OVERVIEW OF THE MULTIVIEW HIGH EFFICIENCY VIDEO CODING (MV-HEVC) STANDARD

Miska M. Hannuksela¹, Ye Yan², Xuehui Huang², and Houqiang Li²
¹Nokia Technologies, ²University of Science and Technology of China

ABSTRACT

This paper reviews the multiview extension (MV-HEVC) of the High Efficiency Video Coding (HEVC) standard. MV-HEVC is capable of multiview video coding with or without accompanying depth views. The key design concepts and design elements of MV-HEVC are described in the paper. Furthermore, the features and characteristics of MV-HEVC compared to other standardized video codec extensions for three-dimensional (3D) video coding are reviewed.

Index Terms— HEVC, MV-HEVC, 3D video coding

1. INTRODUCTION

While stereoscopic content is most commonly used in today's 3D video content and services, an increasing amount of attention has been paid to depth-enhanced and multiview 3D video. Depth or disparity information can be used in depth-image-based rendering (DIBR) [1] to synthesize views as decoder-side post-processing. When applied with stereoscopic displays, view synthesis can be used for adjusting the disparity between the displayed views according to viewers' preferences and view conditions, such as viewing distance and display size, in order to reach as comfortable 3D experience as possible. Depth-enhanced multiview video together with DIBR can also be used with multiview auto-stereoscopic displays to generate a required number of views for displaying from the received views. The multi-view-video-plus-depth (MVD) format refers to a data format with more than one texture view, each paired with the depth view of the same viewpoint [2].

The two most recent international video coding standards, namely the Advanced Video Coding (AVC) standard (also known as H.264) [3] and the High Efficiency Video Coding (HEVC) standard (also known as H.265) [4], were initially intended for two-dimensional (2D) video. Multiview extensions for both AVC and HEVC have been developed, referred to as Multiview Video Coding (MVC) [3][5] and Multiview HEVC (MV-HEVC) [4], respectively. For depth-enhanced multiview video coding one can use AVC extensions referred to as the multiview video and depth coding (MVC+D) [3][6] and the multiview and depth video with enhanced non-base view coding (3D-AVC) [3], and the HEVC extensions MV-HEVC and 3D-HEVC [7][8]. A fundamental principle of MVC+D and MV-HEVC is to re-use the coding tools of the underlying 2D coding, i.e. AVC and HEVC, respectively, so that implementations can be realized by software changes to high-level syntax in the slice header level and above. On the contrary, 3D-AVC and 3D-HEVC introduce new low-level coding tools and aim at improved compression efficiency compared to the MVC+D and MV-HEVC, respectively.

Section 2 of this paper reviews the design of MV-HEVC. Compared to earlier reviews, such as [9], the paper describes the design of the approved MV-HEVC standard. Section 3 of the paper presents a comprehensive feature comparison between 3D video coding standards. Finally, conclusions are provided in Section 4.

2. MULTI-LAYER HEVC EXTENSIONS

The development of the scalable extensions (SHVC) [4][9][10] of HEVC started when the MV-HEVC project was already ongoing. In an early phase of the SHVC development, it was decided to use an approach requiring only high-level syntax changes as well as inter-layer processing. Furthermore, since SHVC and MV-HEVC both shared the fundamental principle of using only the HEVC coding tools for slice data, their design was unified [11][12][13]. This high-level syntax principle is achieved by enabling the inclusion of pictures originating from direct reference layers in the reference picture list(s) used for decoding pictures of predicted layers, while otherwise these inter-layer reference pictures are treated identically to any other reference pictures. Eventually, both SHVC and MV-HEVC were released as parts of HEVC version 2 [4] and share the same specifications of multi-layer extensions. As one consequence of this unified approach, views are treated as scalable layers like any other types of scalability.

An elementary unit in the HEVC bitstreams is called a network abstraction layer (NAL) unit, consisting of a header and a payload. The NAL unit header contains a 5-bit NAL unit type, consisting of a header and a payload. The NAL unit header contains a 5-bit NAL unit type, 6-bit layer identifier called nuh_layer_id, and a 3-bit temporal sub-layer identifier. A video parameter set (VPS) NAL unit specifies a mapping of the layer identifiers to values of scalability dimensions, including a dependency identifier for SHVC, a view order index for MV-HEVC and 3D-HEVC, a depth flag for 3D-HEVC, and an auxiliary
layer type identifier, referred to as AuxId, which may be used in any multi-layer bitstream. A selectable number of scalability dimensions can be used in a multi-layer bitstream. The VPS also indicates a mapping from each view order index to a view identifier value, where each unique view identifier value represents a distinct viewpoint. Temporal sub-layers are used for temporal scalability, i.e. they provide the capability of extracting sub-bitstreams of different picture rates. For more information on the high-level syntax and temporal scalability of HEVC, the reader is referred to [14].

Auxiliary picture layers, originally proposed in [15], enable multiplexing of supplemental coded video into a bitstream that also contains the primary coded video. It was decided to enable depth views with the auxiliary picture layer mechanism in MV-HEVC. AuxId value equal to 2 associated, in VPS, with a layer indicates that the layer represents a sequence of depth pictures. More detailed properties on the depth auxiliary layers, such as the depth or disparity range represented by the sample values, can be provided with the depth representation information supplemental enhancement information (SEI) message [4]. Depth auxiliary layers are also associated with a view order index and view identifier; hence, multiview depth coding is supported in MV-HEVC. It is noteworthy that 3D-HEVC uses the depth flag scalability dimension to indicate depth views. This is because 3D-HEVC enables the use of depth information for non-base texture views. Auxiliary picture layers, on the other hand, are not allowed to affect the decoding of the primary video. As a consequence, depth views of MV-HEVC and 3D-HEVC are not compatible.

As scalable multi-layer bitstreams enable decoding of more than one combinations of layers and temporal sub-layers, the multi-layer HEVC decoding process is given as input a target output operation point, specifying the output layer set (OLS) and the highest temporal sub-layer to be decoded. An OLS represents a set of layers, which can either be necessary or unnecessary layers. A necessary layer is either an output layer, meaning that the pictures of the layer are output by the decoding process, or a reference layer, meaning that its pictures may be directly or indirectly used as a reference for prediction of pictures of any output layer. The VPS includes a specification of OLSs, and can also specify buffering requirements and parameters for OLSs. Unnecessary layers are not required to be decoded for reconstructing the output layers but can be included in OLSs for indicating buffering requirements for such sets of layers in which some layers are coded with potential future extensions.

Multi-layer HEVC extensions support hybrid codec scalability, in which the base layer is coded as a separate non-HEVC bitstream. The multi-layer HEVC decoding process inputs decoded base layer pictures and certain properties for them. The base layer bitstream and HEVC enhancement bitstream are separate, and it is up to the systems layer to handle synchronization between the bitstreams as well as the connection between the base and enhancement layer decoding processes. The hybrid codec scalability feature enables 3D services that are compatible with 2D AVC decoding. For example, broadcast services could provide conventional 2D service for AVC-capable devices and build 3D capability using either MV-HEVC or 3D-HEVC on top of the AVC service.

While earlier video coding standards specified profile-level conformance points applying to a bitstream, multi-layer HEVC extensions specify layer-wise conformance points. To be more exact, a profile-tier-level (PTL) combination is indicated for each necessary layer of each OLS, while even finer-grain temporal-sub-layer-based PTL signaling is allowed. Consequently, decoder capabilities can be indicated as a list of PTL values, where the number of list elements indicates the number of layers supported by the decoder. Non-base layers that are not inter-layer predicted can be indicated to conform to a single-layer profile, such as the Main profile, while they also require so-called independent non-base layer decoding (INBLD) capability to deal correctly with layer-wise decoding. The Multiview Main profile was specified for MV-HEVC and suits both non-base texture views and non-base depth views.

Table I shows an example of a bitstream containing three texture views and three depth views, both coded with the so-called IBP inter-view prediction order, where left-side view (I view) is coded independently of other views, the right-side view (P view) may utilize inter-view prediction from the I view, and center view (B view) may be predicted from both the I and P views. As can be seen the view order index values of the respective views (left, center, right) of texture and depth are identical. The VPS is used to specify the mapping view order index and AuxId values to nuh_layer_id values, on which Table I shows one example where depth views follow in decoding order the texture views, while other mappings and decoding orders would also be possible. The reference layers are indicated in the VPS according to the IBP inter-view prediction pattern. In this example, four OLSs are specified in the VPS. The 0th OLS contains only the texture base view conforming to the Main profile. The first and second OLSs represent a stereoscopic video with wide (side views) and narrow (adjacent views) baseline, respectively. As can be seen from Table I, the P view is not an output layer but is a necessary layer in the narrow baseline OLS. Each necessary non-base view conforms to the Multiview Main profile. In the fourth OLS, all texture and depth views are output. The independently coded depth view is indicated to conform to the Main profile with the INBLD capability, while all predicted views comply with the Multiview Main profile.
3. COMPARISON OF MVD CODING STANDARDS

Table II presents a comparison of the features of the MVD coding scenarios overviewed in Section 1. The following sub-sections provide further details on the information provided in Table II with a specific focus on MV-HEVC. The goal of this section is to assist in selecting a 3D coding format that suits the properties and purposes of the used video acquisition equipment, display, and application.

3.1. Rate-distortion (RD) performance

The RD performance between 3D-HEVC and MV-HEVC was analyzed in [16] with 3-view MVD coding using view synthesis optimization (VSO) [8] in the encoding. The JCT-3V common test conditions [17] were used, according to which the RD performance of the texture views as well as three synthesized views between each pair of coded views is compared. It was found that 3D-HEVC reduced the bitrate by about 14% and 20% for coded texture views and synthesized views, respectively, compared to MV-HEVC in terms of the Bjontegaard delta bitrate (dBR) [18]. No results for a 2-view scenario were provided in [16], but it can be assumed that the bitrate reduction of 3D-HEVC compared to MV-HEVC would be roughly halved. Furthermore, it is mentioned in [16] that only one depth view is coded for MV-HEVC. Encoder-side optimizations and verification of the RD performance difference between MV-HEVC and 3D-HEVC are therefore subjects of further studies.

3.2. Data format

Uncompressed depth information can be obtained for camera-captured content generally in two ways, either through depth estimation or by using depth sensors. In a depth estimation process, the respective pixels in adjacent views are searched, resulting into a disparity picture or, equivalently, to a depth picture. Many depth sensors produce pictures having a significantly lower spatial resolution than the resolution produced by typical color image sensors. Depth sensors and color image sensors are typically located adjacent-ly, i.e. the viewpoint of the depth pictures differs from the viewpoint(s) of the color image sensors. The data format called unpaired MVD, introduced in [17], comprises texture and depth views that need not represent the same viewpoints. The unpaired MVD format provides flexibility in depth acquisition and reduced complexity in encoder-side pre-processing.

All depth-enhanced coding formats support the MVD and the unpaired MVD data formats. When using the unpaired MVD format texture coding tools using depth views have to be selectively turned off in 3D-AVC and 3D-HEVC. As there no dependencies from depth views to texture views or vice versa in MVC+D and MV-HEVC, the selection of the transmitted views e.g. for bitrate adaptation can be performed flexibly e.g. resulting into bitstreams with an unequal number of texture and depth views. 3D-AVC and 3D-HEVC require more careful view extraction to ensure that cross-component prediction dependencies are obeyed in the transmitted bitstream.

When only a limited amount of receiver-side disparity adjustment is required, unpaired MVD can be used to improve RD performance as studied in [20] and [21] for MVC+D and 3D-AVC, respectively. As no similar study was made for MV-HEVC, an experiment was performed for
this paper. JCT-3V common test conditions (CTC) [17] were used for coding MV-HEVC content with two texture views and one depth view. The baseline was adjusted by 10% in the decoding end through the view synthesis arrangement described in [20]. The results are reported in Table IIIa, indicating that with the exception of one sequence, the transmission of one depth view provides improved compression efficiency. The unpaired MVD format can therefore be justified to be used with MV-HEVC from the RD performance point of view, when the disparity adjustment needs are limited.

3.3. Mixed spatial resolution

Mixed resolution between texture and depth views may be desirable to achieve reduced computational complexity in cases where the acquired depth resolution is inherently smaller than the texture resolution, e.g. when a depth sensor has been used to acquire a depth map sequence. As can be observed from Table II, the codecs are similar when it comes to supporting a mixed resolution between texture and depth views. 3D-HEVC includes such depth-based texture coding tools that require the same spatial resolution being applied both in texture and depth pictures. Hence, in order to support unequal spatial resolutions between texture and depth views, 3D-HEVC encoders need to turn off depth-based texture coding tools.

A proper selection of depth resolution can provide compression benefits when considering both coded and synthesized views as studied in [22] and [23] in the context of MVC+D and 3D-AVC, respectively. Early results for MV-HEVC were provided in [24], while for this paper we performed a new set of simulations with a recent version of the reference software, HTM 12.1 [19], with depth pictures having half of the resolution vertically and horizontally compared to the texture pictures. The resampling was performed as explained in [25], and the JCT-3V CTC [17] was used. As can be seen from Table IIIb, the bitrate is reduced by more than 5% on average, while the RD performance of the synthesized views remain unchanged. Even though VSO was not used in these simulations, they are indicative that a depth resolution lower than texture resolution can provide a good trade-off between RD performance and complexity.

Mixed-resolution stereoscopic video has been broadly studied e.g. in [26][27][28][29], as it could potentially achieve a comparable quality with symmetric-resolution stereoscopic video but reduce computational complexity thanks to a smaller number of samples processed. The compared MVD coding extensions use pictures of other views without resampling as inter-view reference pictures, and hence none of them supports mixed-resolution coding when inter-view prediction is applied. However, both MV-HEVC and 3D-HEVC support views having different resolutions, when no prediction takes place between the views.

3.4. Compatibility

The MVD coding extensions provide seamless compatibility with 2D video decoding. The HEVC extensions have the additional benefit that they can use a base view of any coding standard, such as AVC. This feature enables i) an efficient use of hardware designs including both AVC and HEVC decoders for stereoscopic services as well as ii) establishment of 3D video services providing compatibility with AVC-capable 2D clients while using HEVC-based coding technology for non-base views and depth views to keep the additional bitrate as low as possible. Similarly, it may be desired to support clients with stereoscopic decoding capability and other clients with multiview decoding capability with the same content or service. It can be seen from Table II that 3D-AVC and 3D-HEVC leave it up to the encoder to determine whether the second view conforms to MVC and MV-HEVC, respectively, or requires dedicated coding tools of 3D-AVC and 3D-HEVC, respectively.

4. CONCLUSIONS

The paper reviewed the design of the MV-HEVC standard, which introduces only high-level changes over the HEVC standard, facilitating implementations with software updates to existing HEVC codecs. Compared to the 3D-HEVC approach of introducing dedicated 3D coding tools, MV-HEVC is inferior in rate-distortion performance but is easier to implement and provides more flexibility in terms of the data format and scalability, as explained in the paper.
REFERENCES


[19] HTM reference software, online: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/