Standardization Status of 360 degree Video Coding and Delivery

R. Skupin¹, Y. Sanchez², Y.-K. Wang³, M. M. Hannuksela⁴, J. Boyce⁵, M. Wien⁶

¹Fraunhofer HHI, Germany, ²Qualcomm Incorporated, USA, ³Nokia Research Center, Finland, ⁴Intel, USA, ⁵RWTH Aachen, Germany

robert.skupin@hhi.fraunhofer.de

Abstract—The emergence of consumer level capturing and display devices for 360 degree video creates new and promising segments in entertainment, education, professional training, and other markets. In order to avoid market fragmentation and ensure interoperability of 360 degree video ecosystems, industry and academia cooperate in standardization efforts in this field. In the video coding domain, 360 degree video invalidates many established procedures, e.g., concerning evaluation of the visual quality, while the specific content characteristics offer potential for higher compression efficiency beyond the current standards. Likewise, 360 degree video puts stricter demands on the system level aspects of transmission but may also offer the potential to enhance existing transport schemes. The Joint Collaborative Team on Video Coding (JCT-VC) as well as the Joint Video Exploration Team (JVET) already started investigations into 360 degree video coding while numerous activities in the Systems subgroup of the Moving Picture Experts Group (MPEG) started to investigate application requirements and delivery aspects of 360 degree video. This paper reports on the current status of the outlined standardization efforts.

Index Terms—360 degree video, omnidirectional video, HEVC, projection, region-wise packing, standardization

I. INTRODUCTION

Equipment for capturing 360 degree video has been available for some time but was usually custom made, lacking in resolution and/or located in the upper price range. Means for delivery and rendering of this type of content were also tailored to the needs of specific niches. Recently, however, a lot of this equipment reached consumer level price range, mainly triggered through the emergence of a new generation of Head Mounted Displays (HMDs), e.g., based on mass market mobile phone components. As evidenced by the growth of user captured content on today’s over-the-top (OTT) video streaming platforms, new and promising segments of entertainment, education, professional training, and other market have emerged.

In order to avoid the fragmentation of these new market segments and to ensure interoperability of 360 degree video ecosystems, many relevant standard development organizations (SDOs) have started work in their scope in the context of 360 degree video. For example, 3GPP initiated the work on Virtual Reality (VR) media services over 3GPP with a technical report [1], and the Digital Video Broadcasting (DVB) Project followed up on a promising study mission report by establishing a VR related commercial module [2]. In this paper however, we focus on the Video Coding Experts Group (VCEG, ITU-T Q6/16) and the Moving Picture Experts Group (MPEG, ISO/IEC JTC 1/SC 29/WG 11). VCEG and MPEG started investigations into the topic and are expected to lead, among all SDOs working on 360 degree video related standardization, in the areas of 360 degree video coding and delivery. In their respective joint groups, i.e., the Joint Collaborative Team on Video Coding (JCT-VC), responsible for developing the High Efficiency Video Coding (HEVC) standard and its extensions [3], and the Joint Video Exploration Team (JVET) that investigates new video coding approaches for coding efficiency beyond HEVC [4], coding of 360 degree video has gained attention. Delivery and other system aspects are the main focus of MPEG Systems subgroup's work on the Omnidirectional MediA Format (OMAF), which is envisioned to become Part 2 of the emerging ISO/IEC 23090 MPEG-I group of standards on Immersive Media.

This paper reports on the status of the VCEG and MPEG standardization activities outlined above and is structured as follows. Section II reports on the current status of the activities related to 360 degree video coding in JVET and JCT-VC, while Section III provides details on the activities in the groups that focuses on system aspects such as MPEG-I and OMAF.

II. 360 DEGREE VIDEO CODING

Some of the established paradigms and procedures in video coding need adaptation to the context of 360 degree video. One reason is the additional stitching and projection of video content on the capture side and the corresponding steps on the render side, which can be considered as encoder pre- and decoder post-processing. Apart from the need for signalling of projection metadata, further implications arise.

For example, unlike traditional video coding, the visual angle of a video sample or pixel, i.e., its contribution to visual quality, may vary significantly depending on position within the underlying projection, e.g., in the Equirectangular Projection (ERP), whereas the traditional Cube Map Projection (CMP) suffers less distortion in this sense. In addition to that, the user is typically presented only a video subset, which is often referred to as the viewport. This requires quality evaluations in context of 360 degree video to be different from traditional rectangular video as detailed in [5]. Sufficiently high
quality within the viewport can potentially be achieved through more efficient viewport-dependent media coding and delivery schemes.

A. Joint Collaborative Team on Video Coding (JCT-VC)

In the context of 360 degree video, JCT-VC started the process of standardizing signalling means in the form of additional Supplemental Enhancement Information (SEI) messages for HEVC in late 2016 [6], which are expected to be finalized in late 2017. SEI messages do not have normative impact on the decoding process, i.e., they can be stripped from the coded bitstream without altering the decoder output, however, the conveyed information may be either helpful or essential for rendering and other post-decoder operations, or may be used by the system layer.

First, two projection related SEI messages provide information necessary to enable remapping of coded video samples onto a three-dimensional coordinate space in angular and Cartesian coordinates for use when rendering a 360 degree ERP and CMP video. These SEI messages which are referred to as the Equirectangular Projection SEI message and the Cubemap Projection SEI message also include signalling of a rotation before projection such that a particular viewport can be placed on a certain position on the projected picture. One use case for this signalling is to benefit from increased coding efficiency by shifting video areas to less distorted regions of the projected picture, e.g. the ERP equator, in order to allow for better motion compensated prediction. Stereoscopic 360 degree video is supported based on the regular frame packing arrangement signalling in HEVC.

In addition to projection, an SEI message for region-wise packing (RWP) is specified which is an intermediate step between projection and en-/decoding. RWP enables image manipulation (e.g. scaling, translation, rotation, and mirroring) of any rectangular region of the projected picture and can be used to circumvent weaknesses of projections such as oversampling towards the poles in ERP as depicted in the top of Figure 1. RWP further allows to describe videos with a preferred viewport orientation in which certain content areas are subsampled as shown in the bottom of Figure 1 based on CMP.

As capturing, delivering, and consuming spherical video with a coverage of less than 360 degree are an important usage scenario as well, e.g. to reflect extraction of a recommended viewport, means for indicating limited coverage of a projected picture are given as well. For ERP, the Equirectangular Projection SEI includes signalling of azimuth and elevation boundaries, while for CMP, RWP can be used to indicate coverage restrictions by omitting regions.

Third, the Omnidirectional Viewport SEI message allows signalling of the spherical coordinates of one or more viewports recommended for display. This SEI message can be used to indicate distinguished Regions-of-Interest in a spherical video such as a director’s cut or a generally recommended viewport.

Last, two SEI messages relate to independently encoded spatial picture subsets within HEVC bitstreams, so called Motion-Constrained Tile Sets (MCTS), which allow trivial rewriting of these subsets into HEVC conformant bitstreams in the coded domain without entropy coding impact. Using the MCTS Extraction Information Sets SEI message and the MCTS Extraction Information Nesting SEI message, a specified sub-bitstream extraction process can be carried out, e.g., for extraction of a recommended viewport from a full 360 degree video bitstream coded using MCTS.

B. Joint Video Exploration Team (JVET)

The work on 360 degree video related topics in JVET started in early to mid 2016 with the contribution of the first test sequences and focused mainly on the evaluation of visual quality and the impact of projections on coding efficiency. Since then, JVET has established an extensive description of the proposed projection formats and the conversion between them [7] as well as a Common Test Conditions document allowing comparable experiments [8]. From the set of all proposed projections, nine were integrated into the 360Lib software [9] for conversion between formats and further evaluation. The integrated projections range from traditional ones such as ERP and CMP, which are well established in the field of computer graphics, to more sophisticated projection types using relatively complex geometries, e.g., an Icosahedron with 20 surface planes and 30 edges, or projections with a preferred viewing direction such as the Truncated Square Pyramid. Along with the introduction of various projection formats, a continuous effort to improve on the individual packing formats of each projection has been made where applicable. Note that this packing, i.e., the arrangement of geometry faces on the video picture plane, is intended to minimize visual discontinuities between adjacent surfaces arranged in a compact manner as opposed to traditional frame-packing used in stereoscopic video to place multiple camera views on a single picture plane.

JVET also investigated several video quality metrics incorporating specific aspects of 360 degree video. In traditional video, the whole picture plane is presented to the user and, hence, each coded video sample contributes equally to the objective quality of a picture, e.g. for calculation of PSNR. As 360 degree projections potentially vary the importance of samples over the picture plane that is also only partially presented to the user, new strategies for quality evaluations were proposed. First, weighting the contribution of individual samples and second, usage of non-uniform sampling of the picture plane. As a third approach, comparisons of rendered rectilinear viewports were considered.
Relative few contributions thus far have dealt with coding tools or encoder optimization related to 360 degree video, out of which none were integrated into the official JVET test model so far [7]. One approach seen in several contributions is to modify encoder control of the quantization parameter, depending on known projection format characteristics such as the visual angle of samples in ERP much like the weighting in quality metrics [10], [11]. Another coding tool proposal targets projections based on geometries with planar surfaces by extending the packed surfaces in a geometrically correct fashion before motion estimation [12]. The improved cross-surface motion compensated prediction reportedly led to gains in overall compression efficiency. Comparatively simpler padding strategies targeting picture boundaries or surface plane edges were also proposed.

Within the April 2017 joint Call for Evidence (CfE) towards coding efficiency improvements beyond HEVC, 360 degree video coding was specified as one out of three test cases, the other two being Standard Dynamic Range (SDR) and High Dynamic Range (HDR) content. Responses to the CIE proved existence of such technology and subjective tests showed a bitrate reduction by 40-50% compared to HEVC for SDR and HDR but lower reduction in the 360 degree video test case [13]. A consequently planned joint Call for Proposals is expected to attract submissions for evaluation in April 2018.

III. SYSTEM INTEGRATION OF 360 DEGREE VIDEO

From the systems perspective, interoperable 360 degree video streaming services not only require media coding format definitions, i.e., the video and audio codecs conformance points, but also need signalling and encapsulation of 360 degree video content in a container file format as well as in a streaming delivery format. This allows clients to distinguish 360 degree video from traditional video, and identify the media properties that allow correct rendering and an overall satisfactory user experience.

A. MPEG-I

The MPEG standards related to immersive media technologies are envisioned to form the ISO/IEC 23090 MPEG-I group of standards. At the time of this writing, MPEG-I is expected to consist of seven parts: 1) Technical Report on Immersive Media, 2) OMAF, 3) Immersive Video, 4) Immersive Audio, 5) Point Cloud Compression, 6) Metadata, and 7) Metrics. The OMAF standard will specify storage and delivery formats suitable for 360 degree video applications. The associated requirements document [14] lists a set of requirements for the first edition of the OMAF specification, amongst which is a requirement for building on top of existing MPEG storage and delivery formats. This specifically includes MPEG Dynamic Adaptive HTTP Streaming (DASH) [15], MPEG media transport (MMT), and the ISO Base Media File Format (ISOBMMF) [16]. Furthermore, it includes requirements related to the coded video bitstreams in initial OMAF systems as for example the distribution of 360 degree video content beyond 4k@60fps, i.e., exceeding the conformance point of HEVC Main10 Profile, Main Tier, Level 5.1.

B. Omnidirectional Media Format (OMAF)

In late 2015, contributions of related use cases and requirements ultimately led to the start of the OMAF work in early 2016 for specifying an omnidirectional media format focusing on 360 degree video (referred to as omnidirectional video in OMAF) and its associated audio [17]. The established timeline for the work of OMAF targets to address basic mono- and stereoscopic 360 degree video services first in a so-called phase 1a of MPEG-I by the end of 2017. This type of applications is often characterized as three Degrees of Freedom (3DoF), i.e., enabling changes of yaw, pitch and roll of a rendered client viewport only. More elaborated services that also allow for slight or significant translational changes of the user position within the 360 degree video can be characterized as 3DoF+ and 6DoF, respectively, and are expected to be addressed in the later phases 1b and 2 of the MPEG-I work.

Figure 2 illustrates the architecture of OMAF services depicting the flow of video and audio signals from the capturing side (top) to rendering and presentation on client side (bottom). Amongst others, the OMAF specification defines conformance points for the encoded media signals (E) and the ISOBMMF encapsulated segments (F). In addition to the projected omnidirectional video processing workflow depicted in Figure 2, OMAF also enables storage and transmission of video from cameras with one or more fisheye lens omitting stitching and projection steps on the capturing side. The circle images captured by each fisheye lens are directly placed on a common video picture before encoding. Along with such pictures, the fisheye video properties are signalled to enable correct rendering at the receiver side. The main use case for the fisheye video support is low cost equipment enabling user generated VR content.

Given its simplicity and wide adoption in the industry, ERP and CMP are the only adopted projection format at the time of writing, while several other projection formats were under consideration for OMAF [18]. OMAF also relies on RWP and specifies respective signalling in file format boxes. In addition to signalling of projection format and rotation, content coverage, and RWP, OMAF defines a further set of metadata signalling means that enable application optimizations. For instance, the initial viewpoint metadata enables the content author to indicate that

Figure 2: OMAF architecture overview [17].
a particular viewport orientation is recommended when random accessing from a point in the stream or even when playing the video continuously. Further signaling of recommended viewport metadata enables the content author to indicate that a particular viewport trace is recommended for playback such as a director’s cut for best VR story telling. The region-wise quality ranking enables 360 degree video streaming optimizations in a viewport-dependent fashion, e.g. by stream switching or based on MCTS with multiple qualities or resolutions. The guard band signalling allows including additional pixels at projection geometry edges when packing surface planes within the picture before encoding, which reduces scene artifacts in rendered 360 degree video due to projection or RWP. It is worth noting that OMAF requires relevant metadata such as projection characteristics to be carried simultaneously in file format boxes as well as in SEI messages in the elementary stream to address the needs of a wide range of implementation environments.

At the time of this writing there are two HEVC media profiles in the OMAF standard: a viewport-independent baseline profile and a viewport-dependent baseline profile. Both profiles specify a media profile for video conformant to HEVC Main 10 Profile, Main Tier, Level 5.1. The viewport-independent profile only supports video using ERP without emphasis on a given viewing direction. The encoded bitstream is encapsulated in a single ISOBMFF track.

The viewport-dependent profile supports video using ERP and CMP while utilizing advanced file format tools in the form of extractors as defined in Annex A of [16]. Along with other tools, extractors enable MCTS based viewport-dependent 360 degree video streaming while requiring only a single HEVC conforming bitstream to be decoded. Each of the tiles is offered as a subpicture carried in one HEVC track with sample entry type 'hvc1', and the file format parser constructs a single HEVC conforming bitstream out of an extractor track with sample entry type 'hvc2' and a set of subpicture tracks based on explicit reconstruction [16]. The reconstruction is achieved by replacing parameter sets and slice headers, without the need of modifying entropy-coded syntax elements individually, hence processing is lightweight. Requiring only a single HEVC conforming bitstream to be decoded may be desirable because many devices do not support simultaneous decoding of multiple separate bitstreams, and there is no conformance point definition of device capabilities or characteristics across bitstreams for simultaneous decoding. Note that reconstruction may arrange subpictures or tiles differently than in the original projected picture, possibly including resampled subpictures for emphasizing a viewport orientation as depicted on the bottom of Figure 1.

The DASH integration of OMAF is based on an extensibility tool of DASH that allows additional property descriptors to be defined. Hence, OMAF defines descriptors that expose 360 degree video related information in the Media Presentation Description (MPD) to inform clients about the media properties.

The projection information is signalled using the ProjectionFormat descriptor. This descriptor indicates the projection type used, e.g., type 0 indicates ERP. A client can determine whether it supports the projection format of the offered 360 degree video bitstream and whether to request the content. Besides, as aforementioned, there are usage scenarios in which less than the full 360 degree of a sphere is covered. In this case, the ContentCoverage descriptor can be added to the MPD. When region-wise packing is used, its usage is indicated by the RegionwisePacking descriptor. This descriptor serves as a high-level description of the RWP by indicating the type of RWP used in the offered video, e.g. based on rectangular regions. Finally, when viewport-dependent solutions are used, some areas are emphasized in comparison with other. In order for a DASH client to choose a representation dependent on the current user viewport, the RegionQualityRanking descriptor can be added to the MPD to give an indication of the quality of individual areas within the video picture.

REFERENCES
[1] 3GPP TR 26.918 v.1.0.0, “Virtual Reality (VR) media services over 3GPP (Release 15)”, June 2017
[9] 360Lib, online: https://jvet.lti.fraunhofer.de/svn/svn_360Lib/trunk