MPEG-I CODING PERFORMANCE IN IMMERSIVE VR/AR APPLICATIONS

Gauthier Lafruit[1], Arnaud Schenkel[1], Christian Tulvan[2], Marius Preda[2], Lu Yu[3]

[1]LISA, Université Libre de Bruxelles, Belgium

[2]ARTEMIS, Telecom SudParis, CNRS Samovar, France

[3]Institute of Information and Communication Engineering, Zhejiang University, China

ABSTRACT

After decennia of developing leading-edge 2D video compression technologies, MPEG is currently working on the new era of coding for Immersive applications, referred to as MPEG-I. It ranges from 360-degree video with head-mounted displays to free navigation in 3D space, with head-mounted and 3D light field displays.

Two families of coding approaches, covering typical industrial workflows, are currently considered for standardisation – Multiview + Depth Video Coding and Point Cloud Coding – both supporting high-quality rendering at bitrates of up to a couple of hundreds of Mbps.

This paper provides a technical/historical overview of the acquisition, coding and rendering technologies considered in the MPEG-I standardization activities.

INTRODUCTION

The MPEG standardisation committee is currently working on MPEG-I coding technologies to support immersive applications, MPEG-I (1), where multimedia content can be viewed from various viewpoints, different from the camera acquisition viewpoints, therefore supporting free navigation around regions of interest in the scene, e.g. circling around a player in a sports event, similar to The Matrix bullet time effect, Karthikeyan (2).

MPEG-I ranges from 360-degree video on head-mounted displays (extension of existing video codecs with Supplemental Enhancement Information (SEI) messaging for the projection format, and the Omnidirectional Media Format – OMAF - to be standardized by end 2018) supporting head movements with 3 Degrees of Freedom (3DoF), extensions thereof supporting motion parallax within some limited range around the central viewing/camera position (referred to as 3DoF+, expected to be standardized beginning 2019), as well as larger ranges of freedom of movement, eventually achieving full 6 Degrees of Freedom (6DoF) allowing any user viewing position in 3D space, with standards to be accepted by industry around 2020, Koenen (3).

Competitive coding technologies for advanced VR/AR and light field display devices are under study, encompassing EquiRectangular video Projection (ERP), MultiView + Depth (MVD) Coding, as well as Point Cloud Coding (PCC), where the former two are familiar to video-based production workflows (e.g. 3D film production) and the latter to 3D graphics-based workflows (e.g. 3D game production), both steadily evolving towards Cinematic VR/AR.

MPEG has issued several Calls for Test Material, Exploration and Core Experiments for comparing the relative merits of technologies from industrial proponents around the world, supporting 3D extensions of High Efficiency Video Coding (HEVC), Sullivan et al. (4), and Versatile Video Coding (VVC), MPEG Press Release (5), for MultiView + Depth (MVD) Coding in video production, as well as Octree- and kd-based 3D data representations used for Point Cloud Coding (PCC) in early versions of Lidar devices, Schnabel et al. (6).

VVC, which will be finalized in 2020 and will probably have inborn-support for 360-degree video. It is planned that 3DoF+ will be supported in the short term by market-existing 2D video codec devices adding supplementary metadata, while 6DoF may need enhanced coding tools in the longer term to handle even larger volumes of data. In that respect, the maturity of existing technologies for PCC, assessed after a Call for Proposal issued by MPEG in 2017, conducted the committee to start building the technical specifications for this coding approach with the target to publish the final standard early 2020.

The MVD video coding technologies for MPEG-I are under exploration in the MPEG Video Group, while PCC technologies are studied in MPEG 3DG (3D Graphics Group). Both types of technologies are grouped under the MPEG-I umbrella since they contribute to the common goal of addressing immersive applications. The two subgroups, however, have historically started their activities independently of each other, using their own data sets and Common Test Conditions (CTC), but we will see in the remainder of the paper that cross-fertilisation has led to technologies showing stunning similarities.

Both, the MPEG-I Video and MPEG-I Graphics coding technologies, are even expected to reach similar bitrates of around a couple of hundreds of Mbps for high-end Cinematic VR/AR productions, irrespective of the technological specificities of the proposed coding approaches. The coding technology choices will hence merely depend on the workflows (purely video vs. computer graphics based special effects) of the industrial players and their immersive product features (3DoF+ versus 6DoF) flooding the market.

MPEG-I PROCESSING & CODING PIPELINE

Figure 1 shows the generic processing and coding pipeline in a typical MPEG-I immersive application, seamlessly integrating video- and graphics-based approaches.

The input camera feeds are pre-processed for colour correction, distortion removal, depth estimation and/or point cloud extraction, before being compressed and transmitted, eventually accompanied by some meta-data. These data streams reach a specific bitrate that is the first (horizontal) axis in the system’s performance figure.

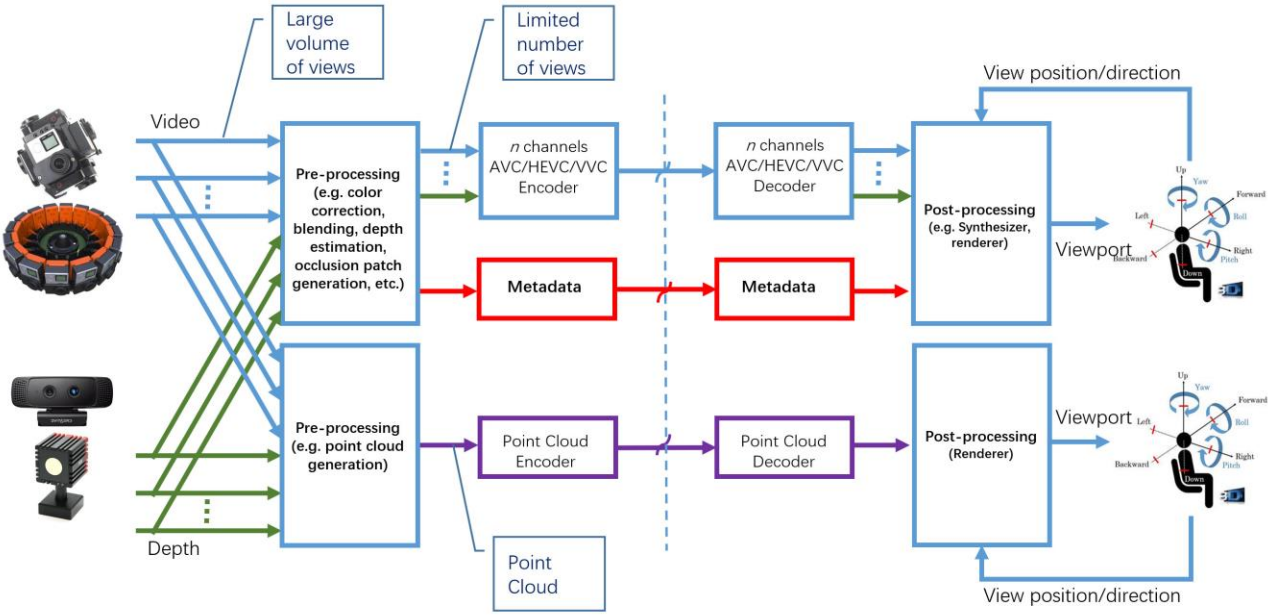

Figure 1 – Video & Graphics based Workflow of MPEG-I

The decoder (the right side of Figure 1) unpacks, decodes and extracts the data in Video- or Graphics-based data representation formats, and finally a renderer does an additional post-processing to obtain an animated image sequence that is displayed on the screen or Head Mounted Device (HMD). The quality of this image sequence in objective (PSNR, SSIM) or subjective (MOS) metrics, Yusra et al. (7), is the second (vertical) axis of the system's Quality-Bitrate performance figure, representing the average quality of the animated image sequence at a series of bitrate points. This is the most important performance figure that any coding standardization committee studies for continuous improvement.

In contrast to classical 2D video coding, the renderer does much more than placing the decoded data as pixels on the screen. For instance, in MPEG-I Video, the images decoded from the bitstream will be interpolated by a Reference Intermediate View Synthesizer (RIVS) to create any virtual view to the scene, hence providing the 3DoF+ or 6DoF immersive experience to the user. In MPEG-I Graphics, however, a point cloud is created from the decoded bitstream - i.e. a collection of coloured points in 3D space - which are projected on the screen through a typical OpenGL 3D graphics pipeline. Since the points are not connected and may possibly leave gaps, they are enlarged to disks with Splatting, Botsch et al. (8), through the rendering (post-processing) module of Figure 1.

The next subsections provide more details on the various modules of Figure 1 for the MPEG-I Video and MPEG-I Graphics processing pipelines, indicating their differences and commonalities.

MPEG-I Video Multiview + Depth Coding

In the MPEG-I Video pipeline, the various colour camera views are transmitted with mild pre-processing (e.g. distortion removal and colour correction) to the coder, and processed after decoding through the Reference Intermediate View Synthesizer (RIVS) at the

renderer side, for creating any virtual viewpoint in response to the user's spatial viewing position. Typically, RIVS requires a depth map per camera input for synthesizing any intermediate view with depth image-based rendering techniques, Sun et al. (9). Consequently, all camera feeds and their corresponding depth maps are transmitted through the network, as in the example of Figure 2 for the Technicolor Painter test sequence, which is one of the many Multiview + Depth video test sequences used in MPEG-I, Panahpour Tehrani et al. (10).

The creation of these depth maps in the pre-processing module is not part of the coding standard and is the sole responsibility of the content provider, who may use active depth sensing or passive depth estimation techniques (e.g. stereo matching). MPEG-I Video recommends to use its Depth Estimation Reference Software (DERS), Wegner (11), with a recent extension to Enhanced DERS (eDERS), Senoh et al. (12), for this purpose.

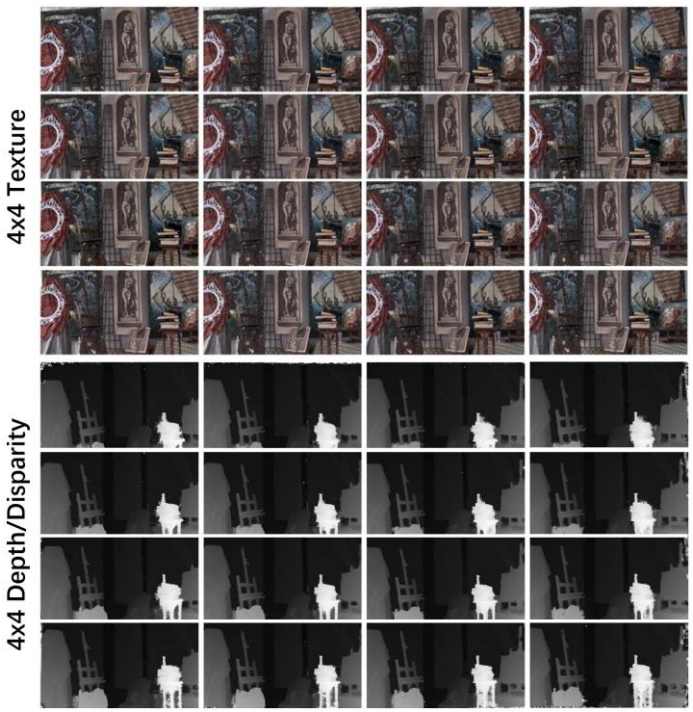

Figure 2 – Technicolor Painter Multiview capture (top) with estimated depth (bottom), Courtesy Technicolor.

The RIVS module - if not used at the encoder (cf. next paragraph as a counter example) – is strictly speaking also not part of the coding standard, though it has (similar to the depth estimation/sensing) a huge impact on the final rendering quality, and all benchmarking decisions. It has therefore been extensively studied over the past years, starting with View Synthesis Reference Software (VSRS), (11), that was originally developed for Horizontal Parallax Only (HPO) autostereoscopic displays with small disparity ranges (hence subject to improvement), its extensions to Enhanced VSRS (eVSRS), Senoh et al. (13), and recently more advanced implementations, Doré et al. (14) and Fachada et al. (15), that surpass VSRS and that – at the time of writing - are under consideration for replacing VSRS as new RIVS module.

Though the pre-processing and post-processing modules of Figure 1 are not part of the coding standard, they are considered in all MPEG-I experiments, since they impact the Quality-Bitrate performance figure of the coding system. Moreover, the redundancies between the Multiview images of Figure 2 might be exploited for better coding, using RIVS as a view prediction. Such View Synthesis Prediction (VSP) will predict a camera view using its adjacent camera inputs, and only the difference image (the residual) is actually coded and transmitted through the network.

Such VSP experiments for better coding have been conducted in the past, but did not lead to conclusive results yet, Baroncini et al. (16), probably because of the - at that time -

mediocre VSRS performances. The new RIVS performances might bring new hope to VSP.

Finally, instead of coding all captured views, packed occlusion patches (as explained in the next section) may be generated by pre-processing and coded as 'pre-processed video' together with supplemental metadata to indicate camera parameters, layout and packing of such video. More results will come out in the upcoming months through the MPEG-I activities.

Meanwhile, using unaltered MVD coding techniques based on HEVC, it is expected that 0.04 bits per refreshed pixel are needed (including the depth maps), Hinds et al. (17), bringing for a typical setup of 16 to 25 camera feeds in UHD (3840 x 2160 pixels), a total of 150-240 Mbps for 30 fps. In applications with Head Mounted Devices requiring much higher frame rates (at least 90-120 fps, i.e. 3 to 4-fold), the total bitrate will increase, but probably less than the corresponding frame rate ratio (expected to be a factor 2).

MPEG-I Graphics Point Cloud Coding

Since MPEG-I Graphics uses point clouds as data representation, the early coding activities of the 3DG group were oriented towards Octree- and kd-based coding used in the very first Lidar devices (6). The basic principle is that the points are grouped into a hierarchical structure of branches and leaves that allows for better difference/residual coding between a representative point and its direct neighbours in a group, cf. Figure 3. This method yields compression performances of one order of magnitude for static scenes, and it was very difficult to further extend its performances to the temporal axis with leaves that jump from one branch to another in the octree, even after a simple translation of an unaltered object in space.

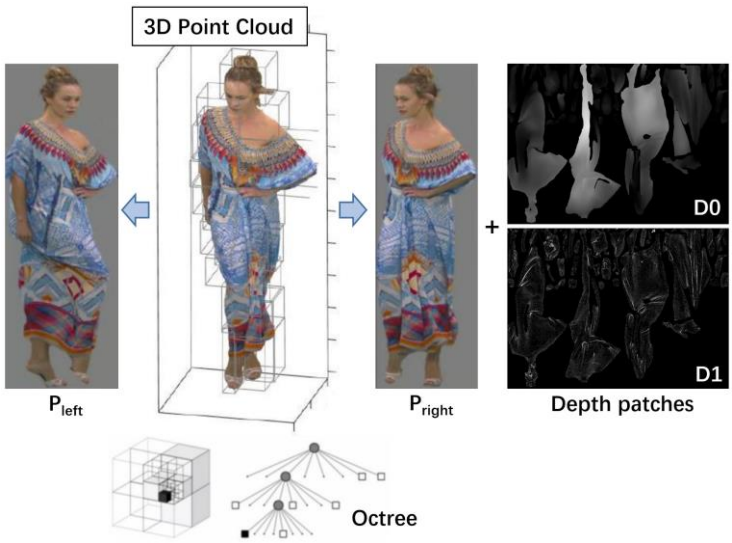

Figure 3 – 3D Point Cloud, its octree and projections (P_{left}, P_{right}, …) with Depth patches D0 and D1, Courtesy 8i.

Specifically, for dynamic point clouds, it was therefore proposed to find existing codecs that could already well exploit the temporal changes of the data, leading to the usual suspect: the video codec. The point cloud (typically for a single object) is segmented in subsets - called patches - and each patch is projected onto different planes in space with respect to its local orientation, cf. Figures 3 and 4, together with its depth maps (i.e. the distance from each point to the projection plane – called D0), and the so-obtained images are coded with already widely-accepted 2D video codecs (e.g. AVC or HEVC).

One may object that it makes little sense to start from a multi-camera acquisition providing images, out of which a point cloud is typically created by photogrammetry, Blizard (18), which in turn is projected back into a Multiview + Depth projection. However, be aware that in practice, the extraction of a point cloud of natural scenery from images (the pre-processing module in Figure 1) requires many different viewpoints to be acquired, typically in the order of hundred(s) of images, while – once a high-quality point cloud is extracted – a lower number of well-chosen projection directions (e.g. one order of magnitude less) may be sufficient to well-code the point cloud in its whole.

Nevertheless, note that in this point cloud projection process, there may be some occlusions that cannot be handled properly – e.g. when two points in space are projected on the same point in the projection plane, e.g. under the arms of the persons. For this case a second depth map (D1) was introduced which encodes the delta between the two points along with the projection axis, cf. Figure 3. One may observe that the 2D distribution of pixels in the patch image is not compression friendly, i.e. the 2D space is not uniformly occupied. To handle this situation, an occupancy map consisting in a binary mask of useful pixels is also encoded and transmitted.

This patch concept is actually extended over all regions of the object – similar to the texture UV mapping of 3D graphics objects, Aguiar (19) – even where there are no occlusions at all, leading to the typical structure of Figure 4(c), which corresponds to the meta-data in Figure 1. This has the advantage that traditional video codecs can be used, making MPEG PCC straightforward to be supported by a huge set of devices already available in the market.

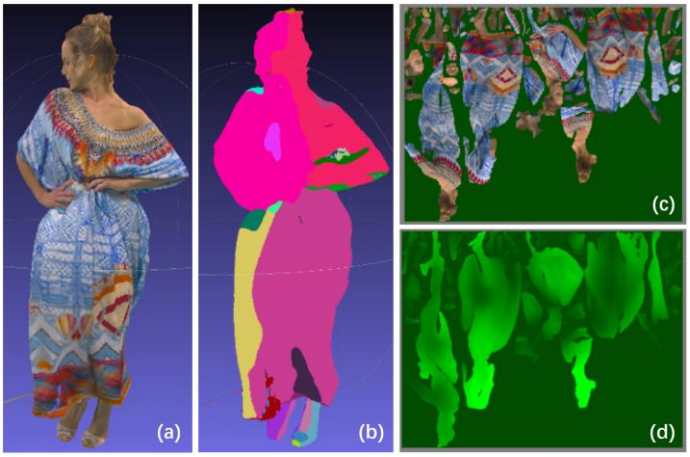

Figure 4 – (a) One projection of a point cloud, (b) its segmentation in (c) texture and (d) depth patches, Courtesy Apple Inc.

Experiments are still under consideration to best distribute the patches temporally so as to keep the highest coherence over time, and hence the best exploitation of redundancy in the codec for higher coding gains, Budagavi (20).

In summary, though PCC in MPEG-I Graphics is similar to MVD in MPEG-I Video, there are subtle differences around the use of patches and depth coding. Interestingly, MPEG-I Video is – at the time of writing – seriously considering also using patches in its MVD coding (e.g. to identify sensitive occlusion regions), bringing PCC and MVD even closer to each other in their technological concepts.

W.r.t. the coding performances in PCC, a bitrate of 10-20 Mbps at 30 fps has been observed – per object – on the extensive point cloud animation test set used in MPEG-I Graphics, Preda (21) and MPEG 3DG (22), rendered on a UHD display. It's important to indicate that these figures are obtained for single-object PCC coding, hence the total bitrate for scenes with multiple objects is increased accordingly with the number of objects.

For simple scenes with a dozen of objects, 120-240 Mbps at 30 fps is hence required, which is the same performance figure as reported with MVD coding in MPEG-I Video. As a result, not only do the coding approaches of MPEG-I Video and MPEG-I Graphics share a lot of similarities, they also share comparable coding performances.

FUTURE MPEG-I EXPERIMENTS

Since the first Working Draft issued after the Call for Proposal in July 2017, MPEG-I PCC continued to evolve by integrating new tools to increase the coding efficiency: the lossless mode is now supported by grouping the miss-projected points in a special patch, alternative approaches for encoding the occupancy map were proposed and time-consistent packing is investigated. While the activity is still ongoing, it is expected that 20% of coding gain will be obtained before publishing the Committee Draft (one of the last stages before final standardization) in October 2018.

MPEG-I Video has set up Common Test Conditions (CTC) in April 2018 for 3DoF+ and 6DoF Visual with test material, anchor definitions, objective and subjective evaluation methods. The group expects that evidence will be shown in the following MPEG meeting that a limited number of multiple 2D coded video streams together with metadata on camera parameters, layout and packing information of the video streams can provide the user with interactive experience of motion parallax in a 3D scene (3DoF+). In that case, a Call for Proposal on such metadata will be issued. Benefits of new coding tools for 6DoF Visual in the defined CTC are also expected from running exploration experiments. This may initialize formal standardization activities on compression of 6DoF content.

Convergence between MPEG-I Video and MPEG-I Graphics

The previous sections clearly suggest that MPEG-I Video and MPEG-I Graphics share a lot of technologies, with one noteworthy difference: while MPEG-I Video takes great care in the view synthesis (RIVS in the post-processing module of Figure 1), MPEG-I Graphics heavily relies on a proper point cloud extraction (the pre-processing module in Figure 1). This difference not being part of the standard itself, the boundaries between MVD coding in MPEG-I Video and PCC in MPEG-I Graphics clearly vanish.

In view of the apparent convergence between MVD coding in MPEG-I Video and PCC in MPEG-I Graphics, it is appealing to consider better comparing both technologies following a strict scientific approach that uses exactly the same test data and common test conditions. Unfortunately, the acquisition workflow (prior to and in the pre-processing module of Figure 1) for Multiview data and Point Cloud data being very different, it is not obvious to go from one representation format to the other for multiple-objects scenes. For example, the ULB Unicorn data set provided to MPEG-I in both representation formats, (10) and (22), has actually been acquired with very different camera acquisition positions: strictly planar for MVD, very much intrusive in the scene for PCC.

It is hence expected that industrial workflows (video- vs. graphics-based special effects) may have the biggest impact in choosing one or the other of these two coexisting representation and coding formats. Eventually, MPEG-I Video and MPEG-I Graphics may merge together to represent different objects from different workflows in a video-graphical 3D scene.

CONCLUSIONS

Two MPEG-I coding approaches – Video- and 3D Graphics-based – are eventually very similar in their technology and coding performances, reaching a couple of hundreds of Mbps at 30 fps for enabling immersive VR/AR applications in the 3DoF+ to 6DoF range. Calls for Proposals and Committee Drafts are the next upcoming milestones in the future MPEG-I standardization activities.

ACKNOWLEDGEMENTS

The authors would like to thank all the MPEG-I experts for their contributions to this work, as well as Innoviris, the Brussels Institute for Research and Innovation, Belgium, for supporting the work w.r.t multi-camera acquisitions, processing and rendering (contract number 2015-DS-39a, 3DLicorneA).

Figures 2 to 4 are altered reproductions with permission from the MPEG-I contributors, i.e. Technicolor, 8i and Apple Inc.

REFERENCES

1. MPEG-I, 2018. Coded Representation of Immersive Media. ISO/IEC 23090. https://mpeg.chiariglione.org/standards/mpeg-i 2018.

2. Karthikeyan, K.C., 2017. How The Matrix Bullet Time Works? https://geekswipe.net/art/films/how-matrix-bullet-time-works/ August 12, 2017.

3. Koenen, R., 2018. MPEG Standardization Roadmap. ISO/IEC JTC1/SC29/WG11. MPEG2018/N17506. April, 2018. San Diego, US.

4. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T., 2012. Overview of the High Efficiency Video Coding (HEVC) Standard. IEEE Transactions on Circuits and Systems for Video Technology. December, 2012. Volume 22, issue 12, pp. 1649 to 1668.

5. MPEG Press Release, 2018. Versatile Video Coding (VVC) project starts strongly in the Joint Video Experts Team. ISO/IEC JTC1/SC29/WG11. MPEG2018/N17482. April, 2018. San Diego, US.

6. Schnabel, R., Klein, R., 2006. Octree-based Point-Cloud Compression. Proceedings of Symposium on Point-Based Graphics, Eurographics. July, 2006. pp. 111 to 121.

7. Yusra, A., Al-Najjar, Y., Der Chen Soong, Dr., 2012. Comparison of Image Quality Assessment: PSNR, HVS, SSIM, UIQI, International Journal of Scientific & Engineering Research, August, 2012. Volume 3, Issue 8, pp. I041 to I045.

8. Botsch, M., Spernat, M., Kobbelt, L., 2004. Phong splatting. Proceedings of the First Eurographics Conference on Point-Based Graphics. 2004. Switzerland. pp. 25 to 32.

9. Sun, W., Xu, L., C. Au, O., Chui, S.H., Kwok, C.W., 2010. An overview of free viewpoint Depth-Image-Based Rendering (DIBR). Proceedings of the Second APSIPA Annual Summit and Conference. December, 2010. Singapore. pp. 1023 to 1030.

10. Panahpour Tehrani, M., Kroon, B., Nikitin, P., Senoh, T., Wegner, K., Lafruit, G., Tanimoto, M., Sun, Y., 2018. Overview of MPEG-I Visual Test Materials. ISO/IEC JTC1/SC29/WG11. MPEG2018/N17606, April, 2018. San Diego, US.

11. Wegner, K., 2018. List of tools for MPEG-I Visual Activities on 6DoF. ISO/IEC JTC1/SC29/WG11. MPEG2018/N17607. April, 2018. San Diego, US.

12. Senoh, T., Hara, K., Kawakita, M., Tetsutani, N., Yasuda, H., 2018. Updated eDERS to Higher Precision. ISO/IEC JTC1/SC29/WG11. MPEG2018/m42525, April, 2018. San Diego, US.

13. Senoh, T., Hara, K., Kawakita, M., Tetsutani, N., Yasuda, H., 2018. Enhanced VSRS to Four Reference Views. ISO/IEC JTC1/SC29/WG11. MPEG2018/m42526, April, 2018. San Diego, US.

14. Doré, R., Fleureau, J., Chupeau, B., Briand, G., 2018. 3DoF+ Intermediate View Synthetizer proposal. ISO/IEC JTC1/SC29/WG11. MPEG2018/m42486, April, 2018. San Diego, US.

15. Fachada, S., Bonatto, D., Schenkel, A., Lafruit, G., 2018. Depth Image-Based View Synthesis with Multiple Reference Views for Virtual Reality. 3DTV-CON. June, 2018.

16. Baroncini, V., Tanimoto, M., Stankiewicz, O., 2016. Results of the Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation. ISO/IEC JTC1/SC29/WG11. MPEG2016/N16128. February 2016, San Diego, US.

17. Hinds, A., Doyen, D., Carballeira, P., Lafruit, G., 2017. Toward the realization of six degrees-of-freedom with compressed light fields. IEEE International Conference on Multimedia and Expo. July, 2017.

18. Blizard, B., 2014. The Art of Photogrammetry: Introduction to Software and Hardware. http://www.tested.com/art/makers/460057-tested-dark-art-photogrammetry/ 2014.

19. Aguiar, G., 2017. Blender Beginner Tutorial - UV Mapping (Part 4). https://www.youtube.com/watch?v=Vj0k4-I33lQ September 17, 2017.

20. Budagavi, M., Llach, J., and Mammou, K., Clare, G., Litwic, L., 2018. PCC Core Experiments for Category 2. ISO/IEC JTC1/SC29/WG11. MPEG2018/N17346. January, 2018. Gwangju, Korea.

21. Preda, M., 2017. Report on PCC CfP answers. ISO/IEC JTC1/SC29/WG11. MPEG2017/N17251. October, 2017. Macao, China.

22. MPEG 3DG, 2017. Call for Proposals for Point Cloud Compression V2. ISO/IEC JTC1/SC29/WG11. MPEG2017/N16763. April, 2017. Hobart, Australia.