

CIVIT DATASET: STEREOSCOPIC 3D-360 VIDEOS OF TYPICAL MEDIA PRODUCTION USE CASES

Filipe Gama¹, Sergio Moreschini¹, Ilmari Huttu-Hiltunen²,
Olli Suominen¹, Robert Bregovic¹ and Atanas Gotchev¹

¹Tampere University, Tampere, Finland; ²Rakka Creative Oy, Tampere, Finland

ABSTRACT

360-video allows for exploring visual scenes from all directions outwards in the form of a monoscopic panorama. It is one of the most widely-used types of media for head-mounted displays, where the panorama can be presented as a single viewpoint to both user's eyes. Alternatively, stereoscopic 3D-360 video content aims at generating two panorama videos, one for each eye of the user, thus providing a more realistic and immersive experience. However, the capture, compression, storage and visualization of this type of content are more challenging and require further research. To assist researchers working in the related areas, this paper presents a dataset comprising six stereoscopic 3D-360 videos captured by the Nokia OZO camera system. We overview the capture and processing pipeline, describe the specifics of the used device and list the captured scenes along with their characteristic features.

Index Terms — 360-video, virtual reality, stereoscopic, 3D-360 video, head-mounted display, dataset

1. INTRODUCTION

In recent years, we have experienced a fast grow of multimedia technologies that provide immersive experiences for users. Among others, 360-videos and Virtual Reality (VR) content are becoming more popular over time. Nowadays, 360-videos are considered one of the main applications for Head-Mounted Displays (HMDs). As a result, industry and the scientific community have been trying to improve and address different problems related to capturing, distribution and visualization of 360-video content [1–3].

360-videos, also known as omnidirectional videos, provide three Degrees-of-Freedom (3-DoF) in which the user can look around from the same viewpoint. In order to create 360-video content, the camera system senses the scene within a Field-of-View (FoV) of 360 degrees. This task is performed by systems with multiple cameras pointing at different directions. Regardless the type of system used to capture the scene, the output signal is typically represented as a panorama image (e.g., equirectangular projection). Though other projection functions exist and result in different image representations (e.g., cube format), panoramas are more convenient in terms of compatibility with existing file formats, encoders, streaming architectures, and content delivery networks [2]. Whenever a user watches a 360-degree video on a HMD, a portion of the panorama image, a.k.a. viewport, is displayed according to the user's head motion. In case there is only a single panorama available on the HMD, no depth perception is provided since the same viewport is shown simultaneously to the left and right eye of the user. This is the most common situation in 360-videos. However, the perception of depth may be obtained by



Figure 1. Nokia OZO [4].

generating two panoramas from different perspectives of the same scene to which we refer as (stereoscopic) 3D-360 videos. In such situations, different viewports are shown to each eye of the user.

Capturing and processing 3D-360 video content requires more sophisticated and expensive systems compared to conventional 360-video systems. Since 2015, several 360-video capturing systems have been developed and manufactured with different specifications, targeting a certain functionality or application. Nevertheless, only a restrict list of devices aimed at 3D-360 video content from the very beginning. One of the earliest products developed in this area came from Facebook with Facebook Surround 360, GoPro and Google with GoPro Odyssey and Nokia with Nokia OZO. More recent devices differ from these ones only in terms of their specifications such as resolution, frame-rate and device's size.

Throughout this paper, all 3D-360 video content was generated using Nokia OZO (shown in Figure 1) and their video editing tool called OZO Creator [4]. We present six 3D-360 videos of typical media production use cases in this paper. We have provided for each scene a stitched video file, fisheye images from each camera, and two different audio files that can be used to improve the current end-to-end 3D-360 systems.

The rest of the paper is organized as follows. Section 2 describes the main concepts related to 3D-360 video and respective capturing systems. Section 3 covers the specifications of the 3D-360 video capturing system used to produce our data. The contents of the dataset are discussed in Section 4 and we conclude our work in Section 5.

2. BACKGROUND

Underneath every 3D-360 capturing system there is a complex pipeline to capture, process and generate the content for HMDs. Unlike monoscopic 360-video, 3D video requires depth. Depth is often obtained by capturing the scene from different perspectives by moving the capturing system [5] or by using at least two cameras - the camera equivalent to our left and right eyes. However, depth can also be acquired by other means e.g., depth sen-

sors. Nowadays, depth sensors such as time-of-flight are capable of providing accurate depth estimations at high frame-rate and under low light conditions. This is a clear advantage over depth estimations from stereo/multi-cameras. Some commercial devices already integrate such technology in their products like in [6].

In 3D-360 capturing systems, we may highlight at least three main challenges: the camera hardware, the data management unit, and the stitching and rendering software. In terms of camera hardware, the cameras must be globally synchronized. Since this particular capturing systems deal with multiple cameras, it is crucial to ensure that each frame represents the same time instant. In addition, the number of cameras, the type of lens and the rig's structure are some of the specifications taken into account to achieve the best performance without compromising the integrity of the hardware or the output image quality. The data management unit is responsible for the integrity and control of the data flow among the cameras. It is also responsible for the centralization of the information so that it can be accessible in later stages. Finally, the stitching and rendering software deals with the fusion of several images into a seamless image. This last challenge is perhaps the most difficult to achieve because it requires sophisticated computational photography and computer vision techniques to accomplish it [7, 8]. The stitching and rendering software is composed by several blocks that play a major role in the quality of the output image. Figure 2 shows one of the most typical pipelines to stitch and generate a seamless panorama image. Firstly, all raw input images acquired by each camera are converted to RGB images. Typically, this process implies mutual camera color correction, anti-vignetting, gamma and tone curve mapping, sharpening, and demosaicing. Secondly, other image aberrations are taken into account at this stage including lens distortions, specially radial distortion. Once the images are compensated for the distortions, they are mapped to a polar coordinate system using e.g., equirectangular projection. This mapping not only eliminates redundant information that may exist between the images but also expresses all the information of each pixel on the image in the view space in terms of azimuth and elevation θ as shown in Figure 3. After this step, image registration takes place. Though many different algorithms to register images exist, feature-based algorithms are widely used in this context due to their robustness [7].

The feature-based algorithm contains three main steps: feature detection, feature matching and alignment. For feature detection and matching, SIFT algorithm [9] is often chosen because of its robustness and properties (i.e., scale, rotation, and translation invariant). Nevertheless, once an initial set of feature correspondences is computed, additional post-processing and refinement is necessary in order to find a set that produces a high-accuracy alignment. Typically, high-accuracy alignment requires iterative procedures that use least squares estimate over a subset of n correspondences (a.k.a. inlier correspondences). Moreover, if the capturing system is composed by several cameras, we also need to ensure global consistency (through Bundle Adjustment) in order to minimize the mis-registration between all pairs of image. Once we have computed the global alignment, we may also need to perform local adjustments such as parallax removal to reduce double images and blurring due to local mis-registrations.

After image registration, one has to decide how to produce the final stitched image (panorama). This involves selecting a final compositing surface (e.g., spherical) and sometimes the reference view i.e., determine which part of the scene will be centered in the final view. It also involves selecting which pixels contribute to the final composite and how to optimally blend these pixels to minimize visible seams, exposure differences, blur, and

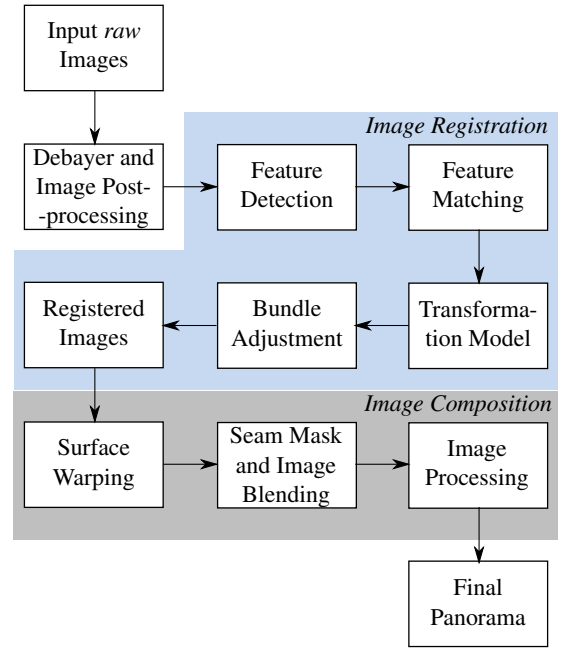


Figure 2. Conventional pipeline of a 2D 360-degree panorama formation.

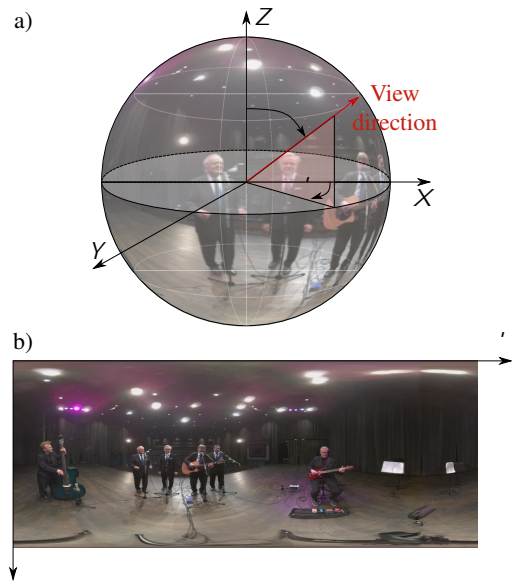


Figure 3. 360-degree panorama mapping (equirectangular projection): a) wrapped sphere; b) rectangular texture.

ghosting. Apart from some small modifications, 3D-360 capturing systems require the same stitching and rendering process as described above. In this particular case, the pipeline shown in Figure 2 is partly replicated and combined in order to obtain both left and right views. In other words, each view requires a dense set of rays that generates a smooth and continues signal along θ and ϕ directions, see Figure 4. This is possible thanks to the generation of novel views (a.k.a. virtual cameras or virtual views) using the real cameras information and view interpolation methods. View interpolation based on images is a broad topic in computer vision and therefore, several methods exist in the literature [10]. Like in the previous case, the last step of the stitching and rendering process is devoted to the formation of a stereo panorama image i.e., a single seamless image composed by both the left and right

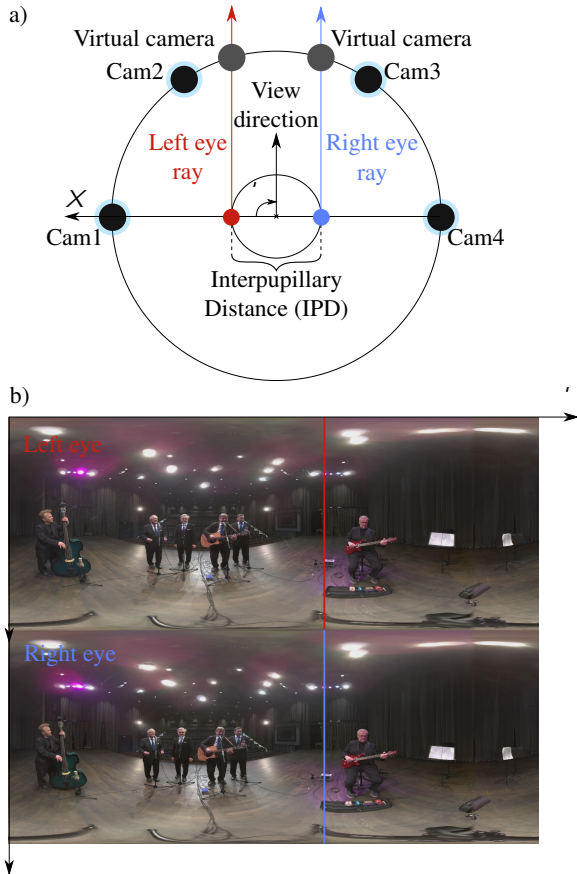


Figure 4. 3D-360 panorama rendering scheme: a) topology of Nokia OZO camera (cross-section -“equator”) with the left and right target views displaced by the inter-pupillary distance (IPD); b) generated panorama views for the left and right eyes.

panoramas.

After the creation of a seamless stereo panorama, the image is ready to be displayed on a HMD. The main task of the HMD is to select the correct viewport for each eye based on the user’s head motion, rectify the views and compensate for additional HMD lens distortions.

3. 3D-360 CAPTURE DEVICE

In the previous section, we covered the basic principles of how (3D-)360 video content is generated from cameras pointing outwards along a circular or spherical rig. The content presented in this paper was generated using Nokia OZO system illustrated in Figure 1 and Figure 5. Nokia OZO is equipped with 8 cameras. Each camera is paired with a wide-angle 195-degree FOV lens. Four cameras located along the “equator” (*Cam#1 - 4*), two on the top (*Cam#5 - 6*) and other two on the bottom (*Cam#7 - 8*) as depicted in Figure 5. Each camera produces images with 2048 x 2048 resolution at 30 frames per second. The system creates stereoscopic panoramas in almost all directions vertically and horizontally. However, due to the design of the system, the rear part of the system where the storage compartment is located occludes part of the area covered by the cameras. As a result, the user does not perceive depth in this direction.

In addition to the stereoscopic panorama, Nokia OZO also has eight microphones integrated around its body for recording spatial audio, which can be exported as a 4.0, 5.0, 7.0, or 8.0 channel

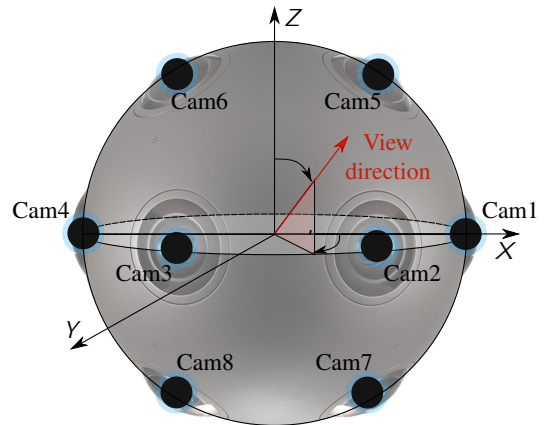


Figure 5. Nokia OZO layout: *Cam#1 - 4* are located along a cross-section called “equator”.

48kHz 24-bit PCM WAV file, first-order Ambisonic spatial audio, and as Nokias OZO Audio Interchange format.

The camera records both video and sound either to its own internal storage (500GB Digital Cartridge), or to an external Blackmagic HyperDeck Studio Pro 2 or Blackmagic HyperDeck Shuttle recorder via a 1.5Gbps HD-SDI cable.

The Nokia OZO compresses the eight 2K x 2K pixel images captured by the cameras into a single compressed data stream that emulates a 10-bit 1920 x 1080 full HD resolution video frame. Sound captured by the microphones is recorded as 8 discrete PCM channels, synced with the video. To post produce the sound and images from these raw camera recordings, the data is exported and edited using the Nokia software called OZO Creator.

4. DATASET

The images for the dataset were generated by an automated procedure available in OZO Creator. OZO Creator itself is conceptually similar to a standard post-production workflow where it is possible to edit, color correct and add visual effects to the content. Moreover, it is also possible to perform sound mixing and transcoding to deliverable formats. The best practices to stitch and edit the content are described in more detail in the reference manual [4].

The presented dataset consists of four main events (see Figure 6):

- **Pole vault** – Men’s pole vault international match in Tampere, Finland, 2016.
- **The concert** – Live music performed by Digiveikot feat Heikki Koivunen and Jouko Helatie: Tiistai keskiviikko, 2017.
- **The climbing** – Demonstration event of professional wall climbing in Tampere, Finland, 2017.
- **The clapping song, the circle, the bicycle** – Short takes of the 3D-360 movie Devils Lungs, directed by Alla Kovgan and produced by Hanna Pajala-Assefa / Loikka Kontakti, 2018.

Unlike others, the sequence “Pole vault” contains footage from two camera positions. For each set, we provide 4 or 8 fisheye images in OpenEXR file format, first-order Ambisonic spatial audio in WAV file format and the respective stereo panorama images with first-order Ambisonic audio that were generated using the fully automated stitching pipeline of OZO Creator. The dataset is available for download at www.civit.fi.

tion due to the limitation of the hardware and use of wide-angle lenses; robust, efficient and fully-automatized stitching and rendering pipelines; generation and compression of high resolution video at high frame-rate to overcome potential issues on the visualization phase (e.g., HMDs) since the data is displayed very close to the user's eyes; user interaction because "truly" immersive experiences require more than 3 degrees-of-freedom; adaptive high dynamic range content in order to meet closer the characteristics of the human eye.

6. ACKNOWLEDGMENT

The work in this paper was funded from the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

This research was supported by CIVIT (Centre for Immersive Visual Technologies).

7. REFERENCES

- [1] Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryavy, and Thomas Olsson, "Visualizing big data with augmented and virtual reality: challenges and research agenda," *Journal of Big Data*, vol. 2, 12 2015.
- [2] M. Zink, R. Sitaraman, and K. Nahrstedt, "Scalable 360 video stream delivery: Challenges, solutions, and opportunities," *Proc. IEEE*, vol. 107, no. 4, pp. 639–650, April 2019.
- [3] C. Anthes, R. J. Garca-Hernandez, M. Wiedemann, and D. Kranzlmüller, "State of the art of virtual reality technology," *2016 IEEE Aerospace Conference*, pp. 1–19, March 2016.
- [4] Nokia, "Ozo + post production workflow," https://docs.ozo.nokia.com/learn/OZO_Post_Workflow_Documentation.pdf, Accessed: 04-09-2017.
- [5] Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin, "6-dof vr videos with a single 360-camera," *Proc. 2017 IEEE Virtual Reality (VR)*, pp. 37–44, 2017.
- [6] NCTech, "Lasiris vr - 3d reality capture," <https://www.nctechimaging.com/lasiris/>, Accessed: 18-02-2018.
- [7] Richard Szeliski, "Image alignment and stitching: A tutorial," *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, Jan. 2006.
- [8] Debabrata Ghosh and Naima Kaabouch, "A survey on image mosaicing techniques," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 1 – 11, 2016.
- [9] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.
- [10] Shing Chan, Heung-Yeung Shum, and King-To Ng, "Image-based rendering and synthesis," *Signal Processing Magazine, IEEE*, vol. 24, pp. 22 – 33, 12 2007.

Figure 6. 3D-360 dataset scenes preview: a) Pole vault; b) The concert; c) The climbing; d) The clapping song; e) The circle; f) The bicycle.

5. CONCLUSION

In this paper we presented a (3D-)360 video dataset of different events that can support the research community to improve and develop further solutions towards the capturing, distribution and visualization of this type of content. In the future, more datasets will be needed in order to solve some of the challenges and open problems related to (3D-)360 systems, including: image distor-