A STUDY OF LIGHT FIELD STREAMING FOR AN INTERACTIVE REFOCUSING APPLICATION

Martin Alain, Cagri Ozcinar, Aljosa Smolic

V-SENSE Project, School of Computer Science and Statistics, Trinity College, Dublin

ABSTRACT

Light fields are able to capture light rays from a scene arriving at different angles, which allows post-capture rendering applications such as interactive viewpoint selection or refocusing. However, this additional angular information comes at the price of a significant increase of the data volume compared to traditional 2D images. While light field compression is still an ongoing research effort, showing impressive compression gain with the latest coding standard, light fields are in practice often stored on remote servers to avoid consuming unnecessary storage of the user devices. A typical cost-effective solution for light field visualisation is then to render the requested image on the server and transmit the result to the user. Another trivial solution would be to directly send the light field to the user and perform the rendering process directly on the client side to avoid transmission delay. While the latter solution seems instinctively less optimal and is usually discarded in previous work because of an expected unacceptable startup delay, we propose a quantitative study to compare both solutions in terms of rate-distortion (RD) performance. A counterintuitive finding of this paper is that accepting a reasonable startup delay (a few seconds) can provide a significant improvement of the RD performances.

Index Terms— Light Fields, Streaming, Refocusing

1. INTRODUCTION

Light fields emerged as a new imaging modality, enabling to capture all light rays passing through a given amount of the 3D space [1]. Compared to traditional 2D imaging systems which only capture the spatial intensity of light rays, the common two-plane parameterisation of light field also contains the angular direction of the rays. A light field can be represented as a 4D function: $\Omega \times \Pi \rightarrow \mathbb{R}, (s,t,u,v) \rightarrow L(s,t,u,v)$ in which the plane Ω represents the spatial distribution of light rays, indexed by (u,v), while Π corresponds to their angular distribution, indexed by (s,t). A practical way to visualise a light field is to consider it as a matrix of $M \times N$ views, also called sub-aperture images (SAI), where each image represents a 2D slice of the light field over the spatial dimensions (u,v). Applications of light fields notably include rendering novel images, either corresponding to new viewpoints [1], and/or with new focus distance and depth-of-field [2–4].

Light field applications need a lot of storage space for accommodating the sheer size of the data representation, requiring an investigation of light field streaming strategies. Light fields, for instance, can be stored on remote servers, such as cloud servers, and only a requested viewpoint can be rendered on the server side, and transmitted to the client instead of having to deliver the whole light field. In [5], light field streaming is explored with a focus on free viewpoint rendering application. As the rendering method used in this research work only relies on a few SAIs to render a different viewpoint [1], a rate-distortion (RD) criterion was proposed to optimise packet scheduling for the transmission of these SAIs. The rendering is then performed on the user side.

However, refocusing can typically depend on all the light field SAIs, which prevents the direct application of the method described above. Thus, interactive light field streaming with a refocusing application was later studied in [6–8], where only a subset of the SAIs is first transmitted based on the assumption that transmitting the full light field would cause a substantial startup delay. New refocused images are then either estimated as a sparse linear combination of the subset of SAIs available, or generated on the server side and then transmitted to the user, together with a new SAI to grow the subset of SAIs available. This progressive framework can reduce the accumulated rate compared to a system where every refocused image is rendered on the server and transmitted to the client. However, since performance is only evaluated based on the accumulated rate, improvement only occurs when a sufficient number of images have been sent, and the practical gain of such a method is difficult to predict.

In this paper, we focus on light field streaming with a refocusing application, similar to the work of [6]. While it was assumed in the previous work that a trivial direct transmission of a full light field would incur an unacceptable startup delay, in this paper, we provide a comparison between this solution, denoted as scenario A, and an interactive streaming solution, denoted as scenario B. In the interactive streaming solution, each requested image is rendered on the server and then transmitted to the user. Two main reasons motivate this study. First, while scenario A is commonly overlooked in most research works, in recent literature, we did not find any quantitative study based on the last video coding standard, high efficient video coding (HEVC) [9], which provides high efficient performance for light field compression by exploiting the pixel correlation between the SAIs [10, 11]. Second, while this was not explicitly taken into account in previous work on refocusing from a compressed light field, refocusing methods are known to have a de-noising effect [4, 12], which can reduce the compression noise. In addition, recent works such as [6–8] focused on dense light fields captured

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776.

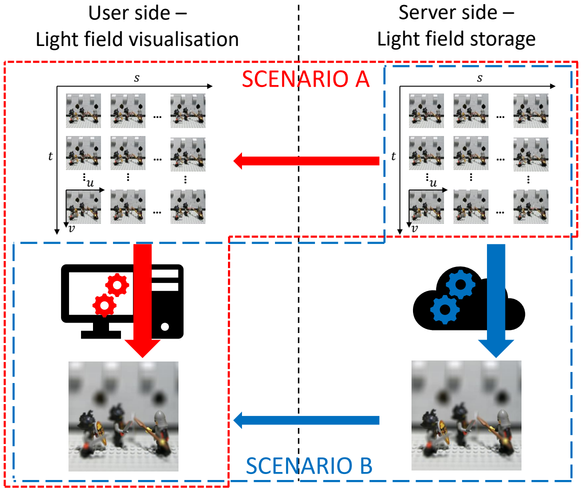

Fig. 1: Scenario A: the light field is streamed to the user side for rendering before visualisation. Scenario B: the light field is rendered in the cloud and the result is streamed to the user side for visualisation.

with a plenoptic Lytro Illum camera [3], while we propose here to study the streaming scenarios for different types of light field datasets, such as the Stanford [13] and Technicolor [14], captured with a gantry and a camera array respectively.

The remainder of the paper is organised as follows. Section 2 describes the technical details of our experimental setup. Section 3 presents the experimental results and analysis. Finally, the paper is concluded with directions for future research in Section 4.

2. EXPERIMENTAL SETUP

We consider two different streaming scenarios in the context of an interactive light field refocusing application, as illustrated in Fig. 1. The user can request any focus distance independently from the ones previously requested, *e.g.,* as in a point-and-click refocusing application. We also do not make any assumption about the number of images that the user will request.

In the first scenario, denoted as scenario A, all the light field SAIs are sent to the user, and the rendering of newly refocused images can then be performed on the user side. However, this scenario will incur a delay, d_A, before the user can start visualising any refocused image, but a negligible delay is then occurring during visualisation. While it has been assumed in previous work that this delay would not be tolerable in the sense of user experience, our goal is to quantify this delay and study the RD performance of this scenario when the startup delay is considered acceptable. In this scenario, we did not consider the transmission of all possible rendered focal stack images as it usually contains more images than the original light field. In addition the images from the focal stack exhibit less pixel correlation than the original SAIs, and it would altogether require a higher transmission bitrate.

In the second scenario, denoted as scenario B, we consider an interactive streaming solution, where the requested refocused image is rendered on the server before being sent to the user. In this scenario, the user does not have to wait before starting visualisation. However, the transmission delay d_B has to be taken into account. Thus the constraint on this delay is quite strong to ascertain interactivity.

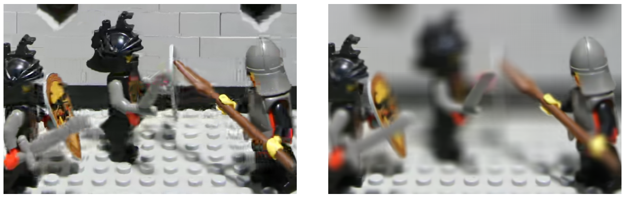

Fig. 2: The de-noising effect of refocusing: a crop of the center image of a compressed light field (left) and the corresponding refocused image (right) (Best viewed zoomed).

For compression in scenario A, all SAIs are encoded in a way to exploit the natural light field redundancies, and the whole light field is transmitted at once. As light field compression is still an ongoing research topic and no standard is yet available, we used the latest MPEG video compression standard, HEVC [9]. The input video is created by temporally stacking the SAIs following a snake scanning order, which is one of the baseline solutions used in light field compression research [15, 16].

In scenario B, each refocused image is encoded separately using the intra prediction of HEVC. As we can not predict the order in which the user is going to request the next image, in this scenario, we can not rely on an inter-image prediction based compression.

For refocusing, we use the well-known shift-and-sum algorithm [3], where the refocused image is obtained as linear combination of shifted SAIs:

$$I_r(u,v) = \sum_{s,t} A(s,t)L(s,t,u+\delta*(s-s_r),v+\delta*(t-t_r)), \quad (1)$$

where (s_r, t_r) are the indices of the reference SAI to be refocused, A is a filter that defines the synthetic aperture, and δ is a disparity value which controls the focus distance. The size and shape of the synthetic aperture notably controls the depth-of-field (DoF) and nature of the bokeh. In this paper, we consider a simple box filter for the synthetic aperture, defined as:

$$A(s,t) = \begin{cases} 1, & \text{if } s \le s_A, t \le t_A \\ 0, & \text{otherwise} \end{cases}$$

where $s_A \times t_A$ is the size of the synthetic aperture, which is studied in section 3.2.

From equation 1, it is clear that the refocused image we obtain is the result of a filtering operation along the angular dimension. It is known that the light field refocusing operation has de-noising properties [4, 12], thus refocusing applied on compressed SAIs will help reducing the distortion due to compression artefacts. Such de-noising effect is illustrated in Fig. 2, and analysed more in depth in section 3.2.

3. RESULTS

To evaluate the performances of the proposed streaming scenarios, we considered three light field datasets with different spatio-angular characteristics. First, we used a dataset of six light fields captured with a Lytro Illum, drawn for the EPFL [17], INRIA [18], and VSENSE [19] datasets. The SAIs were decoded with the modified Matlab Light Fields toolbox [20] proposed in [19] to correct colour issues. Note that the last de-noising step of this pipeline was not used here. The light fields consist in 15×15 SAIs of resolution 625×434. Light fields captured with Lytro Illum have a very dense angular sampling, and thus a very high pixel correlation between close SAIs. Second, we used six light fields from the Stanford dataset [13], captured with a Gantry. These light fields consist of 17×17 SAIs of resolutions 1400×800, 1280×1536, 1536×1152, and 1024×1024. The spatial resolution is higher than for the Lytro Illum ones, but the angular sampling is not that dense, thus close SAIs exhibit less pixel correlation. Finally, we used five light fields from the Technicolor dataset [14], captured with a camera array. These light fields consist of 4×4 SAIs of resolution 2048×1088 and have the highest spatial resolution, but a very sparse angular sampling, thus exhibiting the lowest pixel correlation between close SAIs.

To encode each content for each streaming scenario, we used *libx265* in the FFmpeg software (*ver.* N-92755-g3f08ed3) [21]. In scenario A, we used an inter-predictive coding condition of HEVC, with the default settings of FFmpeg in which group of picture size was set to 12. In scenario B, we independently encoded each rendered focal stack image using the constant rate factor and intra only settings of FFmpeg.

The acceptable latency d_a^B for scenario B is fixed to 100ms to allow for interactivity [22,23]. For scenario A, we study different values of acceptable latency d_a^A: 1s, 2s, 3s, and 5s.

To evaluate the RD performances, the RGB PSNR is computed on refocused images, using refocused images rendered from the source light field as a reference. For scenario A, the bitrate is computed as the size of the compressed light field SAIs divided by the acceptable delay. For scenario B, the bitrate of a single refocused image is evaluated as the size of this compressed image divided by the acceptable delay. For multiple refocused images, the individual bitrates are averaged. This bitrate evaluation was preferred over the strict bitrate accumulation, as it is independent of the number of images transmitted, which we assume cannot be known beforehand. Bitrate and PSNR values are averaged over all the refocused images of the focal stacks.

3.1. Comparison of scenarios A and B

We show in Figs. 4 and 5.e RD curves for the Bikes, Birthday, and LegoKnights light fields, as they are representative of their corresponding datasets. Detailed results for all test light fields are available online[1]. Note that in this section we use the full synthetic aperture to render the focal stacks, *i.e.* $s_A \times t_A = M \times N$. Refocused images thus have a shallow DoF, and the large amount of blur present in the image (see Fig. 3) explains the high PSNR values we obtain.

The main conclusion from our experiments is that for all datasets it is possible to obtain better RD performances with scenario A if the acceptable delay d_a^A is long enough, here 3s or 5s. The extent of the RD improvement depends on d_a^A and varies for each dataset: For the Lytro Illum and Stanford datasets, RD performances of scenario A are on par with scenario B when d_a^A is set to 2s. Scenario A is outperformed by scenario B when $d_a^A = 1$s. Overall, we observed that the RD improvement was slightly better for the Lytro Illum than for the Stanford dataset.

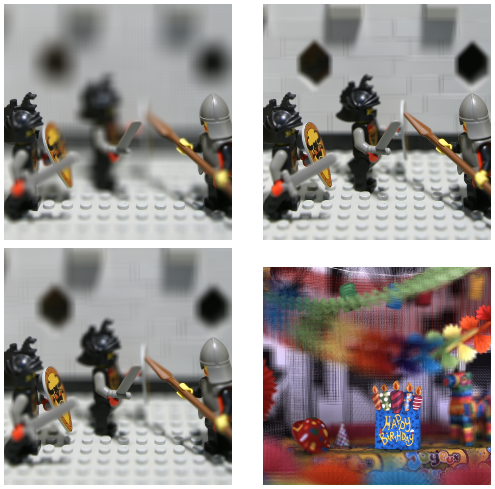

Fig. 3: Influence of the synthetic aperture size on the depth-of-field: 17×17 (top left), 9×9 (bottom left), and 5×5 (top right). Note that the DoF is also depending on the baseline, *e.g.* the 4×4 Technicolor light fields (bottom right) have a very shallow DoF due to their wide baseline.

For all light fields of the Technicolor dataset, scenario A outperforms scenario B for all values of d_a^A by a significant margin.

Thus it is clear that the RD performances depend more on the angular resolution rather than the spatial resolution, and scenario A can be very beneficial for light fields containing a small number of SAIs. Such behaviour can be expected, since less SAIs means a smaller bitrate for scenario A. The spatial resolution does not have an impact as the size of the refocused images is the same as the SAIs.

3.2. Influence of the synthetic aperture size

In this section we study the influence of the synthetic aperture size $s_A \times t_A$ on the RD performances. Reducing $s_A \times t_A$ means increasing the DoF of the refocused image, as illustrated in Fig. 3. Note that DoF is also related to the baseline of the light field, *e.g.* refocused images rendered from the Technicolor dataset with a full aperture (4×4) still have a shallower DoF than refocused images obtained from the Stanford dataset with a 5×5 aperture (see Fig. 3). For scenario A, reducing $s_A \times t_A$ also means reducing the de-noising effect and hence decreases the image quality. However, reducing $s_A \times t_A$ is also more challenging for scenario B as more high frequency details have to be encoded, which increases the bitrate.

We conducted our experiments on the Stanford dataset, which offers the highest angular resolution. We first consider the case where the user only requires a fixed aperture size. In scenario A, the number of SAIs to be transmitted is then only equal to the aperture size $s_A \times t_A$. RD performances for the representative LegoKnights light field are shown in Fig. 5 for a fixed aperture of size 5×5 (5.c), 9×9 (5.d), and 17×17 (5.e). Results show that the resulting reduced bitrate greatly compensate the decreased quality due to the weaker de-noising, and scenario A clearly outperforms scenario B for all values of d_a^A.

Second, we consider a more interactive scheme, where the user can dynamically choose the aperture size. In this case, all

[1] https://v-sense.scss.tcd.ie/research/light-fields/lf-streaming/

the light field SAIs have to be transmitted for scenario A. RD performances for LegoKnights are shown in Fig. 5.f, and show similar results to the fixed full size aperture (5.e).

4. CONCLUSION

This paper investigated the streaming performance of the direct transmission of a full light field scenario and an interactive transmission scenario with a refocusing application. Several test datasets were used corresponding to light fields of different spatio-angular characteristics. The de-noising effect of refocusing was also taken into account, as well as the impact of the synthetic aperture size. The study showed that it is possible to obtain better RD performances for all datasets by transmitting the whole light field when the startup delay is long enough, *e.g.,* 3s or 5s. RD gains can even be achieved with a short acceptable delay of 1s or 2s when the light field contains a smaller number of images.

Based on this finding, we plan to investigate more sophisticated streaming scenario, which could combine the two scenarios studied in this paper, *e.g.* to be robust to changes in network bandwidth conditions, or design predictive mechanisms for the interactive transmission. We also plan to assess more precisely the acceptable delay values with subjective tests. In addition, this next study will take into account more parameters such as the encoding, decoding, and rendering delays. Finally, we also wish to combine the refocusing application with novel viewpoint rendering.

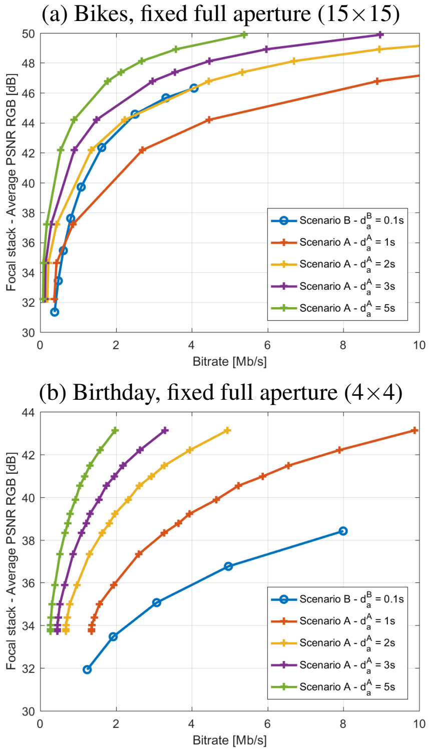

Fig. 4: RD curves representative of the Lytro Illum (a) and Technicolor (b) datasets.

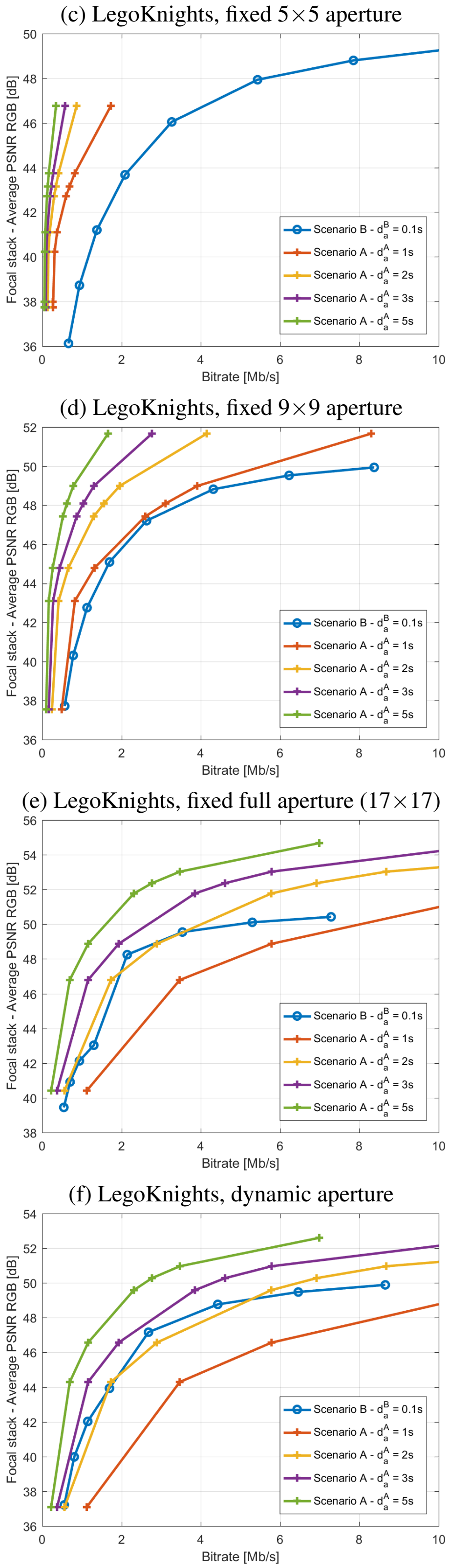

Fig. 5: RD curves representative of the Stanford dataset. Different aperture sizes are tested for the LegoKnights light field.

5. REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering", in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1996, SIGGRAPH '96, pp. 31–42, ACM.

[2] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields", in *Proc. SIGGRAPH*, 2000, pp. 297–306.

[3] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light Field Photography with a Hand-Held Plenoptic Camera", Tech. Rep., CSTR 2005-02, Apr. 2005.

[4] D. G. Dansereau, . Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras", *ACM Trans. Graph.*, vol. 34, pp. 15:1–15:20, 2015.

[5] P. Ramanathan, M. Kalman, and B. Girod, "Rate-distortion optimized interactive light field streaming", *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 813–825, June 2007.

[6] E. Peixoto, B. Macchiavello, E. Mintsu Hung, C. Dorea, and G. Cheung, "Progressive communication for interactive light field image data streaming", in *Proc. ICIP*, Sep. 2017, pp. 1925–1929.

[7] W. B. S. de Souza, B. Macchiavello, E. Peixoto, E. M. Hung, and G. Cheung, "A sub-aperture image selection refinement method for progressive light field transmission", in *Proc. MMSP*, Aug 2018, pp. 1–6.

[8] E. Peixoto, B. Macchiavello, E. M. Hung, and G. Cheung, "Progressive sub-aperture image recovery for interactive light field data streaming", in *Proc. ICIP)*, Oct 2018, pp. 3289–3293.

[9] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.

[10] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of Lytro-Illum light fields", in *Proc. IPTA*, Nov. 2015, pp. 494–497.

[11] K. Hara, M. Kawakita, T. Mishina, and M. Teratani, "Coding results for toys and trees on compression of dense representation of light fields", Tech. Rep. MPEG2019/M46376, ISO/IEC JTC1/SC29/WG11, Marrakech, Morocco, Jan 2019.

[12] D. G. Dansereau, D. L. Bongiorno, O. Pizarro, and S. B. Williams, "Light field image denoising using a linear 4D frequency-hyperfan all-in-focus filter", in *Proc. SPIE*, Feb 2013, vol. 8657, p. 86570P.

[13] "The stanford light field archive", `http://lightfield.stanford.edu/lfs.html`, accessed: 08-02-2019.

[14] N. Sabater, G. Boisson, B. Vandame, P. Kerbiriou, F. Babon, M. Hog, R. Gendrot, T. Langlois, O. Bureller, A. Schubert, and V. Alli, "Dataset and pipeline for multi-view light-field video", in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1743–1753.

[15] G. Alves, F. Pereira, and E. A. B. da Silva, "Light field imaging coding: Performance assessment methodology and standards benchmarking", in *Proc. ICME Workshops*, July 2016.

[16] I. Viola and T. Ebrahimi, "Quality assessment of compression solutions for icip 2017 grand challenge on light field image coding", in *Proc. ICME Workshops*, July 2018, pp. 1–6.

[17] M. Rerabek and T. Ebrahimi, "New Light Field Image Dataset", in *8th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2016.

[18] X. Jiang, M. Le Pendu, R. A. Farrugia, and C. Guillemot, "Light field compression with homography-based low-rank approximation", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1132–1145, Oct 2017.

[19] P. Matysiak, M. Grogan, M. L. Pendu, M. Alain, and A. Smolic, "A pipeline for lenslet light field quality enhancement", in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 639–643.

[20] D. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras", in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2013, pp. 1027–1034.

[21] "x265 HEVC Encoder / H.265 Video Codec", `http://x265.org/`, Jan 2018.

[22] ITU-T Recommendation, "Series G: Transmission systems and media, digital systems and networks: One-way transmission time", Tech. Rep. G.114, G.114, Geneva, Switzerland, March 2003.

[23] C Grunheit, A Smolic, and T Wiegand, "Efficient representation and interactive streaming of high-resolution panoramic views", in *2002 International Conference on Image Processing (ICIP)*, Sept. 2002, vol. 3, pp. III–209–III–212 vol.3.