TOWARD THE REALIZATION OF SIX DEGREES-OF-FREEDOM WITH COMPRESSED LIGHT FIELDS

Arianne T. Hinds (CableLabs), Didier Doyen (Technicolor), Pablo Carballeira

ABSTRACT

360° video, supporting the ability to present views consistent with the rotation of the viewer's head along three axes (roll, pitch, yaw) is the current approach for creation of immersive video experiences. Nevertheless, a more fully natural, photo-realistic experience — with support of visual cues that facilitate coherent psycho-visual sensory fusion without the side-effect of cyber-sickness — is desired. 360° video applications that additionally enable the user to translate in x, y, and z directions are clearly a subsequent frontier to be realized toward the goal of sensory fusion without cyber-sickness. Such support of full Six Degrees-of-Freedom (6 DoF) for next generation immersive video is a natural application for light fields. However, a significant obstacle to the adoption of light field technologies is the large data necessary to ensure that the light rays corresponding to the viewer's position relative to 6-DoF are properly delivered, either from captured light information or synthesized from available views. Experiments to improve known methods for view synthesis and depth estimation are therefore a fundamental next step to establish a reference framework within which compression technologies can be evaluated. This paper describes a testbed and experiments to enable smooth and artefact-free view transitions that can later be used in a framework to study how best to compress the data.

Index Terms— Light Field, View Synthesis, Depth Estimation, 6 DoF, Virtual Reality

1. INTRODUCTION

Over the last decades, generations of block-based, standards for video codecs have been developed with the goal to halve the bandwidth required for a given picture quality at a given pixel rate. The current state-of-the-art video codec, HEVC [1] is capable of delivering each refreshed pixel roughly at 0.04 bits [2]. For UHD resolutions refreshed at 50 fps, this amounts to 16 Mbps at a picture quality that is satisfactory for current UHD displays. However, the view that is presented in these displays is constrained to the single view that is captured typically by a camera with a single lens and corresponding image sensor. Multiview displays, on the other hand, feature the ability to select from a number of views of the captured scene.

1.1. Light field acquisition

For the corresponding applications, views are captured by an array of discrete cameras or a single camera with an array of lenses (i.e. lenslet-based cameras). The use of multiple lenses provides information regarding the direction from which rays of light are being reflected from the scene. This directional information combined with colour values recorded on the image sensors comprise the so-called light field.

Arrays of discrete cameras offer the advantage, over lenslet-based cameras, of supporting a variety of configurations (uniformly vs. non-uniformly spaced) as well as larger baselines. The size of the baseline in addition to the density of the array, the distance from the array to the scene, and the field of view captured by each lens, determines the amount of disparity that exists between consecutive images captured by the system. The magnitude of disparity values correlates directly to the smoothness in transitions from any given image to an adjacent view in the system. Smaller vs. larger disparity values generally create smoother transitions from image to image, and result in fewer artefacts for a higher quality of experience in the application.

1.2. Immersive displays

Also contributing to the Quality of Experience (QoE) is the resolution and refresh rate of the corresponding display. HMDs are rapidly evolving to provide higher refresh rates (measured in the total number of pixels per second to refresh), and corresponding bitrates (see Figure 2) so that they can deliver truly immersive VR experiences with visual acuity and resolutions that do not create eye strain. Moreover, the display should deliver views that are consistent with the motion of the viewer at a motion-to-photon latency that prevents motion sickness [3].

1.3. Degrees-of-Freedom (DoF)

The latter requirement of delivering views that are consistent with the motion of the user refers to the Degrees-of-Freedom (DoF) that the display and content support.

Current generations of 360° video and HMDs provide support for 3 DoF, i.e. the 3 rotations of the head (roll, pitch, yaw) but no translation in the x,y,z space. The resulting experience may provide a realistic impression but with a poor

feeling of motion parallax. With VR, the goal is to render 6 DoF, as shown in Figure 1, to get a perfect feeling of immersion. This means providing coherent content with any head movement and preferably at a refresh rate that introduces no noticeable lag.

In order to provide immersive 6 DoF support, visual depth cues, such as stereopsis, binocular occlusions, vergence and motion parallax are required for natural content. Note that these cues are easier to create for synthetic content. A light-field-based solution addresses these challenges by having the ability to render any view from any point, facing any direction, with any field of view, inside a captured-viewing volume, offering full motion parallax and correct view-dependent lighting.

Fig. 1. 6 Degrees-of-Freedom including roll, pitch, yaw of the viewer's head, and x,y,z translation

1.4. Need for more compression

A major challenge to the light field information that describes such a captured viewing volume is the enormous data transmission requirement, which is more than 100 times bigger than conventional 2D video. For instance, Lytro Immerse [4] employs hundreds of high-resolution cameras, delivering more than 100 GB/s of raw video data to provide a comfortable viewing volume for 6-DoF light field VR. Even though only a fraction of this data is rendered on the user's HMD for the current head position, the underpinning light field data stream still represents 1GB/s of uncompressed data for current HMDs (Oculus Rift [5], HTC Vive [6] and PSVR [7]) supporting viewing resolutions around 1 Mpixels per eye at 90 fps. As future HMDs will rapidly grow in resolution to/beyond 8k pixels per eye at 240 fps, one may expect another 100-fold data streaming requirement in bandwidth.

Figure 2 shows with red circles the available interconnectivity bandwidth predicted from Nielsen-Edholm's law over several decades (units along the left vertical axis), under different technologies [8-9]. The blue squares show the expected growth in the number of pixels to be refreshed per second for Multiview displays (units along the right vertical axis) over the coming decades [10]. Future Light Field displays (cf. line (1)) and VR-HMDs (cf. line (2)) have even more demanding pixel refresh rates, corresponding to the upper line in Figure 2.

For a given bitrate budget per pixel – here HEVC's 0.04

bits per pixel, corresponding to 1.4 dB difference in $\log_{10}$ scale – the slopes of these graphs can be compared; the upper lines

Fig. 2. Pixel refresh rate growth (squares – right axis) vs. available bandwidth growth (circles – left axis) in Multiview and Light Field VR

(1) and (2) in Figure 2 yielding the required bitrate, the lower lines (3) the available bandwidth. The combined graphs show that current Multiview applications already require advanced transmission technologies, for reaching bitrates of up to two orders of magnitude higher than broadband internet.

Future 6-DoF Light Field VR displays are much more demanding in their pixel refresh rate, cf. line (2) in Figure 1. Consequently, to overcome bottlenecks, more advanced compression technologies should be developed, reaching at a minimum 10-fold improvement (one unity in $\log_{10}$ scale) over current HEVC-based codecs. Obviously, this is a very challenging target that should be addressed with a different approach than the one used so far in 2D video compression, i.e. fine-tuning motion estimation and entropy coding.

2. A TWO-PART PROBLEM OF COMPRESSION

Many efforts have been invested in improving the compression performance of standards-based video codecs for Multiview video. For example, as a rule of thumb, a 30% bitrate savings can be expected, on average, when exploiting redundancies between all camera views in MV-HEVC over simulcast HEVC [11][12]. In 3D-HEVC, depth maps are added in the transmission chain to support subsequent View Synthesis (VS) functionalities at the decoder. The addition of depth maps, however, increases the bitrate back up to 30% [2], yielding comparable bitrate figures as those for simulcast HEVC. This justifies the use of HEVC performance figures of 0.04 bits per pixel, in particular in Figure 2, as a reference point in all further evaluations.

2.1. Part 1 - Capture the minimal number of views

If, instead of fine-tuning existing video codecs, only half of the camera views are transmitted and the other half are synthesized at the receiver, then a nominal 50% bitrate savings would already be obtained prior to compressing the remaining half of the views. These gains are higher than the aforementioned 30% bitrate gains in HEVC-based. Such a

scenario is illustrated in Figure 3 where for stereoscopic viewing with an inter-ocular distance of 6 cm, mid-point views are synthesized from views that are separated by 12 cm: views from the heavily-shaded cameras are kept, while views from the lightly-shaded cameras are skipped during transmission (or not captured at all), i.e. their corresponding images should be synthesized at the decoder.

Fig. 3. With an inter-ocular distance of 6 cm, stereoscopic viewing with 12 cm inter-camera distance (bright yellow cameras) requires synthesizing mid-views; in (a) the odd images are synthesized from the surrounding even camera views; in (b) even more images are synthesized from a small subset of the input camera views

This simple thought experiment suggests that sending a fraction M of all available N views ($M < N$) may lead to substantial bitrate savings, as long as the quality of the visual rendering at the receiver is not heavily compromised. Given that compression performance is expressed as a function of quality (objective or subjective) vs. bitrate, quality losses both in the compression of transmitted views, and the synthesis of views that are not transmitted, should be avoided.

VS, however, is a process that is closely tied to the specific display device for which the view is being rendered. VS algorithms need to perform under different conditions such as available computational resources or other factors influenced by the interaction of the viewer with the device. VS also serves as a product differentiator for display manufacturers. All of these points suggest that it is not practical to converge on a single VS algorithm in the development of a compression standard for light fields. That is, as with motion estimation algorithms employed in existing standards-based video encoder design, VS should be regarded as a type of “black box” for which both its inputs and outputs are parameterized in the syntax of the standard. Thus, VS should be a *non-normative* part of a compression standard.

The dilemma, however, is that VS can potentially offer substantial savings in the compressed bitrate if implemented well, (cf. Figure 3). Likewise, if poorly implemented, and/or insufficiently parameterized by the standard, it can degrade the resulting image quality. Furthermore, if VS is not well understood prior to evaluating the compression technology (for the transmitted views), then the problem of identifying which image quality artefacts are from VS, as opposed to those artefacts that are introduced by the compression technology, becomes intractable.

2.2. Part 2 – Compress the captured views

Views that are selected for transmission must be compressed. However, in evaluating the performance of candidate compression technologies in the development of the coding standard, it is important to be able to isolate the problem of compression from the VS process, i.e. the VS process should be predictable and well understood so that any artefacts that are observed in the final rendered output can be attributed to the compression algorithm being evaluated and not the VS process. This implies that the evaluation framework must provide a reference VS implementation whose performance and characteristics are understood well enough to serve as a “minimum quality” VS algorithm.

For example, Domański et al. [13] have studied view synthesis along the path connecting physical cameras, using sparse camera arrangements. Their experiments illustrate the dependencies between the inter-camera distance (baseline), depth resolution and the size of the occluded regions, all of which impact the performance of VS. They show that the presence of occluded areas over adjacent camera views negatively impacts the depth estimation and VS results. Consequently, a non-uniform distribution of cameras clustered pairwise over the scene is recommended for achieving better VS results – referred to as the “PSNR pairing gain” – as opposed to a strictly uniform camera distribution.

Such a non-uniform camera arrangement has the additional advantage of recovering higher compression-oriented redundancies within subsets of cameras (each pair of cameras) than the inter-camera redundancies that are only marginally present in a uniform, sparse camera arrangement.

An evaluation of the compression performances can hence be conducted by addressing the two axes in the Quality-Bitrate performance figures as much as possible independently. For example:

- Estimate the view synthesis quality without compression for different camera arrangements
- Estimate the bitrate budget starting from the simulcast scenario, further improved by studying redundancies over subsets of cameras in the camera arrangement
- Evaluate the impact of the compression artefacts on the view synthesis quality, updating (a) in reaching higher levels of confidence

3. VIEW SYNTHESIS AND 6 DOF

The level of immersion that a visualization system can provide is related to the DoF that the user is able to experience. A translation (or rotation) of the user’s head in one direction (one DoF) is made available if the rendering system is able to provide content that is coherent with the user’s movement in that direction. The coherency includes the new viewpoint position as well as parallax information. Allowing one degree of freedom means that the new viewpoint can be anywhere in a particular direction and not necessarily physically captured by any of the cameras. A VS

process is applied to render any viewpoint using available captured/transmitted views and associated depth information. Furthermore, the disocclusion information should be added; i.e. objects that are located behind a foreground object and made visible by this translation action should also be rendered.

SMV systems (e.g. the Holografika Holovizio [14]) that provide only one DoF in the x direction have been demonstrated. These render only a discrete number of synthesized viewpoints. However, for 6 DoF, the 3 translations in the x,y,z directions are added to the 3 rotations of a 3 DoF system. The complexity of VS in the y direction may not be fundamentally different from the VS in the x direction. Nevertheless, the disparity transmitted should be used taking into account the actual arrangement of the cameras after rectification. If cameras are not placed on a regular grid in both the x and y position, the disparity information should also be adapted.

The third axis, i.e. the z direction, is more complex to handle. It is possible to render an intermediate view in between two views in the x or y direction without any knowledge of real sizes. The synthesized view is still coherent in size with the real view. But performing a step-in/step-out in the scene requires more information about the real distance. For example, the immersion may be destroyed if when approaching an object, its modified size is not coherent. The smoothness of the variation is one important parameter, but the relative size of objects perceived is another one.

4. VIEW SYNTHESIS PARAMETERIZATION

Given these challenging requirements to allow viewpoint translation in the x, y, z dimensions, a set of experiments to find a parametrization of the minimum requirements for the physical-view density that can lead to an acceptable VS quality, is needed.

Previous works [15] [16] have analyzed the QoE in SMV video applications with the objective of identifying new aspects that are relevant in the subjective perception of SMV and that do not apply to the evaluation of 2D or fixed-viewpoint stereoscopic video, namely the perception of viewpoint transition.

These works endeavor to provide a parameterization that describes the relationships between camera arrangement, content and the user experience in this view transition. Such a parametrization is a valuable tool to guide the configuration of scene parameters such as depth or density of cameras for an acceptable viewing experience.

Particularly, the Multiview Perceptual Disparity Model (MVPDM) [16] proposes a parametrization of a given content captured by a multi-camera arrangement based on the disparity between adjacent views, instead of based on the angle or camera distance, and thus:

- It aggregates the contribution of different parameters (focal length, baseline, depth,...) that influence the

subjective experience in viewpoint transition, better representing the perception of visual comfort

- It is common to different camera arrangements, such as linear, linear convergent or arc.

Fig. 4. Disparity (d) histogram weighted by $\omega(d)$ used to compute the d_{perc} in the MVPDM. $\omega(d)$ emphasizes the saliency of objects in the scene that are closer to the viewer.

The parameters of the MVPDM are based on a perceptual disparity d_{perc} that is computed by means of a disparity histogram (cf Figure 4) that is weighted by a perceptual weight $\omega(d)$, using:

$$d_{\text{perc}} = \int_{d_{\min}}^{d_{\max}} \text{hist}(\delta) \omega(d) d\delta. \quad (1)$$

d_{perc} aggregates the disparity distribution of a scene onto a representative value that correlates well with the disparity perception of the viewer.

The MVPDM has been assessed using 1D (horizontal) SMV content, which presents similar view density characteristics to light field content in the horizontal dimension. Thus, the MVPDM will be used as a starting point to define a parametrization of the physical view density of light field content, that ensures acceptable quality in synthesized virtual views in the x, y, z dimensions. The MVPDM has already been used to define the minimum comfortable camera density in a view path for FN scenarios [17], setting the number of intermediate virtual viewpoint positions between physical cameras for a comfortable view transition.

One of the objectives of the framework described in Section 5 is to study the adaptation of the MVPDM to the vertical and depth dimensions. We expect that the parametrization that results from that framework will have a high degree of similarity in the x and y dimensions. However, additional perceptual effects are foreseen in the view translation in the depth dimension, which we will need to further investigate to define the VS requirements and parameterization for the depth dimension.

5. EXPERIMENTAL FRAMEWORK

Currently, the compression problem is a multivariate problem consisting of many ways to 1) synthesize views, 2) capture content and 3) estimate depth. All of these problems make it difficult to separate how the compression impacts the overall quality of the system, i.e. the view synthesis and depth estimation also impact overall visual quality.

5.1. Part 1 – 1D camera arrangement

Fig.5: View synthesis quality degradation with increasing baselines (increasing number of skipped camera input views)

Sparse non-uniform camera arrangements (cf. Section 2) provide opportunities to substantially reduce the bitrate, while maintaining a sufficiently high VS quality. Moreover, as outlined in Section 2.2, both aspects can be studied independently, up to a certain level of confidence.

Though Domański et al. [13] have addressed the “PSNR pairing gain” (cf. Section 2.2) obtained in VS with pairs of cameras, the challenge about the admissible baseline between successive camera pairs remains open.

The MPEG standardization committee has therefore explored the impact of increasing baselines on the VS quality, by gradually skipping existing camera input views and synthesizing the corresponding virtual views. Figure 5 shows a typical variation of the VS quality from the current View Synthesis Reference Software (VSRS) [18] applied on a single test sequence [19]. The valleys correspond to quality degradations under various view skip settings.

Unfortunately, the camera number shown in Figure 5 along the horizontal axis of the graph is not a distinctive parameter that permits a fair comparison between different test sequences, taken under various camera and scene settings. We therefore propose to use a very dense 2D light field data set and conduct the same experiments, using the MVPDM metric proposed in Section 4 on the horizontal axis, rather than the camera number.

Figure 6 depicts an example of a representation of the PSNR values of Figure 5 plotted against the main parameter of the MVPDM. The x-axis values represent the value of $\log_{10}(\text{dperc})$ between a virtual view and the closest of the two reference views used to synthesize that particular virtual view. It can be noted that several PSNR values share the same value of dperc, e.g. all of the single cameras between two original views in the 33 view-case (black points in Figure 5). For each camera configuration and sequence, the graph shows that a linear relationship between the average VS quality and dperc (in a log scale). Thus, the VS quality for a

given reference camera setting could be parametrized by such lines in the dperc domain.

The data set [20] is acquired by taking high-resolution snapshots of several static scenes with successive, tiny translations of a camera over a 2D surface, which is itself perpendicular to a point of interest in the scene. VSRS (or any candidate VS) is then used to generate the virtual (skipped) views. Each test sequence yields curves as shown in Figure 6, and, superimposing these curves for all test sequences indicate the robustness of a VS algorithm over different baselines and scene content settings, expressed in the MVPDM metric. This gives insights into the minimal camera density required for achieving a predefined quality level, and the corresponding reduction in total number of input pixels, addressing a large part of the compression challenge.

5.2. Part 2 – 2D camera arrangement

Experiments can also be performed for a 2D camera arrangement to address 6 DoF using the same data set described in Section 5.1. That is, the analysis of the minimal density of cameras in the x direction can be extended to the y direction. Although numerical values may differ, the method to conduct the experiment is the same. A second step in the framework is then to evaluate the minimum density of cameras for a combination of both x and y displacement (but keeping the z position constant). The MVPDM metric should then be adapted for this purpose. The data set is still able to provide a ground truth for any integer combination (x,y) of displacement. The final camera density should then be expressed as a two-dimensional value with similar values in x and y (or not). The density may also be expressed as a radial distance to a center position.

A third step is to consider the z direction, i.e. the step-in/step-out effect. In this case, the 2D static camera data set no longer provides the ground truth, and therefore a more complex scene acquisition process is needed. For example, at a given (x,y) position, a large number of views should be acquired with different fields of view to get a discrete sampling in the z direction. The VS software may then be adapted to compensate for field of view used in the acquisition process (to ensure a real step-in/step-out effect). The camera density in both x and y directions must then be evaluated again to ensure no unwanted effects resulting from the changes to VS to support the step in the z direction. Likewise, the MVPDM metric will also have to be adapted.

6. CONCLUSION

In order to provide a fully natural, photo-realistic experience, immersive VR applications not only must deliver 6 DoF to the viewer, but must also support visual depth cues, such as stereopsis, binocular occlusions, vergence and motion parallax. Light-field-based solutions can address these challenges, however, the data required to do so is

Fig. 6 MVPDM-PSNR graph for the results in Figure 5, and other sequences. PSNR values are plotted against d_{perc} ($\log_{10}$ scale) instead of camera numbers.

significantly more than the data required for 2D natural video. Toward the development of a standards-based compression technology for light fields, view synthesis will play an increasingly important role, and therefore understanding its performance and characteristics is crucial. We suggest a test framework and set of experiments that utilize a set of dense camera data sets, and the MVPDM metric as a basis to parameterize a minimum quality virtual synthesis process. This minimum quality process can facilitate the evaluation of light field compression technologies so that visual quality artefacts resulting from these technologies can be isolated from artefacts that may arise from the view synthesis process.

7. REFERENCES

- [1] ISO/IEC 23008-2:2015 Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding.
- [2] Stankiewicz, Olgierd, Lafruit, Gauthier, Domanski, Marek, "Multiview Video: Acquisition, Processing, Compression and Virtual View Rendering," Academic Press Library in Signal Processing, Editors Chellappa R., Theodoridis S., to be published in Q1 2017
- [3] Kolasinski, E. (1995). Simulator sickness in virtual environments. Technical report, U.S. Army Research Institute for the Behavioral and Social Sciences.
- [4] Lytro Immerse <https://www.lytro.com/immerge>
- [5] Oculus Rift <https://www.oculus.com>
- [6] HTC Vive <https://www.vive.com/us/>
- [7] Playstation VR <https://www.playstation.com/en-us/explore/playstation-vr/>
- [8] Nielsen's law of internet bandwidth, <https://www.nngroup.com/articles/law-of-bandwidth/>
- [9] Bandwidth growth: nearly what one would expect from Moore's law, <https://ipcarrier.blogspot.be/2014/02/bandwidth-growth-nearly-what-one-would.html>
- [10] Future trends of Holographic 3D display, http://www-oid.ip.titech.ac.jp/holo/holography_when.html
- [11] Y.S. Ho, K-J Oh, "Overview of Multi View Coding", *2007 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Service*, pp. 5-12
- [12] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 26, Issue 1, pp. 35-49, Sept. 2015
- [13] M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, "Optimization of camera positions for free-navigation applications," Conference: 2016 International Conference on Signals and Electronic Systems (ICSES), DOI: 10.1109/ICSES.2016.7593833, September 2016
- [14] Holografika Holovizio <http://www.holografika.com>
- [15] P. Carballeira, J. Gutiérrez, F. Morán, J. Cabrera, N. García, "Subjective Evaluation of Super Multiview Video in Consumer 3D Displays", *Seventh International Workshop on Quality of Multimedia Experience, QoMEX 2015*, C. Navarino, Greece, pp. 1-6, May 2015.
- [16] P. Carballeira, J. Gutiérrez, F. Morán, J. Cabrera, F. Jaureguizar, N. García, "Multiview Perceptual Disparity Model for Super Multiview Video", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 113-124, Feb. 2017
- [17] ISO/IEC JTC1/SC29 WG11, "Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation," MPEG 113th meeting, contribution M37296, Geneva, Switzerland, October 2015.
- [18] MPEG SVN repository: <http://wg11.sc29.org/svn/repos/Explorations/FTV/VSR/S/trunk/>
- [19] ISO/IEC JTC1/SC29 WG11, "EE1 and EE2: Bee results" 110th meeting contribution, M35079, Strasbourg, FR, October 2014.
- [20] ISO/IEC JTC1/SC29 WG11, "ULB High Density 2D camera array data set version 1", M40293, Hobart, AU, April 2017.