.77%**	**-0.70%**	**-0.90%**	**-0.19%**	**-0.19%**	**-0.29%**	**-0.41%**	**-0.10%**	**-0.07%**
Overall Class D		-0.69%	-0.75%	-0.88%	-0.07%	-0.02%	-0.41%	-0.20%	0.47%	-0.53%
Overall Class F		-0.08%	-0.18%	-0.17%	-0.16%	0.11%	0.21%	0.01%	-0.27%	0.84%
Encoding Time		106%			109%			110%		
Decoding Time		105%			103%			103%		

TABLE II: BD-Rate(%) results of RMVF tool for 360° category over VTM-2.0

Class	Sequence	Y	U	V
Class S1 (8K)	Skateboard	-0.59%	-1.30%	-1.34%
	ChairliftRide	-1.52%	-1.68%	-1.52%
	KiteFlite360	-0.10%	-0.14%	-0.24%
	Harbor360	-0.03%	-0.08%	-0.06%
	Trolley	-0.06%	-0.05%	-0.04%
	Gaslamp	-0.01%	0.02%	0.00%
Class S2 (8K)	Balboa	-1.58%	-1.76%	-1.46%
	Broadway	-1.37%	-1.26%	-1.37%
	Landing2	-0.54%	-0.56%	-0.85%
	BranCastle2	-0.70%	-0.47%	-0.48%
Overall Class S1		-0.39%	-0.54%	-0.53%
Overall Class S2		-1.05%	-1.01%	-1.04%
Overall		**-0.65%**	**-0.73%**	**-0.74%**
Encoding Time		107%		
Decoding Time		105%		

[18] "Versatile Video Coding (VVC) reference software VTM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute." https://jvet.hhi.fraunhofer.de/, Jan. 2019.

[19] Y. Ye and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib version 7," Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC29/WG 11 - Document: JVET-K1004, Jul. 2018.

[20] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," VCEG-M33, 2001.

TABLE III: Average usage percentage of different merge tools in merge mode of VTM-2.0 for JVET CTC category

Class	Sequence	Normal	Affine	RMVF
Class A1	Tango2	73.7%	1.5%	24.9%
	FoodMarket4	71.8%	0.9%	27.3%
	Campfire	91.6%	1.2%	7.3%
Class A2	CatRobot1	72.4%	3.5%	24.1%
	DaylightRoad2	65.4%	10.3%	24.4%
	ParkRunning3	80.7%	3.5%	15.8%
Class B	MarketPlace	59.0%	11.2%	29.8%
	RitualDance	78.6%	5.5%	15.9%
	Cactus	73.9%	4.6%	21.5%
	BasketballDrive	77.0%	3.9%	19.1%
	BQTerrace	86.4%	1.3%	12.3%
Class C	BasketballDrill	84.6%	2.5%	12.9%
	BQMall	87.5%	1.4%	11.1%
	PartyScene	80.4%	3.0%	16.6%
	RaceHorses	85.7%	3.1%	11.2%
Class D	BasketballPass	71.8%	2.2%	26.0%
	BQSquare	68.1%	5.6%	26.3%
	BlowingBubbles	72.0%	3.6%	24.5%
	RaceHorses	80.6%	3.0%	16.4%
Class E	FourPeople	68.1%	2.0%	29.9%
	Johnny	64.2%	5.2%	30.6%
	KristenAndSara	65.0%	4.5%	30.5%
Class F	BasketballDrillText	85.6%	2.4%	12.0%
	ArenaOfValor	86.3%	3.8%	9.9%
	SlideEditing	90.9%	0.1%	9.0%
	SlideShow	84.7%	0.8%	14.5%

Ramin Ghaznavi-Youvalari (M'16) received his M.S. degree in information technology from Tampere University of Technology (TUT), Tampere, Finland in 2016. He is currently pursuing the Ph.D. degree in signal processing at Tampere University, Tampere, Finland.

He was a research assistant at TUT, from 2015 to 2016, working on virtual reality content compression and streaming. He joined Nokia Technologies in 2016. Currently, he is a senior researcher at Media Technologies Research group of Nokia Technologies and his work is focused on various image and video compression and streaming fields including standardization of Versatile Video Coding (VVC/H.266). His research interests include image and video compression, virtual reality, augmented reality and machine learning.

Alireza Aminlou received his B.S. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 2000, and his M.S. and PhD degrees in electrical engineering from University of Tehran, Tehran, Iran, in 2003 and 2010, respectively.

He was with Multimedia Processing Laboratory, University of Tehran, from 2003 to 2010 in different projects including hardware implementation of JPEG2000 and H.264/AVC codecs. He was visiting researcher in Tampere University of Technology, Tampere, Finland in 2011. Since 2012, he has been with Nokia Research Center and Nokia Technologies, Tampere, Finland contributing to scalable extension of High Efficient Video Coding (HEVC), streaming of Virtual Reality (VR) content, and Versatile Video Coding (VVC) projects. His research interests include hardware implementation, video compression and rate-distortion optimization.

Jani Lainema received his M.Sc. degree in computer science from the Tampere University of Technology, Finland in 1996. He joined the Visual Communications Laboratory of Nokia Research Center in 1996. Since then he has contributed to the designs of ITU-T's and MPEG's video coding standards as well as to the evolution of different multimedia service standards in 3GPP, DVB and DLNA. He is a Bell Labs Distinguished Member of Technical Staff and working currently as a Distinguished Scientist, Visual Media at Nokia Technologies, Tampere, Finland. His research interests include video, image and graphics coding and communications.

PUBLICATION

IV

Comparison of HEVC coding schemes for tile-based viewport-adaptive streaming of omnidirectional video

R. Ghaznavi-Youvalari, A. Zare, H. Fang, A. Aminlou, Q. Xie, M. M. Hannuksela and M. Gabbouj

*IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*2017, 1–6
DOI: 10.1109/MMSP.2017.8122227

Comparison of HEVC Coding Schemes For Tile-based Viewport-adaptive Streaming of Omnidirectional Video

Ramin Ghaznavi-Youvalari[1], Alireza Zare[1], Huameng Fang[2], Alireza Aminlou[1], Qingpeng Xie[2], Miska M. Hannuksela[1], and Moncef Gabbouj[3]

[1] Nokia Technologies, Tampere, Finland
[2] Huawei Technologies Co., Ltd, Shenzhen, China
[3] Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland

Abstract— **Virtual reality applications make use of 360-degree panoramic or omnidirectional video with high resolution and high frame rate in order to create the immersive experience to the user. The user views only a portion of the captured 360-degree scene at each time instant, hence streaming the whole omnidirectional video in highest quality is not efficient. In order to alleviate the problem of bandwidth wastage, viewport-adaptive encoding and streaming schemes have been proposed. In these schemes, part of the captured scene that is within the viewer's field of view is delivered at highest quality while the rest of the scene in a lower quality. In this work, three tile-based viewport-adaptive methods using motion-constrained tile sets (MCTS), region-of-interest scalability and simulcast approach have been studied for streaming omnidirectional content. In the performed experiments with various tiling arrangements, MCTS-based scheme required highest bitrate compared to other methods. The scalable coding scheme provided the highest performance in terms of streaming bitrate saving on average up to 53% and 35% compared to streaming the whole omnidirectional video and MCTS-based method, respectively.**

Keywords— Virtual reality, video coding, HEVC, SHVC, panoramic video streaming.

I. INTRODUCTION

Panoramic or omnidirectional videos are widely used in virtual reality (VR) applications to provide the feel of immersion to the viewer. The conventional way of streaming panoramic video is realized by encoding the video content as a single-layer bitstream, which is transmitted to the receiver and fully decoded. The region of the decoded video corresponding to the current viewport is rendered onto a head-mounted display (HMD), which typically has a field of view (FOV) from 96° to 110° [1]. Thus, transmitting the whole 360-degree panorama content at the highest resolution and quality consumes an unnecessarily high bandwidth of the network.

In order to reduce the bandwidth consumption of the network, viewport-adaptive streaming (VAS) schemes have been developed. In these approaches, the primary viewport that is currently viewed by the viewer is transmitted at highest quality to the end-user device. The parts that are expected not to be visible to the user are transmitted with lower quality. The non-visible parts are sent due to the latency of the encoding and transmission system. If the user turns his/her head to the side/back views of the 360-degree scene, he/she will see the lower quality content for a short period. Based on a sensory feedback from the end-user device, the server transmits the representation that matches the new primary viewport as described above.

In the literature two types of methods have been studied for viewport-adaptive streaming of omnidirectional video: viewport-dependent and tile-based methods. In the viewport-dependent methods, the 360-degree image is first re-projected and packed into the same frame so that a finer sampling density is used for the primary viewport compared to the sampling density for the remaining parts of the 360-degree image. The re-projected and packed VR content is then encoded. Several versions of the same content are encoded, each for a different pre-defined orientation of the primary viewport. Viewport-dependent projections and packing methods were compared in [2]. However, the large number of versions needed for viewport-dependent methods have been considered impractical and resource-consuming, and hence the use of tile-based streaming methods have been suggested instead [3]. In [4], a tile-based VAS method was found to be more efficient in streaming rate utilization when compared against a viewport-dependent method, namely truncated cube map projection.

Tile-based VAS can be realized in several ways, out of which this paper analyzes the following:

1) Motion-constrained tile set (MCTS) based approach, in which tiles are encoded independently of other tiles of the same bitstream.

2) Region-of-interest (ROI) scalable coding approach, in which each tile position corresponds to an enhancement layer that is predicted from the base-quality panoramic video provided as the base layer of the scalable video bitstream.

3) Simulcast approach, in which base-quality panoramic video is encoded conventionally, e.g. for providing compatibility with legacy clients, and additionally tile partitioning is performed prior to encoding and each enhanced quality tile bitstream is coded independently.

While the MCTS-based approach has been studied earlier (see e.g. [5], [6], [7]), to the best of our knowledge this paper is the first to compare the MCTS-based approach with other tile-based VAS methods.

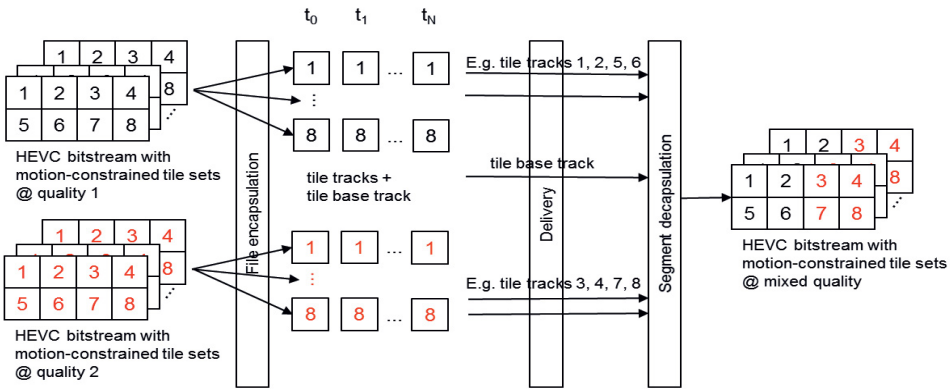


Figure 1. Single layer HEVC bitstream with motion constrained tile sets

The remainder of the paper is organized as follows. Related works are reviewed in Section II. The panorama streaming using VAS methods are studied in Section III. Section IV includes the simulation results and comparison of VAS methods. Finally, Section V presents the conclusion of the work.

II. Related Work

This section briefly reviews recent works for tile-based viewport-adaptive coding and streaming schemes.

In the High Efficiency Video Coding (HEVC) standard, pictures can be divided into several tiles along a grid of tile columns and rows. Spatial prediction and context prediction of the entropy coding do not cross tile boundaries. Furthermore, encoders can make a sequence of tile sets independent of the remaining tiles by constraining motion vectors as first proposed in [8] and referred to as the motion-constrained tile set (MCTS) technique in the HEVC context. The MCTS technique enables selectively choosing different combinations of tiles. For tile-based viewport-adaptive streaming, the received high-quality tiles cover the viewer's current viewport and the low-quality tiles corresponding to the non-viewport area.

In [5], a method was studied for generating a single bitstream for delivering the tile grids to the client in a way that is decodable with single hardware decoder. Further, the paper proposed a Generated Reference Picture (GRP) method for reducing the bitrate in the ROI switching points.

Streaming panoramas using the MCTS concept was studied in [6]. The paper proposed to encode the panorama at the full captured resolution and a lower resolution version using MCTS in the server side. A set of high-resolution tiles which covers the current viewing orientation along with a low-resolution tile set which covers the remaining part of the 360-degree scene are sent to the viewer. In [7], the tile-based streaming method was used with scalable extension of HEVC (SHVC). The tiling method is used both in the base layer and the enhancement layer, each enhancement layer tile is then inter-layer predicted from the co-located tile in the base layer. Furthermore, the authors extended the GRP method [5] for improving the bitrate in viewport switching points. However, the methods described in [7] require MCTS support in the SHVC encoder. Moreover, the enhancement and base layers of the proposed method consist only subset of the 360-degree scene, and hence for the cases in which the viewpoint switching is fast, the method would not be suitable.

The target of this paper is to study efficient streaming of the whole 360-degree scene to the user by utilizing tile-based VAS schemes.

III. Viewport-Adaptive Streaming Schemes

This section describes three tile-based viewport-adaptive streaming approaches for delivering panoramic videos.

A. MCTS-based VAS

This section briefly describes the state of the art VAS method using motion constrained tile set concept that was proposed in [6]. In this method several HEVC bitstreams of the same omnidirectional content are encoded at different qualities using MCTS technique.

Figure 1 demonstrates the MCTS-based approach for multiple quality scheme (a.k.a. quality adaptation). However, the multiple resolution scheme (a.k.a. spatial adaptation) can also be used instead of multiple qualities. As it can be seen from the figure, two versions of the same content in different qualities were encoded using the MCTS technique. Based on the user's viewing orientation, the primary viewport tiles are selected from the high-quality version of the bitstream, while the remainder of the scene is chosen from the low-quality tiles to cover the non-visible areas.

The viewport switching happens in the stream access points (SAPs) corresponding to instantaneous decoding refresh (IDR) pictures in both low- and high-quality bitstreams. Therefore, both low- and high-quality bitstreams require intra random access point (IRAP) pictures in order to provide seamless switching between viewports. However, the frequent IRAP pictures cause significant bitrate increment.

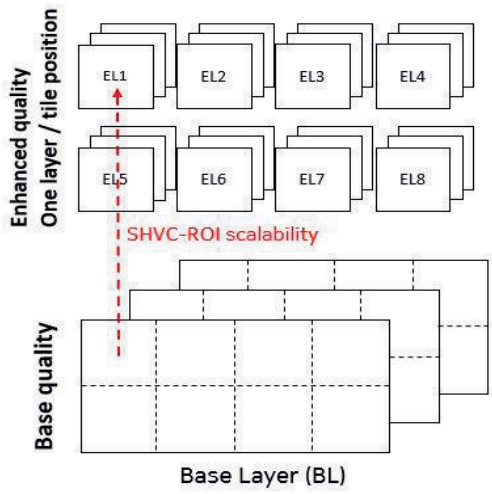


Figure 2. Illustration of SHVC-ROI method for 4x2 tile grid

B. *SHVC Region-of-Interest Coding Scheme*

In order to provide a seamless viewport switching in the MCTS-based method, both the low- and high-quality bitstreams are required to have frequent IRAP pictures. These IRAP pictures increase the bitrate of the low-quality content that are visible only in the cases of fast head movements for a short period of time. In order to alleviate the high bitrate caused by frequent IRAP pictures, the approach presented in this subsection uses infrequent IRAP pictures for the base-quality content. Furthermore, the approach makes use of region-of-interest (ROI) scalability of the HEVC scalable extension (SHVC) and is hence referred to as the SHVC-ROI scheme.

In the SHVC-ROI method, the base layer (BL) is encoded conventionally in 360-degree panorama format. It is not necessary to use HEVC tiles in the base layer. In fact, the base layer can be coded with any codec, such as the Advanced Video Coding standard (H.264/AVC), thanks to the external base layer feature supported by SHVC. Similar to the MCTS-based approach, the high-quality panorama is split to multiple tile segments, but each tile segment is encoded as a separate enhancement layer (EL). The enhancement layers are encoded using temporal prediction from the same layer and inter-layer prediction from the co-located region in the base layer. The sampling density used in the enhancement layer can match that of the base layer, i.e. quality scalability or can be lower than the enhancement layer, i.e. spatial scalability. Figure 2 illustrates the tile segmentation to multiple enhancement layers. While 4×2 tiling is illustrated in Figure 2, the method is not limited to any particular grid.

In the end-user device, the base layer is always received and decoded. Additionally, enhancement layers selected on the basis of the current viewing orientation are received and decoded. Stream access points for the enhancement layers are inter-layer predicted from the base layer, and are hence more compact than similar SAPs realized with intra-coded pictures. Since the base layer is consistently received and decoded, the SAP interval for the base layer is chosen to be longer than that of the enhancement layers. An example of SAP intervals for the base and enhancement layers is illustrated in Figure 3. The Random Access element of MPEG-DASH can be used to announce SAP intervals in the representations.

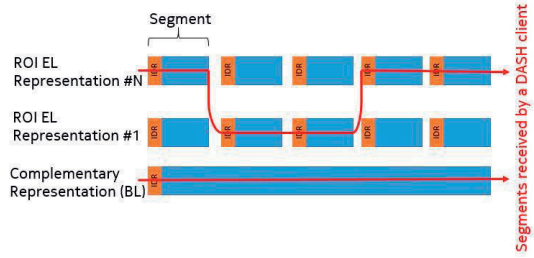


Figure 3. An illustration of switching between ROI EL representations of short segment duration based on viewing orientation.

From the system perspective, the client sends a request based on the viewing orientation to the server. The server responds to the request by transmitting segments of the enhancement layers corresponding to viewport along with the whole low-quality base layer. A single SHVC decoder instance can decode the high-quality viewport as well as the base layer video. Viewport switching can take place at a SAP in the enhancement layer, corresponding to an IDR picture. As IDR pictures are periodically coded, it is likely that there will be a delay until the enhancement layer corresponding to the new viewing orientation is received after the user turns his/her head. However, the base layer which contains the whole 360-degree scene will be available to be displayed to the user.

As it is illustrated in Figure 3, no representation switching happens in the base layer, hence it includes infrequent IRAP pictures. This results in a lower bitrate for the base-quality content compared to the respective bitstream encoded for MCTS-based streaming. However, in the SHVC-ROI approach the entire base layer is transmitted, while in the MCTS-based approach only the non-visible tiles of the base-quality are transmitted.

In addition to lower base-quality bitrate, the continuous transmission of the entire base layer is particularly helpful to avoid rebuffering and playback interruptions in the case of sudden throughput drops. In the proposed method, the transmission of the enhancement layer segments can be stopped at any time by terminating the respective Transmission Control Protocol (TCP) connections. In conventional tile-based streaming, bitrate adaptation choices are constrained by segment duration.

The minimum tile sizes in the HEVC standard is limited to 256 and 64 luma samples for tile column width and tile row height, respectively [9]. The SHVC-ROI method is not limited

Table 1. Tile partitioning chosen for 4K content

	Quality adaptation		Spatial adaptation	
Tiling arrangement	TileColumnWidthArray (MaxCUWidth = 64)	TileRowHeightArray (MaxCUHeight = 64)	TileColumnWidthArray (MaxCUWidth = 64)	TileRowHeightArray (MaxCUHeight = 64)
4×2	[15, 15, 15, 15]	[15, 15]	[8, 7, 8, 7]	[8, 7]
6×3	[10, 10, 10, 10, 10, 10]	[10, 10, 10]	[5, 5, 5, 5, 5, 5]	[5, 5, 5]
12×4	[5, 5, …, 5]	[8, 7, 8, 7]	-	-
12×8	[5, 5, …, 5]	[4, 4, 3, 4, 4, 3, 4, 4]	-	-

to the minimum tile size limitations, since each high-quality ROI is encoded in a separate enhancement layer.

C. Simulcast HEVC Coding Scheme

The SHVC-ROI approach described above reduces the bitrate of the low-quality content due to the infrequent IRAP pictures. However, the proposed method requires SHVC decoder which includes more complexity than the HEVC decoders.

The same technique as in the base layer of the SHVC-ROI method can be utilized in a manner that the bitstream is decodable using HEVC decoder. In this approach, the high-quality panorama is coded using the MCTS-based technique, while the lower quality content is coded conventionally without MCTS. Similar to base layer of the SHVC-ROI method, the lower quality video include longer SAP interval as illustrated in Figure 3. The primary viewports are selected among the high-quality tile sets. The lower quality content is always fully received and decoded in the user side.

The simulcast HEVC scheme benefits from the longer SAP interval in the lower quality content compared to the MCTS-based method. However, compared to the SHVC-ROI approach, this method lacks the inter-layer prediction that was beneficial for improving the rate-distortion performance of the higher quality video. The advantage of this method is the lower decoding complexity compared to the SHVC-ROI technique. The standard HEVC decoder is sufficient for decoding the primary viewports and the whole base-quality stream.

IV. Experimental Results

This section includes a comparison of the described VAS methods in Section III.

A. Experimental Conditions

The HEVC reference software (HM) version 16.15 [10] was used for coding the full-resolution panorama and the MCTS-based approach. For the SHVC-ROI method, the SHVC reference software (SHM) version 12.2 [11] was used.

For experimenting the described methods, 6 monoscopic panorama sequences each consisting 300 frames were used. All the video sequences are 4K JVET test contents for 360-degree video coding [12][13] which are in equirectangular projection (ERP) format.

Table 1 shows the tiling arrangements that were used for performance evaluation of the described VAS methods in Section III. For the spatial adaptation scenario, only 4×2 and 6×3 tile grids in the tiling arrangement list were feasible for low-resolution version of the video due to the minimum tile width of 256 luma samples in the HEVC Main profile [9].

The Main profile random access configuration [14] was used for the simulations. In the spatial adaptation test case, the higher resolution content has 2x spatial resolution difference along both axes compared to lower resolution one. In all the schemes, the contents were coded with the quantization parameter (QP) values of the JCT-VC random access common test conditions [14], except the lower quality version of quality adaptation scenario which had a QP value difference of 7 compared to the higher quality version. The QP difference was chosen to match the bitrate share that would be achieved by 2x spatial resolution difference along both axes. The decoding refresh type was set to IDR picture. In order to have equal viewport switching period, the same SAP interval was used in the equirectangular panorama, the MCTS-based encoding, the EL representations of the tested SHVC-ROI scheme and higher quality version of simulcast HEVC method. The selected SAP intervals were set to 32 pictures to achieve viewport switching period of 1 second. The SAP interval of the BL in SHVC-ROI and lower quality video of simulcast HEVC method representations were 10 seconds for all the test sequences.

B. Quality Assessment Method

The quality of the 360-degree video viewing experience was measured using the quality assessment process introduced in [15]. In this framework, a set of viewport-based representations of the captured scene is uniformly distributed over the sphere. The experienced quality is measured across a set of pre-defined viewing orientations in which the center of viewport may match the center of one of the viewport representation. The closest viewport representation is used to derive a viewport for the non-matching viewing orientation. For quality assessment, 24 (i.e., 360°/15°) uniformly distributed viewport orientations were defined along the equator, in which the viewing center of 12 of those match the viewport representation. Similarly, the same number of viewport orientations were selected in ± 15° latitudes, in total 72 viewing orientations were defined.

For each viewport representation, a set of tiles covering 110°×110° FOV was transmitted at high-quality/resolution while the remaining tiles were taken from a low-quality/resolution bitstream. For each viewing orientation, a viewport is rendered using rectilinear projection with 90°×90° FOV, which is close to the real-world perspective when using HMDs. For that, viewports are generated using cubemap projection.

In case of the SHVC-ROI method, the high-quality tiles are arranged in multiple enhancement layers which match the

corresponding tile grid in the MCTS-based method. However, for the low-quality content, the whole base layer video was transmitted.

The bitrate and quality values were averaged across the all viewing orientations. For quality measurement, peak signal-to-noise ratio (PSNR) is calculated on the front-face of the rendered cubemap. The streaming performance was measured in terms of Bjøntegaard Delta Bitrate (BDBR) criterion [16] for luma pictures.

C. Experimental Results

The quality adaptation and spatial adaptation streaming bitrate results of the described VAS methods in Section III compared to streaming the whole ERP are illustrated in Table 2 and Table 3, respectively. As it can be observed from the tables, the SHVC-ROI scheme superior the MCTS-based and simulcast HEVC (in Table 2 to Table 5 referred to as no inter-layer prediction (No-ILP)) schemes in all test cases.

Among the tiling arrangements, 12×8 tile grid provided the highest bitrate saving compared to other grids. As can be seen from the results in Table 2, the VAS methods with 12×8 tile grid provided on average 53%, 45.8% and 27.8% gain using the SHVC-ROI, simulcast HEVC and MCTS-based methods, respectively. The reason for such high bitrate saving is that by using finer tile grids, the high-quality/resolution area that cover the primary viewport become smaller, and the rest of the 360-degree scene is transmitted in lower quality/resolution. As a result of such tile selection, the streaming bitrate becomes smaller.

The inter-layer prediction gain in the SHVC-ROI scheme can be realized by comparing it to the simulcast HEVC method. As it can be observed from Table 2, in the quality adaptation scenario, the SHVC-ROI method provided around 7% to 15% more bitrate saving compared to the simulcast HEVC approach when using the same tile grid. Similar behavior can be observed in spatial adaptation case in Table 3, in which SHVC-ROI method provided 18% and 10% more bitrate reduction compared to simulcast HEVC method in 4×2 and 6×3 tile grids, respectively.

Based on the experimented results, the SHVC-ROI scheme performs better in the videos that have stationary content. This behavior can be observed for example in the 12×8 tiling arrangement of Table 2, BearAttack sequence with 60% gain and LRRH sequence with 65% gain had the highest bitrate saving among the test sequences due to the stationary content of these sequences. On the other hand, for DrivingInCity and DrivingInCountry sequences which include high global motion, the performance is relatively less than the other sequences.

Table 4 and Table 5 shows the streaming bitrate comparison of the SHVC-ROI and simulcast HEVC schemes compared to MCTS-based method as anchor for quality adaptation and spatial adaptation cases, respectively. The results demonstrated that the SHVC-ROI method improved the streaming performance in the range of 19% to 35% depending on the tile grid. Moreover, the results indicated that the simulcast HEVC method has sequence-wise performance variation compared to MCTS-based method, e.g. in the 4×2 tile grid of quality adaptation case, MCTS-based method outperformed the simulcast HEVC scheme in DrivingInCity and DrivingInCountry sequences. Similar behavior can be observed from 4×2 tile grid of spatial adaptation case in Table 5, in which the MCTS-based method outperformed the simulcast HEVC scheme on average over all sequences by 5.2%.

V. Conclusion

This work provided a comparison of three standard compliant viewport-adaptive streaming methods: HEVC MCTS-based, SHVC-ROI, and simulcast HEVC. In terms of streaming bitrate usage, the SHVC-ROI approach was found to be the best, while the HEVC MCTS-based method required the largest bitrate. The SHVC-ROI scheme reduces the bitrate of the high- and low-quality content by using inter-layer prediction and longer stream access points, respectively. The simulcast HEVC method is similar to the SHVC-ROI scheme but does not use inter-layer prediction and hence provided a smaller bitrate reduction when compared to SHVC-ROI.

The highest rate-distortion performance in all schemes were achieved in the finest tile grid. The experimented results demonstrated that the 12×8 tile grid provides the best streaming bitrate among other used tile grids. More than 50% and 35% bitrate saving were achieved compared to streaming the whole ERP content and the MCTS-based method, respectively, when using 12×8 tile grid in SHVC-ROI scheme. Moreover, in the 12×8 tile grid, the simulcast HEVC approach provided more than 45% and 24% streaming bitrate saving compared to full-ERP and MCTS-based method, respectively.

References

[1] W. Mason, . VR HMD Roundup: Technical Specs. Accessed Apr. 2016, from http://uploadvr.com/vr-hmd-specs/.

[2] K. Kammachi-Sreedhar, A. Aminlou, M. M. Hannuksela and M. Gabbouj, “Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications,” In 2016 IEEE International Symposium on Multimedia (ISM), pp. 583-586, Dec. 2016.

[3] S. Lederer, “Today's and future challenges with new forms of content like 360°, AR, and VR, ” invited talk in MPEG workshop Global Media Technology Standards for an Immersive Age, Jan. 2017, http://mpeg.chiariglione.org/sites/default/files/events/06_Lederer.pdf.

[4] A. Zare, A. Aminlou, and M. M. Hannuksela, "Virtual reality content streaming: viewport-dependent and tile-based techniques," Proc. of IEEE International Conference on Image Processing (ICIP), Sep. 2017.

[5] Y. Sánchez, R. Skupin, and T. Schierl. “Compressed domain video processing for tile based panoramic streaming using HEVC,” In IEEE International Conference on Image Processing (ICIP), pp. 2244-2248, Sep. 2015.

[6] A. Zare, A. Aminlou, M. M. Hannuksela and M. Gabbouj, “HEVC-compliant tile-based streaming of panoramic video for virtual reality applications,” In Proceedings of the 2016 ACM on Multimedia Conference, pp. 601-605, Oct. 2016.

[7] Y. Sanchez, R. Skupin and T. Schierl. "Compressed domain video processing for tile based panoramic streaming using SHVC," In Proceedings of the 3rd International Workshop on Immersive Media Experiences, pp. 13-18. 2015 Oct 30.

[8] M. M. Hannuksela, Y.-K. Wang, and M. Gabbouj, "Isolated regions in video coding," IEEE Transactions on Multimedia, vol. 6, no. 2, pp. 259-267, Apr. 2004.

[9] ITU-T, Recommendation H.265 (04/15): Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services-Coding of Moving Video, High Efficiency Video Coding, May 2017, http://www. itu. int.

[10] High Efficiency Video Coding (HEVC) reference software HM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, May 2017, https://hevc.hhi.fraunhofer.de/.

[11] Scalable Extensions of the High Efficiency Video Coding (SHVC) reference software SHM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, May 2017, https://hevc.hhi.fraunhofer.de/shvc.

[12] J. Boyce, E. Alshina, A. Abbas, Y. Ye, "JVET common test conditions and evaluation procedures for 360° video," ITU-T Joint Video Exploration Team (JVET), document JVET-D1030, Oct. 2016.

[13] J. Ridge, M. M. Hannuksela, E. Aksu, J. Lainema, and A. Aminlou, "Nokia test sequences for virtual reality video coding," ITU-T Joint Video Exploration Team (JVET), document JVET-C0064, June 2016.

[14] F. Bossen, "Common test conditions and software reference configurations," Joint Collaborative Team on Video Coding (JCT-VC), JCTVC-F900, Jul. 2011.

[15] K. Kammachi-Sreedhar, A. Zare, A. Aminlou and M. M. Hannuksela, "Testing methodology for viewport-dependent encoding and streaming," ITU-T Joint Video Exploration Team (JVET), document JVET-D0079, Oct. 2016.

[16] G. Bjøntegard. "Calculation of average psnr differences between RD-curves," document VCEG-M33, 2001, Austin.

Table 2. Streaming bitrate comparison of quality adaptation VAS methods relative to ERP (BD-Rate %)

	4×2 Tiling			**6×3 Tiling**			**12×4 Tiling**			**12×8 Tiling**		
Sequence	MCTS	SHVC	No-ILP	MCTS	SHVC	No-ILP	MCTS	SHVC	No-ILP	MCTS	SHVC	No-ILP
AerialCity	-15.5	-37.0	-20.3	-23.9	-49.1	-37.2	-15.6	-52.8	-39.3	-13.6	-54.3	-47.8
DrivingInCity	-15.2	-19.2	-6.4	-27.6	-33.6	-26.5	-19.8	-37.1	-27.2	-25.2	-36.5	-38.3
DrivingInCountry	-18.4	-23.2	-10.5	-29.7	-36.7	-28.0	-25.5	-40.7	-29.6	-27.9	-44.2	-32.0
PoleVault_le	-21.9	-37.7	-22.8	-31.9	-50.1	-39.7	-32.0	-55.1	-45.6	-37.8	-58.0	-43.4
BearAttack	-21.8	-42.2	-27.2	-31.1	-53.9	-43.5	-28.6	-58.1	-49.0	-31.0	-59.9	-56.4
LRRH	-21.9	-47.1	-29.8	-31.6	-58.4	-46.3	-32.4	-63.5	-53.1	-31.2	-65.3	-56.7
Average	**-19.1**	**-34.4**	**-19.5**	**-29.3**	**-47.0**	**-36.9**	**-25.6**	**-51.2**	**-40.6**	**-27.8**	**-53.0**	**-45.8**

Table 3. Streaming bitrate comparison of spatial adaptation VAS methods relative to ERP (BD-Rate %)

	4×2 Tiling			**6×3 Tiling**		
Sequence	MCTS	SHVC	No-ILP	MCTS	SHVC	No-ILP
AerialCity	-15.9	-36.4	-16.6	-20.6	-45.7	-33.5
DrivingInCity	-15.6	-16.7	0.8	-23.5	-30.8	-19.4
DrivingInCountry	-18.9	-22.1	-1.9	-24.7	-31.7	-19.3
PoleVault_le	-22.6	-35.7	-19.3	-29.7	-44.3	-36.1
BearAttack	-21.4	-43.2	-25.9	-27.9	-50.2	-42.2
LRRH	-23.6	-45.1	-27.9	-30.4	-53.9	-44.3
Average	**-19.7**	**-33.2**	**-15.1**	**-26.1**	**-42.8**	**-32.5**

Table 4. Streaming bitrate comparison of SHVC-ROI and No-ILP quality adaptation VAS methods relative to MCTS method (BD-Rate %)

	4×2 Tiling		**6×3 Tiling**		**12×4 Tiling**		**12×8 Tiling**	
Sequence	SHVC	No-ILP	SHVC	No-ILP	SHVC	No-ILP	SHVC	No-ILP
AerialCity	-25.4	-5.7	-33.1	-17.7	-44.0	-28.4	-47.1	-39.7
DrivingInCity	-4.8	10.2	-8.2	1.4	-21.8	-9.8	-15.5	-17.0
DrivingInCountry	-5.9	9.4	-10.0	2.1	-20.7	-6.2	-23.4	-6.8
PoleVault_le	-20.3	-1.2	-26.7	-11.5	-34.0	-20.0	-32.5	-8.2
BearAttack	-26.1	-6.9	-33.1	-18.0	-41.3	-28.6	-42.4	-36.2
LRRH	-32.3	-10.2	-39.2	-21.4	-46.1	-30.7	-49.5	-37.0
Average	**-19.1**	**-0.7**	**-25.1**	**-10.9**	**-34.7**	**-20.6**	**-35.1**	**-24.2**

Table 5. Streaming bitrate comparison of SHVC-ROI and No-ILP spatial adaptation VAS methods relative to MCTS method (BD-Rate %)

	4×2 Tiling		**6×3 Tiling**	
Sequence	SHVC	No-ILP	SHVC	No-ILP
AerialCity	-24.4	-1.0	-31.6	-16.5
DrivingInCity	-1.5	19.1	-9.6	5.2
DrivingInCountry	-4.1	20.5	-9.5	6.7
PoleVault_le	-17.0	4.2	-20.9	-9.2
BearAttack	-27.7	-5.7	-30.9	-19.9
LRRH	-28.1	-5.6	-33.7	-20.0
Average	**-17.1**	**5.2**	**-22.7**	**-9.0**

PUBLICATION V

Shared coded picture technique for tile-based viewport-adaptive streaming of omnidirectional video

R. Ghaznavi-Youvalari, A. Zare, A. Aminlou, M. M. Hannuksela and M. Gabbouj

IEEE Transactions on Circuits and Systems for Video Technology 29.10 (2018), 3106–3120
DOI: 10.1109/TCSVT.2018.2874179

Shared Coded Picture Technique for Tile-based Viewport-adaptive Streaming of Omnidirectional Video

Ramin Ghaznavi-Youvalari, Alireza Zare, Alireza Aminlou, Miska M. Hannuksela, and Moncef Gabbouj

***Abstract*— Tile-based viewport-adaptive streaming methods have been used in delivering omnidirectional video for virtual reality applications. In these methods, the 360° video is encoded in multiple quality versions by using motion constrained tile set (MCTS) technique. A set of high-quality and low-quality tiles, corresponding to viewport and non-viewport areas respectively, are selected and transmitted to the user. However, these methods require frequent intra random access points to ensure seamless viewport switching capability, very high decoding complexity or multi-layer coding scheme. The frequent intra random access points include very high bitrate in viewport switching points. The high decoding complexity and multi-layer decoder requirements are not aligned with the Omnidirectional MediA Format (OMAF) standard. Such requirements make these methods sub-optimal or impractical for streaming the omnidirectional video. This work studies the current tile-based solutions for delivering the omnidirectional content. Moreover, OMAF-compliant Shared Coded Picture (SCP) based scheme is proposed in this work for streaming the omnidirectional video. The core concept of the SCP-based method is to manipulate the switching point pictures in a way that the frequent intra-coded pictures are no longer required for the viewport switching operations between different quality versions of the content. The experiments illustrated that the SCP-based method outperforms the MCTS-based method on average by 11% to 14% in terms of streaming bitrate reduction with only 4% extra decoding complexity.**

***Index Terms*—Virtual reality, 360° video, coding and streaming, OMAF, HEVC, SHVC, Tile.**

I. Introduction

OMNIDIRECTIONAL video is characterized by the 360° horizontal field-of-view (FOV). A conventional approach for delivering omnidirectional video is realized by encoding the content as a single-layer bitstream and transmitting it to the receiver. The receiver decodes the full 360° bitstream, and the region that corresponds to the user's viewing orientation is rendered and displayed by using a head mounted display (HMD) device. Since the user views only a portion of the 360° video at each time instant (typical FOVs of HMDs vary from 96° to 110° [1]), streaming the whole 360° content in the highest resolution and quality consumes an unnecessary high bandwidth of the network.

Ramin Ghaznavi-Youvalari, Alireza Zare, Alireza Aminlou, and Miska M. Hannuksela are with Nokia Technologies, Tampere, Finland (Emails: {ramin.ghaznavi-youvalari, alireza.zare, alireza.aminlou, miska.hannuksela}@nokia.com).

Moncef Gabbouj is with Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland (Email: moncef.gabbouj@tut.fi).

Tile-based viewport-adaptive streaming (VAS) schemes [2]–[8] have been studied in recent years in order to reduce the delivery bandwidth. In these techniques, the part of the video that is currently viewed by the user (i.e., viewport area) is transmitted in the highest resolution and/or picture quality, and the rest of the 360° scene (i.e., non-viewport area) is delivered in a lower resolution and/or picture quality. If the viewer turns his/her head, the non-viewport areas are displayed only for a short period of time until the next high-resolution viewport is decoded and displayed. Frequent random access points (RAPs) are required to reduce the latency to update the quality of the viewport after a viewing orientation change. Among the tile-based VAS schemes, the motion-constrained tile set (MCTS) based methods have been widely suggested (e.g., [9]–[11]).

The recent Omnidirectional MediA Format (OMAF) standard [12] enables viewport-adaptive streaming through specified media profiles, which requires the VAS schemes to comply the following capabilities:

- Frequent viewport switching
- Single-layer decoding constraint
- Decoding with a single decoder instance
- 4K decoding constraint

This paper proposes a Shared Coded Picture (SCP) method that conforms to the OMAF viewport-dependent media profiles. The SCP-based scheme avoids the need of coding switching points as intra-coded pictures as required in MCTS-based VAS. In contrast, switching points are coded in the SCP-based method with shared coded pictures that are identical in different quality versions of the bitstream and are predicted only from other shared coded pictures. Such a coding scheme decreases the bitrate significantly compared to intra-coded pictures.

Consequently, either a more frequent viewport switching among the bitstreams is allowed or the streaming bitrate is reduced when compared to MCTS-based VAS. In the performed simulations, SCP-based VAS achieved more than 10% streaming bitrate reduction with the same viewport switching frequency and at the same picture quality compared to MCTS-based VAS.

The remainder of the paper is organized as follows. Section II includes background study of streaming omnidirectional video as well as review of related works. Section III provides a brief description of OMAF standard along with the viewport-dependent streaming requirements. Section IV describes the

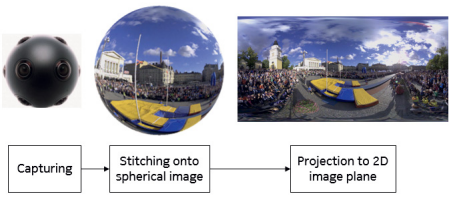


Fig. 1: Projection from spherical coordinates to 2D image plane by equirectangular projection

tile-based VAS schemes that are studied in [7]. The proposed SCP-based VAS method is described in Section V. Decoding complexity analysis of the VAS methods are studied in Section VI. Section VII describes the quality assessment methodology that is used in this work. Section VIII discusses the performance of the described VAS schemes in this work. Finally, Section IX presents the conclusion of the work.

II. Background And Related Works

A. Background

1) Motion-constrained Tile Set Technique: In the High Efficiency Video Coding (HEVC) standard, pictures can be divided into several tiles along a grid of tile columns and rows in a way that the spatial and temporal prediction of the entropy coding do not cross tile boundaries. Furthermore, encoders can make a sequence of tile sets independent of the remaining tiles by constraining motion vectors as first proposed in [13] and referred to as the motion-constrained tile set (MCTS) technique in the HEVC context.

The MCTS technique enables the functionality of selectively choosing and transmitting different combinations of tiles from the bitstreams. Due to the above-mentioned constraints in the prediction, a standard decoder is able to decode each tile independent from other tiles.

2) Viewport-adaptive Streaming: The 360° content can be represented by a sphere that has been mapped to a two-dimensional (2D) image plane by using planar projections (e.g., equirectangle, cube, pyramid, etc.). Figure 1 illustrates an example of the projection from spherical domain to a 2D image plane using equirectangular projection (ERP). The reason for 2D representation requirement of the spherical content is due to the fact that the current compression algorithms operate only for 2D format of the content. Hence, they are not suitable for compressing and delivering the spherical content in its 3D representation.

In order to create the immersive experience to the user in virtual reality (VR) applications, apart from the 360° requirement, the content must be in high resolution, quality and frame rate. Such requirements highlight the role of efficient encoding and streaming of omnidirectional content to the user with the current bandwidth limitations. Therefore, the utilized encoding and streaming method must provide a functionality that the transmitted VR content can be carried out over the users available internet connections (e.g., WLAN, 3G/4G networks, etc.).

In 360° video streaming, the ideal scenario would be to transmit only the user's viewport at that time in the highest resolution. However, temporal prediction in video encoding, segment-based delivery, as well as end-to-end transmission delay cause an inherent latency in the streaming system. Consequently, the transmitted video content cannot be instantly adapted to match exactly to the prevailing viewing orientation. Hence, VAS methods are studied for this purpose. The omnidirectional content representation for VAS can be realized by three categories of methods:

- Viewport-dependent projection schemes
- Region-wise mixed resolution (RWMR) schemes
- Region-wise mixed quality (RWMQ) schemes

The viewport-dependent projection schemes use unequal resampling approach for encoding and streaming the omnidirectional video. In these methods, the viewport area is encoded in the highest sampling density, and the rest of the 360° scene is re-projected in a way that the sampling density is gradually decreasing from the viewport to non-viewport areas. The re-projected non-viewport area is packed into the same image plane as the viewport area. Multiple versions of the same content, each consisting different viewport area is encoded and stored in the server side. For these methods, separate 360° stream for each viewing orientation is required.

In RWMR and RWMQ schemes, a viewport-independent projection, such as ERP, is used in content authoring. In RWMR method, the 2D projected 360° scene consists of different resolutions for the viewport and non-viewport areas (e.g., by resampling, rotating, mirroring, etc. of different parts of non-viewport area). Examples of this method are multi-resolution versions of ERP and cubemap formats [14]. In the region-wise mixed quality schemes, unlike the RWMR methods, the spatial resolution is the same for entire 360° scene, but the quality varies in viewport and non-viewport areas. The RWMR and RWMQ schemes can be transmitted to the user with separate 360° stream for each viewing orientation or as several tile-based streams. In the case of separate 360° stream for each viewing orientation, coding and storing multiple versions of the same content considering different viewport orientations is required.

The performance of various viewport-dependent projection and RWMR methods with separate 360° stream for each viewing orientation was studied in [14], [15]. A study in [6] demonstrated that the tile-based method requires 12% of the storage space, while having similar streaming performance, compared to one of the well-known viewport-dependent projection methods (i.e., truncated square pyramid projection [16]), coded for 30 different viewport orientations as suggested in [17]. A RWMQ approach with separate 360° streams for each viewing orientation was experimented in [18]. A certain number of quality emphasis centers (QECs) were selected, each corresponding to a separate 360° stream of a particular rotation of a cube map. The front cube face was centered at a QEC, while all other cube faces were coded at 25% lower bitrate compared to the bitrate of the front face. The number of

QECs was tuned to reduce storage requirements based on head position traces, which resulted into 5 to 7 distinct bitstreams.

The encoding and delivering of omnidirectional content using tile-based VAS schemes can be done using RWMR or RWMQ approaches. In both scenarios, the viewport area is selected from the encoded versions of the content with high spatial resolution and signal-to-noise ratio (SNR) quality. However, a comprehensive study in [19] demonstrated that when the resolution of the original content is lower than the decoding capacity and the ideal resolution for the display, the tile-based RWMQ scenario provides significant subjective improvements compared to the tile-based RWMR approach, in the same bitrate level.

This paper does not attempt to analyze whether viewport-dependent projection, RWMR schemes, or RWMQ schemes, nor whether several viewport-dependent 360° bitstreams or tile-based delivery are ideal for VAS. The paper studies the family of RWMQ tile-based VAS methods and proposes a new method for that family. Tile-based delivery is suggested in the industry [9] and supported by the OMAF standard [10]–[12] and hence the research problem addressed by this paper is considered important.

B. Related Works

Streaming the omnidirectional video with tile-based techniques are widely studied in recent years. An adaptive method was studied in [3] by real time modulating the quality of each region of the 360° video based on its viewing likelihood by the user.

Streaming panoramas using the MCTS concept was studied in [2]. The paper proposed to encode the panorama at the full captured resolution and a lower resolution version using MCTS in the server side. A set of high-resolution tiles which covers the current viewing orientation along with a low-resolution tile set which covers the remaining part of the 360° scene are sent to the viewer.

These methods [2], [3] provide proper switching and streaming bitrate reduction for the omnidirectional video, however both methods require to have frequent intra-coded pictures in the viewport switching points. The intra-coded pictures include very high bitrate and hence make these streaming methods sub-optimal.

In [4], a method was studied for generating a single bitstream for delivering the tile grids to the client in a way that is decodable with single hardware decoder. Further, the paper proposed a Generated Reference Picture (GRP) method for reducing the bitrate in the ROI switching points. The authors extended the GRP technique to be used in HEVC's extensions. The Multiview Generated Reference Picture (MGRP) and Multi-layer Generated Reference Picture (ML-GRP) were used in multiview extension (MV-HEVC) and scalable extension (SHVC) of the HEVC standard, respectively. However, these methods are suitable for only streaming subset of the 360° scene, and hence are not practical when the viewport switching is in a way that requires support for the whole 360° scene. Moreover, the multi-layer extensions of the HEVC standard are not widely supported in hardware implementations. Thus, using such techniques are not feasible in real practices.

A multi-layer coding scheme was studied in [5], in which the authors used scalable approach for adaptive streaming of the cubemap projected omnidirectional video. In this work, the full omnidirectional base layer which contains the low-quality content is transmitted along with the subset of enhancement layer that includes the viewport area. The authors also used longer IRAP interval in the base layer video compared to the enhancement layer. However, as mentioned above, the scalable codecs are not widely used in hardware implementations, also this method is not compliant with single-layer requirement of OMAF standard. Moreover, streaming the whole 360° base layer video, increases the decoding complexity significantly. Hence, such method is not applicable in streaming the omnidirectional video.

III. Overview of Omnidirectional Media Format

The Omnidirectional MediA Format (OMAF) was recently finalized by the Moving Picture Experts Group (MPEG) [10], [12]. It specifies file and delivery formats for three-degrees-of-freedom 360° audio, video, still images, and timed text. OMAF builds on other MPEG standards and specifies extensions to the ISO Base Media File Format (ISOBMFF) and the Dynamic Adaptive Streaming over HTTP (DASH) for enabling storage and delivery, respectively. OMAF is a toolbox standard from which features can be selectively implemented. To provide a limited number of interoperability options for implementations, OMAF specifies media profiles, each defining a subset of OMAF features in combination with a codec and constraints for the coded media bitstream.

This section reviews those features of OMAF that are considered important in understanding aspects and advantages of the SCP-based method presented in this paper. Selected basic features of ISOBMFF and DASH are first reviewed in Section III-A. OMAF specifies various pieces of metadata, out of which Section III-B presents those that are relevant for viewport-dependent streaming. Finally, Section III-C describes the features included in the HEVC-based viewport-dependent OMAF media profile and how the tile-based streaming is realized with that media profile.

A. Introduction to ISOBMFF and DASH

In ISOBMFF media bitstreams are logically organized into tracks. A track is accompanied by the necessary metadata, such as timing, that is needed for parsing and playback of the media bitstream. Metadata and media data are stored within different container structures in the file. Tracks can be time-wise partitioned into self-containing movie fragments, each consisting of the metadata and the media data for a particular time range.

DASH is based on a client-driven operation, where the bitrate adaptation logic and content selection takes place in clients and consequently the server needs not maintain a state for each client. A Media Presentation Description (MPD) in DASH describes the content available for streaming. The MPD format uses a hierarchical data model, in which an Adaptation Set contains Representations that are alternatives to each other and among which the client can select the one to be streamed

e.g. based on the bitrate. A Representation corresponds to a track in a file format and to a coded media bitstream. Representations are time-wise partitioned into Segments, which are atomic units described in the MPD and consist of an integer number of ISOBMFF self-containing movie fragments. The MPD provides bitrates and other content properties for Representations and either a template for deriving a Uniform Resource Locator (URL) for each Segment or a list of Segment URLs. Per each Segment, the client selects the Representation to be requested, concludes the respective URL from the MPD, and issues an HTTP GET request with that URL. The client can switch from a Representation to another one of the same Adaptation Set at random access or switching positions indicated in the MPD. The server can be an ordinary web server, simply responding to the HTTP GET request with the requested resource.

For on-demand content, Segments may be further divided into Subsegments. An index of the Subsegments is provided in the Segment itself, enabling clients to conclude Subsegment-wise byte ranges for HTTP GET requests. Otherwise Subsegment-based operation is similar to Segment-based operation and not described in further details in this paper due to the page count limit.

B. OMAF Metadata For Viewport-dependent Operation

OMAF specifies region-wise quality ranking (RWQR) metadata that enables clients to select between different options of the same content in a viewport-adaptive manner. The RWQR metadata for ISOBMFF or DASH MPD provides a quality order between indicated regions among all tracks or Representations of the same omnidirectional video source. Based on the RWQR metadata clients can select the track(s) or Representation(s) that cover the viewport at better picture quality.

The region-wise packing (RWP) metadata describes how indicated rectangular regions of decoded pictures map to a 360° picture of a specified omnidirectional projection format, such as ERP. The RWP metadata is provided for a track when it is not fully omnidirectional or when the sampling density varies region-wise.

C. HEVC-based Viewport-dependent OMAF Video Profile

OMAF includes a viewport-dependent video media profile that is based on HEVC and requires client processing for RWP. The bitstream to be decoded is required to conform to HEVC Main 10 profile at Level 5.1, which limits the maximum picture size to 8 912 896 luma samples, e.g. corresponding to 4096×2176 luma samples, and the maximum sample rate, in luma samples per second, to be 60 times of the maximum picture size. This paper uses the phrase ”4K decoding constraint” to refer to the constraints of HEVC Level 5.1. HEVC-based viewport-dependent OMAF video profile enables various viewport-adaptive streaming methods, many of which are summarized in an informative annex of the OMAF specification. Among these schemes is the tile-based viewport-dependent streaming described in [2], which is enabled in OMAF by encapsulating each MCTS sequence as an ISOBMFF track. The MCTS sequences of the same portion of the omnidirectional video content form the Representations of a single Adaptation Set in the MPD. Consequently, an OMAF client selects one MCTS sequence among the alternatives. The selection can be based on the bitrate and the RWQR metadata in a manner that the total streamed bitrate does not exceed the estimated prevailing network throughput and the viewport is covered by MCTSs that have higher quality than the remaining portions of the omnidirectional content.

In order to have clear input and output interfaces for a video decoder, OMAF video media profiles enable the operation with a single decoder instance. The HEVC standard does not specify a decoding process that takes individual MCTS sequences as input. To specify how an HEVC bitstream is formed from MCTS sequences, OMAF takes advantage of extractor tracks specified in HEVC's ISOBMFF encapsulation format [21]. An extractor track enables merging of MCTS sequences to a single HEVC-compliant bitstream by following prescribed instructions, called constructors, provided in the extractor track. An in-line constructor is used to rewrite high-level syntax structures of an HEVC bitstream, and a sample constructor is used to include coded MCTS data by reference. A sample constructor typically refers to an indicated group of tracks included as Representations in one Adaptation Set.

IV. Tile-based Viewport-adaptive Streaming Methods

As it was mentioned in Section II, tile-based approaches are considered to be practical for streaming the omnidirectional video content. We studied and compared three different approaches of tile-based region-wise mixed quality (RWMQ) methods in [7]. This section provides the description for these schemes and discusses the advantages and disadvantages of each method.

A. MCTS-based Coding Scheme

Streaming the 360° video using motion-constrained tile set (MCTS) technique is considered as the state of the art method. In this method, that is studied in [2], [7], the 360° video is divided into multiple tile segments. Each tile is encoded in a way that the spatial and temporal predictions are constrained to the tile boundaries. Moreover, the in-loop filtering operations do not use samples from across the tile boundaries. Such limitations provide us the independent decodability feature for each tile segment using a standard-compliant decoder.

Multiple versions of the same content with different qualities are encoded and stored in the server side. Based on the user's viewing orientation, a set of high-quality tiles that cover the corresponding viewport is selected. The non-viewport areas of the 360° scene are chosen from the lower quality version of the content and transmitted along with the high-quality viewport tiles. Figure 2 illustrates an example of such scheme by using 4×2 tile grid.

The MCTS-based scheme provides similar streaming performance as the viewport-dependent methods with significantly less storage requirements [6]. However, for ensuring the seamless switching between viewports for the user, such schemes

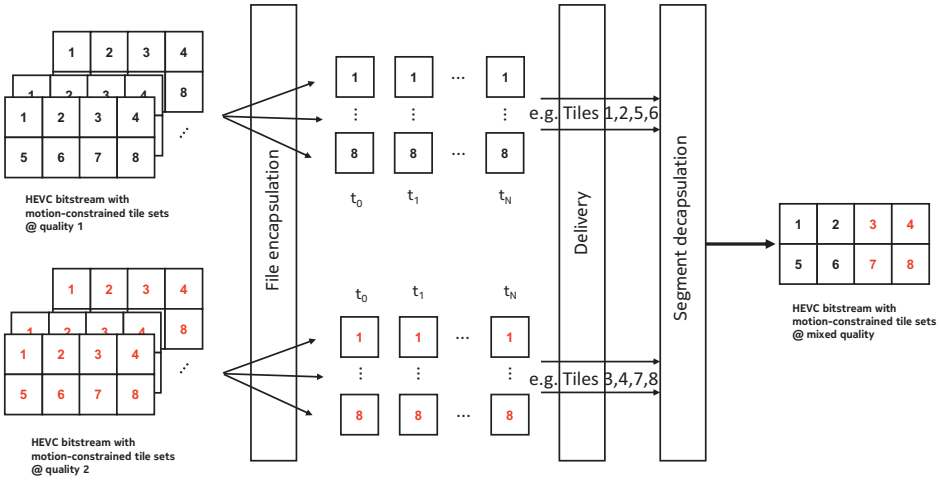


Fig. 2: Single-layer HEVC streaming with motion-constrained tile sets (MCTSs) [20]

required to have frequent intra random access points (IRAPs) for both low- and high-quality versions of the content. The IRAP pictures which are intra-coded, include significantly higher bitrate compared to the inter-coded pictures. Consequently, the overall bitrate for the transmitted bitstreams will increase. This issue is problematic particularly for the low-quality content that are only displayed for a short period of time until the next high-quality viewport is decoded and displayed.

Despite the fact that the MCTS-based scheme is compliant with OMAF standard requirements [22], the high bitrate of this approach in switching points makes this method sub-optimal for delivering the omnidirectional content.

B. SHVC Region-of-Interest Coding Scheme

This Section describes a method for resolving the issue of frequent IRAP pictures in the MCTS-based method by utilizing scalable codecs. Scalable coding approach is a multi-layer scheme in which each layer can be used for coding the video content in different quality and/or resolution than other layers. Hence, it can be used as a VAS scheme where multiple quality versions of the same content are required.

Figure 3 presents such coding scheme for delivering the 360° video by making use of region-of-interest (ROI) scalability of the HEVC scalable extension (SHVC) and is hence referred to as the SHVC-ROI scheme. In this approach, the base layer (BL) contains the low-quality version of the content and the enhancement layer (EL) includes the high-quality content.

The higher quality version of the content is encoded using MCTS technique of the MCTS-based method with conventional IRAP interval. In order to solve the problem of frequent IRAP pictures in the high-quality content, the inter-layer prediction (ILP) feature of the SHVC can be used in a way that each tile segment can be predicted from the co-located ROI in the base layer. As a result, in the switching points, intra-coded pictures are no longer necessary and are replaced by P-coded pictures, which require significantly less bitrate compared to intra-coded pictures.

The use of ILP, brings a requirement of streaming the whole 360° base layer to the user, the portion that covers the non-viewport area and the portion that is used for ILP of viewport area. Thus, it is not required to use tiling in BL since there is no switching is taking place for this content. Moreover, the IRAP interval in BL is longer than that of the enhancement layer representations. The longer IRAP interval resolves the frequent intra-coded picture issue that is used for low-quality content in the MCTS-based method.

The ILP can be constrained to the switching points (i.e., constrained-ILP or CILP) or can be enabled in all EL frames (i.e., full-ILP or FILP). In the case of FILP, higher bitrate reduction can be achieved with the consequence of higher encoding/decoding complexity. The ILP boosts the performance particularly in the switching points where the intra-coded pictures take place. The results in Section VIII, demonstrates that the FILP only provides less than 2% bitrate reduction compared to the constrained-ILP and majority of the bitrate reduction is the result of not using intra-coded pictures in viewport switching points.

The streaming performance of omnidirectional video by using the SHVC-ROI scheme that is shown in Table IV to Table VIII represent the significant bitrate reduction. However, the multi-layer codecs are not widely supported in hardware implementations due to their complexities compared to single-layer codecs. Moreover, due to transmission of the entire 360° video of the BL, the decoder needs to decode relatively more pixels compared to the MCTS-based approach. Such requirement introduces further complexity on top of the multi-

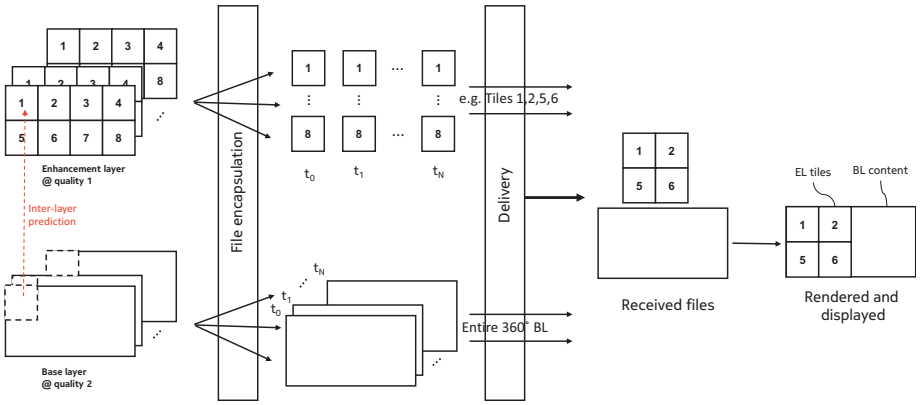


Fig. 3: Multi-layer SHVC streaming with motion-constrained tile sets (MCTSs)

layer requirements. Both the multi-layer scheme requirement and higher decoding complexity are not compliant with OMAF specifications for delivering the omnidirectional video. Hence, such constraints make the SHVC-ROI method impractical for omnidirectional content streaming scenarios.

C. Simulcast HEVC Coding Scheme

As it was mentioned in Section IV-B, scalable decoders are not popular in hardware implementations, hence a method studied in [7] that makes use of functionalities of both the MCTS-based technique and the SHVC-ROI method for decreasing the streaming bitrate of omnidirectional video in a way that the generated bitstream can be decoded with single-layer decoders.

In this technique, the lower quality version of the content is encoded conventionally with longer IRAP interval compared to the higher quality version (i.e. the same method in the BL of the SHVC-ROI scheme). As a result of the infrequent intra-coded pictures, the bitrate of the low-quality content will decrease significantly.

Moreover, the higher quality version of the 360° video is encoded similar to the higher quality version of the MCTS-based coding method in a single-layer scheme with conventional IRAP interval. By using this method, the bitrate of the low-quality content decreases, while the transmitted bitstream can be decoded by using a single-layer HEVC decoder. The difference between this method and the SHVC-ROI scheme is the inter-layer prediction feature. The lack of ILP between low- and high-quality content introduces the frequent IRAP pictures in the high-quality content that makes this scheme sub-optimal compared to the SHVC-ROI method.

Even though the generated bitstream is decodable with the standard single-layer decoders, but the entire 360° decoding requirement of the low-quality content (like the BL of the SHVC-ROI method) is not compliant with OMAF's 4K decoding constraint.

V. Proposed Shared Coded Picture-based Coding Scheme

The described VAS methods in Section IV require either frequent IRAP pictures (e.g. in low- and high-quality versions of the MCTS-based method and in high-quality of the Simulcast method) or include higher decoding complexity (e.g. in the SHVC-ROI and the Simulcast methods) and hence are considered sub-optimal or impractical for encoding and streaming omnidirectional video. Moreover, the SHVC-ROI and Simulcast HEVC schemes are not aligned with OMAF requirements for viewport-adaptive streaming of omnidirectional video.

This section introduces a novel tile-based VAS scheme to alleviate the above-mentioned issues of the three described methods. Shared Coded Picture (SCP) based scheme is proposed for this purpose. The aim of the SCP-based method is to achieve the following benefits:

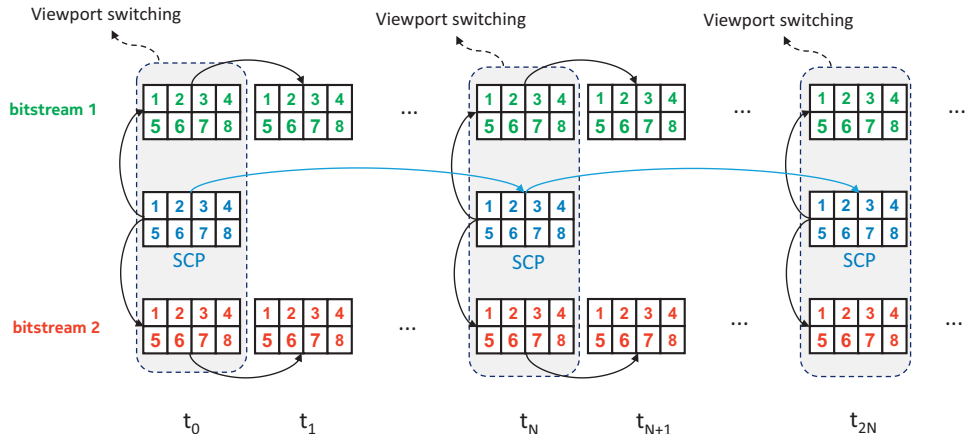
- Enabling viewport switching capability without the need of frequent intra-coded pictures
- Compliant with OMAF's single-layer decoding constraint
- Compliant with OMAF's 4K decoding constraint

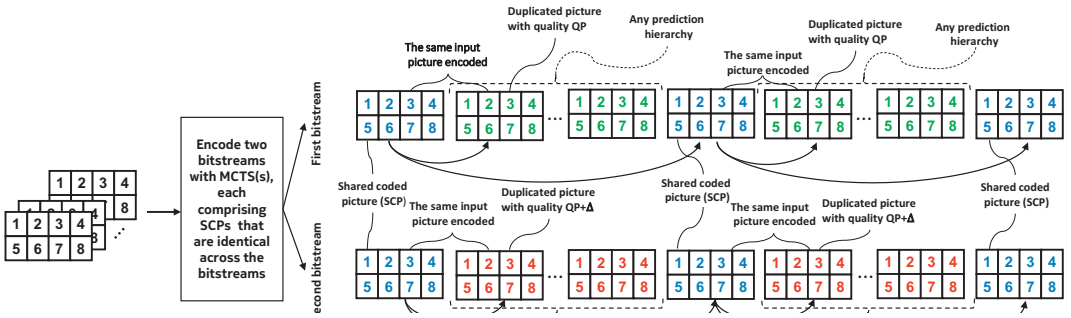
In the conventional MCTS-based method, viewport switching operation between different quality bitstreams is not possible without using intra-coded pictures, in order to act as refresh point to the prediction process. To avoid the intra-coded pictures in these points, the proposed method in this section facilitates such operation by making use of shared coded pictures between different quality bitstreams.

The encoding and streaming process of the omnidirectional video using SCP-based scheme is described in Section V-A. Moreover, the OMAF compliancy of the proposed method is discussed in Section V-B.

A. Encoding and Streaming Process of SCP-based Scheme

The main idea for enabling viewport switching without intra-coded picture requirement is to have certain pictures in





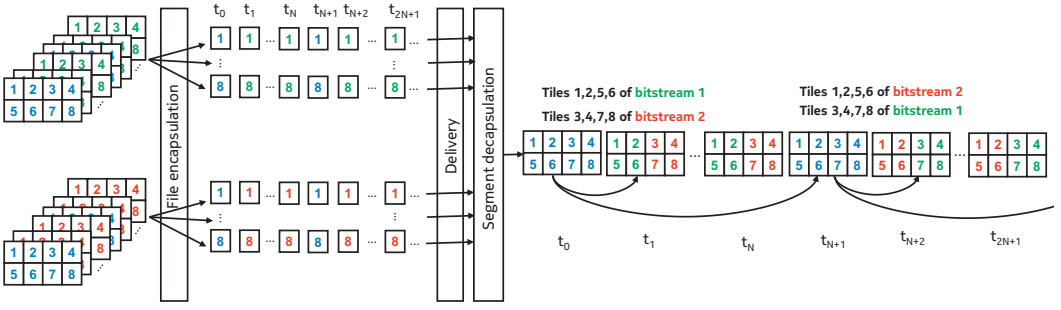


Fig. 6: Streaming the SCP-based viewports

different versions of the bitstream that are identical to each other. Figure 4 demonstrates the core concept of the SCP-based approach.

These pictures that are marked as SCP pictures in the figure, take place at the switching points. Since these pictures are identical in all versions of the bitstreams, viewport switching can happen in these points without the need for having intra-coded pictures. SCPs are typically aligned with Segments used in streaming, while the interval of intra-coded pictures can be selected according to a desired delivery granularity for random access. Since carriage over a reliable protocol, such as the Transmission Control Protocol (TCP), is assumed, the decoded bitstream does not contain errors and longer inter prediction chains caused by the use of SCPs do not cause temporal error propagation. As it can be seen from the figure, the SCPs are P-coded pictures which are predicted based on the previously coded SCPs and hence the required bitrate for the switching points will decrease significantly.

In order to facilitate the coding scheme to this feature, a pre-processing operation is required to be applied to the video sequence prior to encoding. In this process, certain pictures of the sequence (i.e., meant to be used in viewport switching) are duplicated into the sequence. As a result, the video sequence will comprise one extra picture per switching point (SWP) interval. The first picture of the repeated pictures is referred to as shared coded picture (SCP) and the second one as duplicated picture (DP), hereafter.

Figure 5 illustrates the encoding process for the SCP-based scheme that uses 4×2 tiling with MCTS technique. In this figure, the first and second bitstreams represent the high- and low-quality versions of bitstreams, respectively. In the figure, the first two pictures (the blue color picture and the consecutive green/red color picture) of the prediction hierarchy are the SCP and DP pictures that include the same content in both bitstreams but with different qualities. As mentioned above, the SCPs along the bitstreams at a certain time instance, are coded in a way that are identical in both quality versions of the bitstreams.

In the encoding process, the prediction hierarchy is designed in a way that the SCPs are indicated to use only other SCPs as reference pictures. Whereas, the duplicated pictures are predicted only from the corresponding SCPs.

The SCPs in the sequence may be encoded with the same or different quality as its corresponding duplicated pictures. Assuming that the higher quality version of the content is encoded with QP_{HQ} and the lower version is encoded with $\mathrm{QP}_{LQ} = \mathrm{QP}_{HQ} + \Delta$ (i.e., Δ is the quality difference between viewport and non-viewport video content). In this scenario, the SCPs in both high-quality and low-quality content are coded with QP_{LQ}. Furthermore, the remaining pictures (including DPs) inside the interval of the SCPs in both bitstreams are coded by using normal coding hierarchy. Thus, the DPs in the high-quality version of the content are encoded with different quality as their corresponding SCPs (QP_{HQ} for DPs and QP_{LQ} for SCPs), on the other hand, these pictures are encoded using the same quality (i.e., QP_{LQ}) in the lower quality version of the content. Consequently, DPs are encoded efficiently in high-quality content using inter prediction from the low-quality version of the same content (i.e., the corresponding SCP) and are skip-coded from the corresponding SCP in the same quality level in the low-quality version of the content.

The duplication process in the SCP-based scheme, results in extra pictures in the sequence compared to the VAS methods of Section IV. For a period of N (in terms of number of pictures) for SWP interval, the pre-processed video will include 1 extra picture than the conventional one. As an example, a 10 second video at 30 fps will comprise 10 extra frames if the SWPs are set to occur every 1 second.

There are multiple factors affecting the selection of an SCP interval, including the following:

- SCPs increase decoding complexity. The SCP interval must be selected so the additional pixel rate caused by SCPs does not cause a violation of the limits imposed by the decoding Level in use.
- As mentioned above, SCPs are aligned with Segment boundaries. Very short Segments could be inefficient in caching and increase the server load. For low-latency streaming, Segment durations starting from 0.5 seconds have been investigated [23]. The DASH Industry Forum Guidelines state that Segment durations from 1 to 10 seconds are reasonable [24].
- Minimizing any quality degradation that the user perceives because of motion-to-high-quality latency is one of the requirements for MPEG-I Phase 1b [25], which the upcoming version 2 of the OMAF specification should fulfill. Short segments provide the flexibility for more reactive client-driven bitrate adaptation and for smaller motion-to-high-quality latency in viewport-dependent streaming of 360° video. A small motion-to-high-quality latency enables a lower number of tiles to be streamed at high quality, since viewing orientation is less likely to exceed the area of the high-quality tiles.

Multiple versions of the same content are encoded in different qualities and stored in the server side. The selection of the viewport and non-viewport tiles is done like the VAS schemes in Section IV.

Since the SCPs and DPs comprise the same video content, in order to avoid displaying the same content twice to the user, the SCPs are indicated to be non-output pictures. As a response, the decoder does not output the reconstructed SCPs.

The overview of streaming omnidirectional video using the SCP-based method is shown in Figure 6. The viewport tiles are selected among the high-quality version of the bitstream and the remaining 360° scene that is the non-viewport areas are selected and transmitted from the low-quality version of the bitstream, resulting in a mixed-quality bitstream. The switching from one viewing orientation to another viewing orientation occurs in the SCPs that contain single quality tiles and are identical between bitstreams. It can be observed that no intra-coded pictures are required for switching between the bitstreams, hence the rate-distortion performance of the transmitted stream is improved compared to the conventional MCTS-based method described in Section IV-A.

The SCP-based scheme does not require transmitting and decoding the entire low-quality 360° scene when compared

to the SHVC-ROI and Simulcast methods, hence comprises significantly lower decoding complexity.

The main advantage of the SCP-based scheme over VAS methods of Section IV is that it facilitates the seamless viewport switching without using intra-coded pictures. Moreover, the decoding complexity is slightly higher compared to the MCTS-based approach due to the duplicated pictures in the sequence, however the streaming performance of the SCP-based method is significantly higher than the MCTS-based approach. Despite the mentioned advantages of the proposed method, the pixel rate of the sequence increases due to the SCP pictures. This can bring constraints to the coding process when a smaller viewport switching intervals are needed, in which the number of required SCP pictures will increase in proportion to the viewport switching interval. In such scenarios, the reference frame memory requirements will increase consequently.

B. Realization of SCP-based Scheme in OMAF

The presented Shared Coded Picture based method is fully compliant with HEVC, ISOBMFF, and DASH. It is also conforms to OMAF structurally as well as to the HEVC-based viewport-dependent OMAF video profile. Some details of the standard conformance are presented below.

HEVC encoding is controlled to mark the shared coded pictures as pictures that are not output by the decoding process. In practice, the SCPs have the *pic_output_flag* syntax element equal to 0. Likewise, the SCPs are marked as non-output pictures in the file encapsulation, as specified in [21]. Consequently, the SCP-based method causes no changes to HEVC decoding or to the rendering process in the client.

The file encapsulation is done similarly compared to the MCTS-based method. In other words, each MCTS sequence is encapsulated as its own track, and an extractor track is generated to include instructions how to merge an HEVC bitstream from the MCTS sequences. OMAF RWQR and RWP metadata are used with the SCP-based method similarly to how they are used with the MCTS-based scheme. For tracks carrying MCTS sequences, the RWQR metadata indicates the relative quality of output pictures (i.e., excluding SCPs) and the RWP metadata indicates the location of the MCTS on the 360° picture.

The Segments starting with a shared coded picture are marked as bitstream switching points, using the Switching element of the MPD. As opposed to other types of switching points, a bitstream switching point is such that the decoder is not re-initialized but rather the concatenated Segment sequence conforms to the bitstream format, i.e. HEVC in the SCP case. By parsing the Switching element, the client has the capability to switch between Representations at each Segment starting with a shared coded picture.

For example, when the content resolution is 4096×2048, the SCP-based method can be used with the HEVC-based viewport-dependent OMAF video profile for 4K decoding capacity with a minimum interval of one shared coded picture for each 16 conventionally coded pictures. This is enabled by the maximum processing rate of HEVC Level 5.1, which is slightly greater than what would be required for decoding pictures of size 4096×2048 at 60-Hz picture rate.

VI. Decoding Complexity Analysis Of The VAS Schemes

This section studies the decoding complexity of the tile-based VAS schemes of Sections IV and V. The performances of the described schemes are analyzed with various tiling arrangements in the conducted experiments. Table I shows the details of the 5 tile grids that have been used in this work. The selected tile grids cover very coarse tiling (i.e., 8 tiles or 4×2 grid) to a very fine tiling gird (i.e., 96 tiles or 12×8 grid).

Table II demonstrates the average decoding complexity of the described VAS methods in each tiling arrangement. The complexities are categorized as the average percentages of required pixels for decoding from high-quality (HQ), low-quality (LQ) contents and total required pixels for decoding (HQ+LQ) relative to the full 360° omnidirectional video.

As it can be observed from the table, the portion that covers the viewport (i.e., HQ) are the same regardless of the utilized VAS method in the corresponding tile grid. However, this portion varies for the non-viewport area (i.e., LQ) depending on the utilized VAS scheme. In the case of the MCTS-based method, the low-quality area is the complementary part of the transmitted 360° scene. Hence, the decoder needs to decode the 100% of the 360° scene regardless of the tiling arrangement.

In the SHVC-ROI (both FILP and CILP) and Simulcast methods, the entire 360° low-quality content is transmitted to the user. Therefore, the decoder is required to decode extra pixels compared to the MCTS-based approach. As it can be observed from the table, the decoder requirement is around 160% to 132% for the coarsest to finest tile grids, respectively.

In the SCP-based scheme, similar to the MCTS-based method, the low-quality area is complementary to the high-quality parts. Moreover, as it was explained in Section V, this method requires transmitting extra SCP picture to the user for each viewport switching interval. Thus, the decoder is required to decode 1 extra picture in the switching intervals (i.e., usually around 1 second). The decoding complexity for the SCP-based method can be calculated from equation (1) below:

$$Complexity\% = (\frac{SWP + EP}{SWP}) \times 100\% \qquad (1)$$

In this formula, the SWP is the switching point interval and EP represents the extra picture in every SWP. In our experiments, the viewport switching takes place every 24 pictures. Considering the 1 extra picture in every SWP, the complexity is around 104% relative to the 360° omnidirectional content and 4% higher when compared to the MCTS-based method. This extra decoding requirement is regardless of the used tile grid, hence unlike the SHVC-ROI and Simulcast methods, is consistent in all the tiling scenarios. SCP interval of 24 pictures corresponds to 0.8-second Segments for 30-Hz video, which is believed to be a reasonable trade-off between avoiding high server load caused by very short Segments and achieving relatively low motion-to-high-quality latency.

TABLE I: Tile partitioning chosen for 4K content

Tile grid	TileColumnWidthArray	TileRowHeightArray
4×2	[15, 15, 15, 15]	[15, 15]
6×3	[10, 10, ..., 10]	[10, 10, 10]
8×4	[8, 7, 8, 7, 8, 7, 8, 7]	[8, 7, 8, 7]
12×4	[5, 5, ..., 5]	[8, 7, 8, 7]
12×8	[5, 5, ..., 5]	[4, 4, 3, 4, 4, 3, 4, 4]

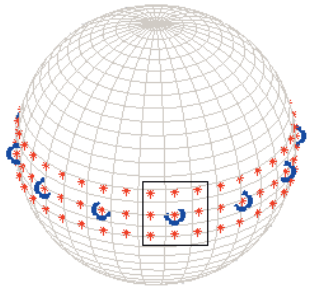


Fig. 7: QAVs used in the simulation. Blue marks: center of viewport-based streams, Red marks: center of QAVs

VII. Quality Assessment Methodology For Omnidirectional Video

The quality of the 360° video viewing experience was measured using the quality assessment process introduced in [26]. In this framework, a set of viewport-based streams of the captured scene is uniformly distributed over the sphere. Each stream contains high-quality content in a pre-defined viewport, while the non-viewport area of the content is presented at a lower quality. The experienced quality is measured across a set of pre-defined viewing orientations called quality assessment views (QAVs) in which the center of viewport may match the center of one of the viewport-based streams. The closest stream is used to derive the non-matching QAVs. In this work, each stream is defined to have a 110°×110° of viewport coverage.

In the MCTS-based and SCP-based VAS methods, to construct each viewport-based stream, a set of tiles covering 110°×110° FOV (i.e., viewport) is transmitted at high quality while the remaining tiles covering the non-viewport area are streamed from lower quality bitstream. In the SHVC-ROI and Simulcast methods, the high-quality tiles are selected in a way to achieve the same 110°×110° coverage as above. However, for the low-quality content, in both techniques the full low-quality is transmitted.

In this experiment, 12 viewport-based streams (shown in Figure 7 with blue marks) located along the equator is defined, each 30° apart. For quality assessment, 24 (i.e., 360°/15°) uniformly distributed QAVs (shown in Figure 7 with red marks) are defined along the equator, in which the viewing center of 12 of those match the viewport-based streams. Similarly, the same number of QAVs are selected in ±15° latitudes, in total 72 QAVs are defined. This arrangement is aligned with the previous studies (e.g., [27] and [28]) show that viewers mainly focus on the equator and very rarely watch top and bottom poles, while looking at HMDs.

For each QAV, a viewport is rendered using rectilinear projection with 90°×90° FOV, which is close to the real-world perspective when using a typical HMD available in the market. For that, viewports are generated using the cubemap projection.

VIII. Experiments

A. Experimental Conditions

The HEVC reference software (HM) version 16.7 [29] was used for coding full-resolution omnidirectional video (i.e., the Anchor), the MCTS -based, the Simulcast, and the SCP-based methods. The SHVC reference software (SHM) version 12.2 [30] was used for the FILP and CILP SHVC-ROI schemes. For ERP to cubemap conversion, the 360Lib tool [31] developed by Joint Video Exploration Team (JVET) was used.

In order to evaluate the performance of the described VAS schemes in Section IV and V, 6 monoscopic omnidirectional sequences were used. All the video sequences are at 4K resolution from JVET test content for 360° video coding [31], [32], which are in equirectangular projection (ERP) format. The number of frames of the test sequences were 300 frames for all VAS schemes, except the SCP-based method that includes 13 extra duplicated pictures as discussed in Section V-A. Main profile random access (RA) configuration [33] was used for conducting the simulations.

The higher quality version of the bitstreams in all described VAS methods were coded with the quantization parameter (QP) values of the JCT-VC RA common test conditions (CTC) [33], i.e., 22, 27, 32, and 37. However, the lower quality version of the content was coded with a QP value difference of 7 compared to the higher quality version of the content. The QP difference was chosen to match the bitrate share that would approximately be achieved by 2x spatial resolution difference along both axes.

The decoding refresh type was set to instantaneous decoding refresh (IDR) pictures. The viewport switching was set to take place in equal SWP intervals for all described VAS schemes as well as the ERP anchor method. The SWP interval is chosen to occur every 24 pictures. The SWP interval for the lower quality version of the content in the SHVC-ROI (i.e. the BL) and the Simulcast methods were decided to be longer when compared to the higher quality version of the content, thus it was set to 10 seconds for all test sequences.

As discussed in Section VI, in order to have thorough analysis of the performances, various tile grids have been used for conducting the simulations. Table I includes the information related to the experimented tile grids.

The performances were analyzed by using the well-known *Bjøntegaard* Delta Bitrate (BDBR) criterion [34] for luma pictures. The negative values are the indications of how much the bitrate is decreased in the same peak signal-to-noise ratio (PSNR). Similarly, the positive values show the bitrate increment for the same quality level.

TABLE II: Decoding complexity of VAS schemes in terms of amount of decoded pixels

Tiling	MCTS-based			SHCV-ROI (FILP/CILP)			Simulcast			SCP-based		
	HQ	LQ	**Total**	HQ	LQ	**Total**	HQ	LQ	**Total**	HQ	LQ	**Total**
4×2	59.9%	40.1%	**100.0%**	59.9%	100.0%	**159.9%**	59.9%	100.0%	**159.9%**	59.9%	44.3%	**104.2%**
6×3	47.7%	52.3%	**100.0%**	47.7%	100.0%	**147.7%**	47.7%	100.0%	**147.7%**	47.7%	56.5%	**104.2%**
8×4	41.7%	58.3%	**100.0%**	41.7%	100.0%	**141.7%**	41.7%	100.0%	**141.7%**	41.7%	62.5%	**104.2%**
12×4	38.7%	61.3%	**100.0%**	38.7%	100.0%	**138.7%**	38.7%	100.0%	**138.7%**	38.7%	65.5%	**104.2%**
12×8	32.4%	67.6%	**100.0%**	32.4%	100.0%	**132.4%**	32.4%	100.0%	**132.4%**	32.4%	71.8%	**104.2%**

TABLE III: Tiling overhead for different tile-grids in terms of BD-Rate (%) relative to without tiling scenario

Sequence	High-quality content					Low-quality content				
	4×2	6×3	8×4	12×4	12×8	4×2	6×3	8×4	12×4	12×8
AerialCity	5.5	7.2	12.3	16.3	22.5	5.7	9.1	14.8	20.5	29.2
DrivingInCity	7.1	5.6	11.3	14.7	18.1	9.2	8.1	15.5	20.7	26.5
DrivingInCountry	5.8	8.1	12.5	16.6	22.2	7.5	11.0	17.1	23.1	32.0
PoleVault_le	0.9	1.7	2.3	3.2	4.8	1.3	2.6	3.6	5.2	8.2
BearAttack	1.0	1.7	2.5	3.7	5.8	1.9	3.3	5.1	7.7	12.7
LRRH	0.6	1.0	1.5	2.0	3.1	1.0	1.7	2.7	3.7	6.0
Average	**3.5**	**4.2**	**7.1**	**9.4**	**12.8**	**4.4**	**6.0**	**9.8**	**13.5**	**19.1**

B. Analysis of the Results

1) Tiling Penalty: The spatial and temporal constraints that were applied in tiling techniques for providing independent decoding functionality for each tile (e.g., in the MCTS-based and SCP-based methods) will affect the compression performance compared to the non-tiled video. And since the tiling technique is not used in the conventional streaming of omnidirectional video (i.e., ERP) method and is partly used in the SHVC-ROI (in the EL) and Simulcast (higher quality version) methods, this section intends to study the tiling penalty in these cases.

Table III includes the compression performance comparison of the omnidirectional video that are used in this work in the tiled and non-tiled scenarios with the same coding configurations. Both low- and high-quality cases are considered with the tile grids of Table 1 in order to analyze the penalties.

As it can be observed from the results, the tiling overhead increases as the number of tiles increase. In the high-quality content, the overhead varies on average from 3.5% in the coarsest tile grid to around 13% in the finest tile grid. However, in the low-quality version of the content, this penalty varies from 4.4% to 19% in the coarsest and finest tile grids, respectively. This indicates that the SHVC-ROI and Simulcast methods that are not utilizing tiling in the lower quality contents, are benefiting from this issue in terms of compression performance compared to other schemes.

The results of Table III also demonstrate that the tiling penalty is significantly higher in the sequences with higher motion compared to the stationary sequences. For example, in the *AerialCity*, *DrivingInCity* and *DrivingInCountry* sequences that include high motion, the penalties are relatively higher compared to *PoleVault_le*, *BearAttack* and *LRRH* sequences that have stationary content. Such behavior can be observed in both low- and high-quality contents regardless of the tile grid.

2) Streaming Performance: Streaming performance of the described VAS schemes in Section IV and V under various tile grids compared to streaming the whole 360° (referred to as ERP) are presented in Table IV to Table VIII. As the results demonstrate, all the schemes provide very high bitrate reduction compared to ERP method.

In terms of various tiling arrangements, the streaming bitrate reduction increases as the tile grids increase. The lowest bitrate reductions were achieved in the 4×2 tiling that is the coarsest grid with the bitrate reductions of 21.9%, 32.6%, 30.9%, 18.1% and 33.5% for the MCTS-based, FILP SHVC-ROI, CILP SHVC-ROI, Simulcast and SCP-based schemes, respectively. Moreover, the highest bitrate reduction among the tile grids was achieved in the finest grid that is 12×8 tiling, in which on average of all sequences 31.6%, 53.5%, 52.7%, 44.3% and 45.2% bitrate reduction were achieved when using the MCTS-based, FILP SHVC-ROI, CILP SHVC-ROI, Simulcast and SCP-based methods, respectively. The reason for such performance difference is that, in the fine tiling arrangement, the high-quality area that covers the viewport becomes smaller when compared to coarse tiling. Thus, larger area from the transmitted 360° scene is covered by low-quality content that requires lower bitrate compared to the high-quality content.

The performance of the inter-layer prediction can be realized by comparing the results of the FILP SHVC-ROI and Simulcast approach. As it can be observed, when the ILP is enabled to perform in all the frames, this feature provides around 10% to 15% gain compared to the Simulcast method that does not include ILP. Moreover, the significant share of this bitrate reduction is due to the use of ILP functionality in viewport switching points. Comparing the the FILP and CILP SHVC-ROI results illustrate that less than 2% extra bitrate reduction is achieved by using the ILP in all frames compared to the case when this feature is used in switching points only.

Among all the experimented schemes, the SHVC-ROI schemes and SCP-based method outperform the MCTS-based and the Simulcast approach in all tiling arrangements. This performance difference is significant particularly when compared to the MCTS-based approach. The FILP SHVC-ROI scheme provides around 10% gain in the worst case (i.e., 4×2 tiling) and around 22% gain in the best case (i.e., 12×8 tiling), compared to the MCTS-based method. However, this range in the SCP-based method is different than scalable coding approach and is around 11% in the worst case (i.e., 4×2 tiling) and around 14% in the best scenario (i.e., 12×4 tiling).

TABLE IV: Streaming bitrate comparison for 4×2 tiling of VAS methods relative to ERP (BD-Rate %)

Sequence	MCTS-based	FILP SHVC-ROI	CILP SHVC-ROI	Simulcast	SCP-based
AerialCity	-19.2	-36.8	-36.4	-20.3	-33.2
DrivingInCity	-18.0	-19.4	-16.7	-5.1	-22.6
DrivingInCountry	-20.9	-21.7	-19.0	-8.3	-24.6
PoleVault_le	-24.4	-34.7	-32.9	-20.7	-35.1
BearAttack	-24.3	-39.9	-38.5	-26.3	-41.5
LRRH	-24.8	-43.0	-42.1	-28.1	-44.0
Average	**-21.9**	**-32.6**	**-30.9**	**-18.1**	**-33.5**

TABLE V: Streaming bitrate comparison for 6×3 tiling of VAS methods relative to ERP (BD-Rate %)

Sequence	MCTS-based	FILP SHVC-ROI	CILP SHVC-ROI	Simulcast	SCP-based
AerialCity	-28.4	-49.2	-48.9	-37.0	-44.6
DrivingInCity	-30.2	-33.4	-31.6	-24.4	-35.7
DrivingInCountry	-31.7	-34.9	-32.7	-25.0	-36.1
PoleVault_le	-35.0	-47.5	-46.2	-37.3	-46.7
BearAttack	-34.2	-52.3	-51.3	-42.3	-52.1
LRRH	-35.2	-55.1	-54.5	-44.1	-54.9
Average	**-32.4**	**-45.4**	**-44.2**	**-35.0**	**-45.0**

TABLE VI: Streaming bitrate comparison for 8×4 tiling of VAS methods relative to ERP (BD-Rate %)

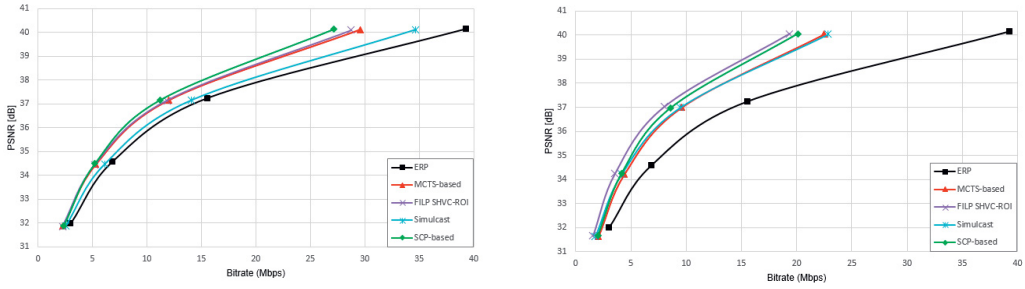
Sequence	MCTS-based	FILP SHVC-ROI	CILP SHVC-ROI	Simulcast	SCP-based
AerialCity	-20.7	-44.2	-44.0	-29.2	-35.9
DrivingInCity	-21.9	-28.0	-25.8	-15.5	-27.4
DrivingInCountry	-25.0	-30.1	-27.4	-17.4	-29.4
PoleVault_le	-31.0	-43.6	-42.2	-32.2	-42.4
BearAttack	-29.5	-47.9	-47.0	-37.2	-47.3
LRRH	-31.6	-51.7	-51.1	-39.8	-51.2
Average	**-26.6**	**-40.9**	**-39.6**	**-28.5**	**-38.9**

TABLE VII: Streaming bitrate comparison for 12×4 tiling of VAS methods relative to ERP (BD-Rate %)

Sequence	MCTS-based	FILP SHVC-ROI	CILP SHVC-ROI	Simulcast	SCP-based
AerialCity	-25.6	-53.5	-53.5	-41.4	-42.2
DrivingInCity	-27.4	-37.8	-36.2	-27.8	-33.9
DrivingInCountry	-31.7	-40.0	-37.8	-29.6	-36.8
PoleVault_le	-38.4	-53.1	-52.1	-44.5	-50.6
BearAttack	-35.9	-57.3	-56.7	-49.3	-54.4
LRRH	-39.2	-61.0	-60.6	-52.2	-59.3
Average	**-33.0**	**-50.5**	**-49.5**	**-40.8**	**-46.2**

TABLE VIII: Streaming bitrate comparison for 12×8 tiling of VAS methods relative to ERP (BD-Rate %)

Sequence	MCTS-based	FILP SHVC-ROI	CILP SHVC-ROI	Simulcast	SCP-based
AerialCity	-20.7	-56.2	-56.4	-44.2	-38.2
DrivingInCity	-23.0	-37.9	-36.3	-27.7	-29.8
DrivingInCountry	-32.1	-44.7	-43.2	-34.5	-37.5
PoleVault_le	-40.0	-57.4	-56.6	-49.9	-52.7
BearAttack	-34.3	-60.9	-60.5	-54.1	-53.4
LRRH	-39.2	-63.4	-63.1	-55.5	-59.6
Average	**-31.6**	**-53.4**	**-52.7**	**-44.3**	**-45.2**



(a) 4×2 tiling (b) 12×8 tiling

Fig. 8: RD-Curves of *DrivingInCountry* sequence for tile-based VAS schemes in a) 4×2 tiling and b) 12×8 tiling

Comparing the SHVC-ROI schemes and the SCP-based method, different performances were observed between coarse and fine tiling. In the coarse tile grids, which are 4×2, 6×3 and 8×4 cases, the SCP-based and SHVC-ROI schemes provide similar streaming performance. In the 4×2 gird, the SCP-based scheme outperforms the FILP and CILP SHVC-ROI methods slightly. However, the scalable schemes outperform the SCP-based method in finer tile grids, from 4% in 12×4 tile grid to 8% in the 12×8 tile grid. The difference is slightly lower when the SCP-based method is compared to the CILP version of scalable approach. Such variations in the performances of the SCP-based scheme in different tile grids can be explained by the tiling overhead that was discussed above. The SCP-based approach requires to have tiling in both low- and high-quality contents, however, on the other hand, tiling only utilized in the EL of the SHVC-ROI schemes and the BL were coded conventionally.

As the results in Table III illustrate, for the low-quality content, this overhead is relatively large for 8×4, 12×4, and 12×8 tile grids with overheads of 10%, 13% and 19%, respectively. Even though some portion of this overhead was compensated by streaming only non-viewport areas from the low-quality content in the SCP-based method compared to streaming the entire 360° of the BL in scalable methods, but since the transmitted non-viewport area in finer tile grids become larger compared to the coarser grids, the overall penalty of tiling has increased consequently.

Furthermore, 12×4 and 12×8 tile grids are not aligned with HEVC Level 5.1 constraints that are required in OMAF standard. Hence, the performance of the SCP-based scheme is roughly equal to the SHVC-ROI schemes in the OMAF-compliant tiling arrangements.

Figure 8 demonstrates the rate-distortion (RD) curves of the VAS methods for the *DrivingInCountry* sequence in the coarsest and finest tile grids. Based on the RD-Curves of the 4×2 tiling in Figure 8a, the proposed SCP-based scheme outperforms other VAS methods particularly in high bitrates. Similar performance behavior was observed in 6×3 and 8×4 tiling, however, due to the page limit count, these RD-Curves were not included in the paper. For the finest tile grid (12×8 tiling) in Figure 8b, the SHVC-ROI scheme provides better RD performance than the other VAS schemes.

3) Storage Requirements: Table IX shows the storage requirements for the described VAS schemes in this work compared to the conventional ERP encoding method. As mentioned in VIII-A, 4 quality levels are considered in the experiments according to the common test condition [33], for the high-quality version of the content. Moreover, the corresponding low-quality version of the content is coded with QP differences of 7. The storage requirements in the table are calculated based on the aggregation of the required bitrates in different quality levels for both low- and high-quality versions of the content.

Among the schemes, the MCTS-based method has the largest storage requirements on average with 134% storage of the ERP. This can be explained by the frequent intra-coded pictures in both versions of the content in the MCTS-based scheme. On the other hand, the FILP SHVC-ROI does not include frequent intra-coded pictures and hence requires the lowest storage compared to other on average by 98% storage of the ERP. In the CILP version of the scalable method, the storage requirement is slightly higher than of the FILP scenario, as expected.

Similarly, the Simulcast method requires to have frequent intra-coded pictures in the high-quality content, hence requires larger storage (118%) than the scalable schemes and lower storage compared to the MCTS-based method.

The SCP-based scheme requires 122% of the ERP storage as it can be realized from the Table IX. Even though the SCP-based method has less storage requirement compared to the MCTS-based method, due to the duplicated pictures in the video sequence the storage requirement is larger when compared to the scalable and Simulcast approaches.

Table X summarizes the features of the described VAS schemes in this work, in terms of compliance with OMAF requirements and streaming performance. As can be seen, and explaind in above, only the MCTS-based and SCP-based schemes are fully compliant with the OMAF's requirements for delivering the omnidirectional content. However, the streaming performance of the SCP-based method is significantly higher than the MCTS-based scheme.

IX. Conclusion

In this work, current tile-based viewport-adaptive streaming solutions for transmitting the omnidirectional video were studied: MCTS-based, SHVC-ROI and Simulcast HEVC. These methods, were considered sub-optimal or impractical for such delivery applications due to the frequent intra-coded picture requirements, multi-layer coding schemes or decoding complexity considerations.

Shared Coded Picture (SCP) based streaming scheme was proposed in this work for resolving the issues of VAS methods. The proposed method, enables the viewport switching functionality without the need for using frequent intra-coded pictures. Such functionality is achieved by shared coded pictures technique in different quality versions of the content. Moreover, the proposed method is perfectly aligned with the viewport-dependent streaming requirements of the OMAF standard. The conducted experiments demonstrated that the SCP-based method outperforms the MCTS-based VAS scheme by 11% to 14% in terms of streaming bitrate reduction with only 4% decoding complexity increase. Compared to the Simulcast method, the SCP-based approach provides up to 15% streaming bitrate reduction in coarse tiling and around 1% in the finest tiling. The proposed scheme, provided similar performance as the FILP SHVC-ROI method in the OMAF-compliant tiling scenarios. Furthermore, the decoding complexity of the SCP-based method is 56% to 28% lower from 4×2 to 12×8 tiling arrangement respectively, when compared to the SHVC-ROI and Simulcast.

As a future work, we are planning to conduct a study that addresses the effect of tiling granularity and low/high-quality tile decision in the case of user's head motion. Furthermore, future studies could investigate other aspects of the SCP-based method such as: study on the effect of different SCP intervals

TABLE IX: Storage requirements for the described VAS schemes relative to ERP (%)

Sequence	MCTS-based	FILP SHVC-ROI	CILP SHVC-ROI	Simulcast	SCP-based
AerialCity	137.4%	95.9%	96.7%	115.3%	119.0%
DrivingInCity	141.4%	112.8%	118.0%	130.8%	144.7%
DrivingInCountry	134.1%	108.1%	112.4%	126.8%	139.1%
PoleVault_le	132.5%	94.6%	97.5%	116.2%	120.4%
BearAttack	131.2%	93.6%	96.0%	110.6%	109.0%
LRRH	131.9%	86.7%	88.0%	108.8%	103.4%
Average	**134.7%**	**98.6%**	**101.4%**	**118.1%**	**122.6%**

TABLE X: Summary of VAS schemes

Feature	MCTS	SHVC	Simulcast	SCP
Single-layer	Yes	No	Yes	Yes
4K decoding	Yes	No	No	Yes
Frequent switching	Yes	Yes	Yes	Yes
Streaming performance	Medium	High	High	High
Compliant with OMAF profile	Yes	No	No	Yes

and its relation to motion-to-high-quality delay, extending SCP technique for delivering the stereoscopic omnidirectional content, utilizing SCPs in region-wise mixed resolution (RWMR) tile-based methods, etc.

REFERENCES

[1] W. Mason, "VR HMD roundup: Technical specs." http://uploadvr.com/vr-hmd-specs/. [Accessed April 2018].

[2] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "HEVC-compliant tile-based streaming of panoramic video for virtual reality applications," in *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 601-605, ACM, 2016.

[3] P. Rondao Alface, J.-F. Macq, and N. Verzijp, "Interactive omnidirectional video delivery: A bandwidth-effective approach," *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 135-147, 2012.

[4] Y. Sanchez de la Fuente, R. Skupin, and T. Schierl, "Video processing for panoramic streaming using HEVC and its scalable extensions," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 5631-5659, 2017.

[5] A. T. Nasrabadi, A. Mahzari, J. D. Beshay, and R. Prakash, "Adaptive 360-degree video streaming using scalable video coding," in *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 1689-1697, ACM, 2017.

[6] A. Zare, A. Aminlou, and M. M. Hannuksela, "Virtual reality content streaming: Viewport-dependent projection and tile-based techniques," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1432-1436, IEEE, 2017.

[7] R. Ghaznavi-Youvalari, A. Zare, H. Fang, A. Aminlou, Q. Xie, M. M. Hannuksela, and M. Gabbouj, "Comparison of HEVC coding schemes for tile-based viewport-adaptive streaming of omnidirectional video," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-6, IEEE, 2017.

[8] R. Ghaznavi-Youvalari, M. M. Hannuksela, A. Aminlou, and M. Gabbouj, "Viewport-dependent delivery schemes for stereoscopic panoramic video," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2017, pp. 1-4, IEEE, 2017.

[9] S. Lederer, "Today's and future challenges with new forms of content like 360° AR and VR," in *MPEG workshop Global Media Technology Standards for an Immersive Age*, 2017.

[10] D. Podborski, E. Thomas, M. Hannuksela, S. Oh, T. Stockhammer, and S. Pham, Virtual reality and DASH, in *International Broadcasting Convention, IBC*, 2017.

[11] R. Skupin, Y. Sanchez, Y.-K. Wang, M. Hannuksela, J. Boyce, and M. Wien, "Standardization status of 360 degree video coding and delivery," in *Visual Communications and Image Processing (VCIP)*, 2017 IEEE, pp. 1-4, IEEE, 2017.

[12] *ISO/IEC 23090-2, Information technology - Coded representation of immersive media (MPEG-I) Part 2: Omnidirectional Media Format (OMAF)*.

[13] M. M. Hannuksela, Y.-K. Wang, and M. Gabbouj, "Isolated regions in video coding," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 259-267, 2004.

[14] K. Kammachi-Sreedhar, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications, in *2016 IEEE International Symposium on Multimedia (ISM)*, pp. 583-586, IEEE, 2016.

[15] Y. Wang, R. Wang, Z. Wang, and W. Gao, "Asymmetric circular projection for dynamic virtual reality video stream switching," in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2726-2730, IEEE, 2017.

[16] G. Van der Auwera, M. Coban, Hendry, and M. Karczewicz, "AHG8: Truncated Square Pyramid Projection (TSP) for 360 video," *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0071*, October 2016.

[17] E. Kuzyakov and D. Pio, "Next-generation video encoding techniques for 360 video and VR," 2016.

[18] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1-7, IEEE, 2017.

[19] I. D. D. Curcio, H. Toukomaa, and D. Naik, "Bandwidth reduction of omnidirectional viewport-dependent video streaming via subjective quality assessment," in *Proceedings of the 2nd International Workshop on Multimedia Alternate Realities*, pp. 9-14, ACM, 2017.

[20] M. M. Hannuksela, "OMAF: viewport dependent video coding schemes," *ISO/IEC JTC1/SC29/WG11 (MPEG) document M39864*, January 2017.

[21] *Information technology - Coding of audio-visual objects Part 15: Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format*, 4 ed., February 2017.

[22] A. Zare, A. Aminlou, and M. M. Hannuksela, "6K effective resolution with 4K HEVC decoding capability for OMAF-compliant 360° video streaming," in *The 23rd Packet Video Workshop 2018*, ACM, 2018.

[23] T. Stockhammer, "DASH for TV live and low-latency services," in *Mile High Video Workshop*, August 2017.

[24] *Guidelines for Implementation: DASH-IF Interoperability Points*. version 4.2, April 2018. available: https://dashif.org/guidelines/.

[25] R. Koenen and M.-L. Champel, "Requirements MPEG-I phase 1b," *ISO/IEC JTC1/SC29/WG11, MPEG document N17331*, Jan. 2018.

[26] A. Aminlou, K. Kammachi-Sreedhar, A. Zare, and M. Hannuksela, "Testing methodology for viewport-dependent encoding and streaming," in *ISO/IEC JTC1/SC29/WG11 (MPEG) document M39081*, October 2016.

[27] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 31-36, IEEE, 2015.

[28] A. Singla, S. Fremerey, A. Raake, P. List, and B. Feiten, "Measurement of user exploration behavior for omnidirectional (360°) videos with a head mounted display," in *ITU-T Joint Video Exploration Team (JVET), document: JVET-H0050*, Oct.2017.

[29] High Efficiency Video Coding (HEVC) reference software HM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute. https://hevc.hhi.fraunhofer.de/., May. 2018.

[30] Scalable Extensions of the High Efficiency Video Coding (SHVC) reference software SHM. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute,. https://hevc.hhi.fraunhofer.de/shvc., May. 2018.

[31] E. Alshina, J. Boyce, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360 degree video," *JVETG1030, m41362*, Aug, 2017.

[32] J. Ridge, M. M. Hannuksela, E. Aksu, J. Lainema, and A. Aminlou, "Nokia test sequences for virtual reality video coding," in *ITU-T Joint Video Exploration Team (JVET), document JVET-C0064*, June 2016.

[33] F. Bossen, "Common test conditions and software reference configurations," in *Joint Collaborative Team on Video Coding (JCT-VC) of ITUT SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 5th meeting*, Jan. 2011, 2011.

[34] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," *VCEG-M33*, 2001.

Ramin Ghaznavi-Youvalari (M'16) received his M.S. degree in information technology from Tampere University of Technology (TUT), Tampere, Finland in 2016. He is currently pursuing the Ph.D. degree in signal processing at TUT.

From 2015 to 2016, he was a research assistant at TUT and an external researcher at Nokia Technologies, working on virtual reality content compression and streaming. From 2016 to 2017, he was working as a research intern at Nokia Technologies on video compression domain. Since 2017, he has been a researcher at Nokia Technologies and his work is focused on various image and video compression and streaming fields including standardization of Versatile Video Coding standard (VVC/H.266). His research interests include image and video compression, virtual reality, augmented reality and machine learning.

Alireza Zare is a Research Scientist at Nokia Technologies, Tampere, Finland. He received M.S. degree in Information Technology from Tampere University of Technology (TUT) in 2017. He is also pursuing the PhD at Laboratory of Signal Processing at TUT. His research field is mainly focused on Video Coding and Streaming and Machine Learning.

Alireza Aminlou received his B.S. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 2000, and his M.S. and PhD degrees in electrical engineering from University of Tehran, Tehran, Iran, in 2003 and 2010, respectively.

He was with Multimedia Processing Laboratory, University of Tehran, from 2003 to 2010 in different projects including hardware implementation of JPEG2000 and H.264/AVC codecs. He was visiting researcher in Tampere University of Technology, Tampere, Finland in 2011. Since 2012, he has been with Nokia Research Center and Nokia Technologies, Tampere, Finland contributing to scalable extension of High Efficient Video Coding (HEVC), streaming of Virtual Reality (VR) content, and Versatile Video Coding (VVC) projects. His research interests include hardware implementation, video compression and rate-distortion optimization.

Miska M. Hannuksela (M'03) received his Master of Science and Doctor of Science degrees from Tampere University of Technology, Finland, in 1997 and 2010, respectively.

He is currently Nokia Bell Labs Fellow and the Head of Video Research in Nokia Technologies. He has been with Nokia since 1996 in different roles including research manager/leader positions in the areas of video and image compression, end-to-end multimedia systems, as well as sensor signal processing and context extraction. He has published more than 160 conference and journal papers and hundreds of standardization contributions. He is or has been an editor in several video and systems standards, including H.264/AVC, H.265/HEVC, High Efficiency Image File Format (HEIF), ISO Base Media File Format, and Omnidirectional Media Format. His current research interests include video compression and immersive multimedia systems.

Dr. Hannuksela received the award of the Best Doctoral Thesis of the Tampere University of Technology in 2009. He has co-authored several papers awarded in international conferences. He was an Associate Editor of the IEEE Transactions on Circuits and Systems of Video Technology from 2010 to 2015.

Moncef Gabbouj (F'11) received his BS degree in electrical engineering in 1985 from Oklahoma State University, Stillwater, and his MS and PhD degrees in electrical engineering from Purdue University, West Lafayette, Indiana, in 1986 and 1989, respectively.

Dr. Gabbouj is a Professor of Signal Processing at the Department of Signal Processing, Tampere University of Technology, Tampere, Finland. He was Academy of Finland Professor during 2011-2015. He held several visiting professorships at different universities. Dr. Gabbouj is currently the TUT-Site Director of the NSF IUCRC funded Center for Visual and Decision Informatics. His research interests include Big Data analytics, multimedia content-based analysis, indexing and retrieval, artificial intelligence, machine learning, pattern recognition, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Dr. Gabbouj is a Fellow of the IEEE and member of the Academia Europaea and the Finnish Academy of Science and Letters. He is the past Chairman of the IEEE CAS TC on DSP and committee member of the IEEE Fourier Award for Signal Processing. He served as Distinguished Lecturer for the IEEE CASS. He served as associate editor and guest editor of many IEEE, and international journals.

Dr. Gabbouj was the recipient of the 2017 Finnish Cultural Foundation for Art and Science Award, the 2015 TUT Foundation Grand Award, the 2012 Nokia Foundation Visiting Professor Award, the 2005 Nokia Foundation Recognition Award, and several Best Paper Awards. He published two books and over 700 journal and conference papers and supervised 45 doctoral and 58 Master theses.