Volograms & V-SENSE Volumetric Video Dataset

Rafael Pagés[1], Konstantinos Amplianitis[1], Jan Ondrej[1], Emin Zerman[2] and Aljosa Smolic[2]
[1]Volograms Limited, Guinness Enterprise Centre, Taylor's Lane, Dublin 8, Ireland
[2]V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland
{rafa, kostas, jan}@volograms.com, {emin.zerman, smolica}@tcd.ie

Keywords: Volumetric video, 3D reconstructions, Augmented Reality, Virtual Reality

Abstract: Volumetric video is a new form of visual media that enables novel ways of immersive visualisation and interaction. Currently volumetric video technologies receive a lot of attention in research and standardization, leading to an increasing need for related test data. This paper describes the Volograms & V-SENSE Volumetric Video Dataset which is made publicly available to help said research and standardisation efforts.

1 Introduction

Volumetric video is a media format that allows reconstruction of dynamic 3D objects from real life and their visualization in immersive applications such as augmented reality and virtual reality. Generally, the volumetric video is captured in dedicated studios using many cameras that are looking inwards and capturing synchronized videos (as further described in Section 2). The volumetric video is then generated in 3D form which changes with respect to time [4]. Examples of 3D models for a sample volumetric video are shown in Fig 1. The generated volumetric video can be represented with either colored point clouds or textured meshes [6].

There are different approaches and methods for content creation. The number of cameras that are looking inwards a studio range from 4 Microsoft Kinect cameras [7], 12 synchronised RGB cameras [4], and 32 cameras (i.e., 16 stereo cameras) [8] to 106 cameras which include both RGB cameras and infrared cameras [9]. Many different techniques are used for 3D reconstruction.

Contrary to many different volumetric video generation techniques, there have not been many volumetric video datasets. The 8i dataset [10] is a commonly used one for the point cloud representation of volumetric video. There are only two other publicly available volumetric video datasets: vsenseVVDB [11] and vsenseVVDB2 [6]. The former provides two volumetric videos in coloured point cloud format with four different point cloud densities. The latter provides four new volumetric videos in both coloured point cloud and textured 3D mesh formats.

Even though there have been many studies that focus on volumetric video creation, the number of datasets that provide publicly available volumetric video is very limited. With new MPEG standardisation activities on compression and quality assessment of point clouds and dynamic meshes as well as intensified research in this area, there is a need for new datasets.

This paper introduces the Volograms & V-SENSE Volumetric Video Dataset, which releases three new volumetric videos in differing characteristics (i.e., different texture and different movement characteristics) with different durations. The following sections describe the content creation, applications, and details of the new dataset.

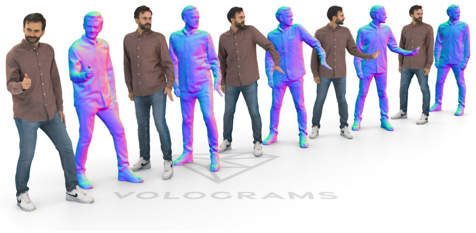

Figure 1. Example of 3D models (textured and non-textured) from different time instances of a volumetric video.

2 Volumetric Video Content Creation

2.1 Capture Setup

The volumetric capture studio, located in Dublin, Ireland, is a cubic room with an approximate capture volume consisting of a 2m radius length cylinder. The studio contains an aluminium frame covered in green fabric, which also covers the ceiling. The floor of the capture space is also green. See Fig. 2 below.

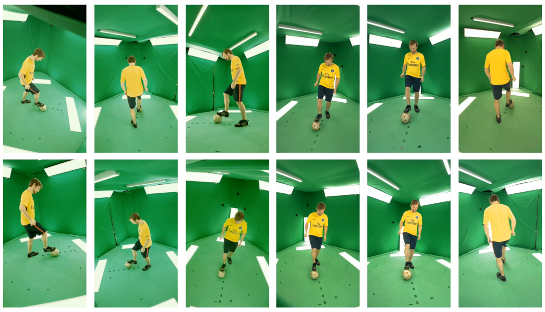

Figure 2. Different viewing angles of the capture studio in Dublin, Ireland.

2.2 Cameras & Capture capacity

There are 12 cameras mounted to the aluminium frame:

- 6 Blackmagic Micro Studio FullHD cameras.
- 6 Blackmagic Micro Studio 4k cameras.
- 12 Olympus 7-14mm f/2.8 Pro M.Zuiko Digital ED lenses.
- 6 Blackmagic Video Assist.

The capture capacity of the studio is approximately 60 minutes and the cameras are synchronised using a BMD Sync Generator and a 20-port video/sync amplifier.

2.2.1 Lighting

The studio is evenly lit by white light (4400K) emanating from an arrangement of 10 LED industrial light panels (0.3m x 1.2m) located at different heights.

2.2.2 Geometric calibration

The cameras are modelled using the pinhole camera model [3] including radial and tangential distortions, which are used to remove the distortion from the images after calibration.

To calibrate the cameras, a calibration totem is used with the following characteristics:

- A set of cubes placed with multiple orientations, to guarantee different planar surfaces. These cubes can be simple cardboard boxes or similar, making it very simple to build by third parties. The height of the totem needs to be similar to a person's height.
- Unique random patterns generated using the approach by Li et al [1], and attached to each planar surface of the totem.

An example of the calibration totem is provided in Fig. 3.

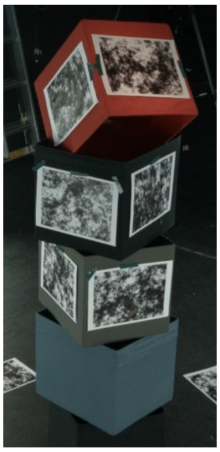

Figure 3. Calibration totem example.

The totem is placed in several positions inside the studio and captured with all the cameras. The automatic calibration process is done by running a multiple structure from motion (SfM) system, which takes the different positions of the totem separately, but uses all the sets of point matches in a single SfM and bundle adjustment problem.

2.2.3 Radiometric calibration

To perform radiometric calibration, a Macbeth ColorChecker (24 squares) is used.

2.2 3D Reconstruction Pipeline

To generate the 3D data, our volumetric pipeline uses a proprietary algorithm to combine the best of SfM, MVS (Multi View Stereo) and volume estimation to guarantee the generation of both detailed and complete 3D human models, even when a reduced number of cameras is used in the capture. The global approach is described in the work of Pagés et al [4].

Firstly, a foreground segmentation algorithm separates the performer from the background on each of the cameras, and the resulting foreground masks are used to estimate a scene volume, through the visual hull. In the following stage, a multi-depth estimation algorithm computes a depth map for each of the cameras, by using a MVS algorithm in the volume constrained by the visual hull, which improves accuracy and performance.

Furthermore, the resulting point cloud is fused with the estimated scene volume in an intelligent way by statistically analysing the differences between the MVS point cloud and the volume point cloud (the vertices of the volume mesh) in the voxel space, identifying missing geometry and keeping only the information that is denser and more accurate [4]. Next, a volume-constraint Poisson Surface Reconstruction [5] process is applied to obtain the final detailed mesh, avoiding the connection of small gaps in the model. Lastly, a re-meshing process to reduce the polygon count of the resulting models.

Once a model per frame has been obtained, it is necessary to apply a key-framing and temporal consistency process: a fundamental step with two key purposes. The first step assures the meshes are temporally coherent, reducing the flickering and other temporal artefacts. The second is identifying temporal redundancies, reducing the amount of information to store and enabling smaller files. For this, we analyse the motion of the sequence through the optical flow and use the flow correspondences to drive a robust ICP algorithm that registers meshes in a tracking sub-sequence.

The last step of the pipeline is generating a set of UV maps and colouring them. We use D-charts to generate the texture atlases and the method by Pagés et al. [2] to blend the colour information from the different cameras.

2.3 Applications

The generated volumetric video can be used in many different applications such as education, museums, (or cultural heritage), tour guide, entertainment, telepresence, or teleconference applications, etc.

One of the main applications of volumetric video is telepresence. That is, the users can project their 3D images onto another person's reality and can feel "present" in that reality. One of the first efforts that uses volumetric video to achieve this was Microsoft's holoportation system which used HoloLens head-mounted displays [12], more recently, newer systems also developed to do similar tasks [13]. In another instance, Trinity College Dublin's then Provost was recorded and his image was shown at the Trinity Business and Technology Forum 2018 [14].

Another interesting venue of application is the cultural heritage applications. With volumetric video, Samuel Beckett's "Play" could be reenacted in AR or VR (i.e., Virtual "Play") [15], Jonathan Swift's likeness can reappear at Trinity College Dublin Library [16], or James Joyce's novel character Stephen Dedalus from Ulysses can be brought to life in VR [17]. A sample of these projects can be seen in Fig. 4.

Other applications can include education, empathy building (e.g., the creative experiment of "Bridging the Blue") [18], or entertainment (e.g., Awake: Episode One) [19].

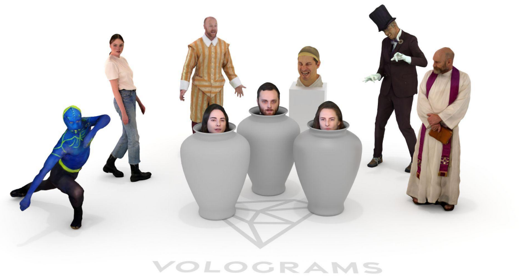

Figure 4. Some projects using Volograms technology

3. Details of the Dataset

This dataset includes three sequences featuring three different characters, each of them captured with a different purpose and application in mind. The three sequences feature male characters with varying skin colour, clothing, stature and range of movements.

Rafa

This sequence shows a performer, Rafa, who does a quick electro-move in a five seconds clip. Rafa wears a standard shirt and jeans, which is representative of many volumetric captures done nowadays. He also finishes his moves with a thumbs up gesture, which shows the reconstruction accuracy with fingers. Rafa was captured at V-SENSE's 12 camera studio in Dublin, Ireland. Meshes are ~40k polys/frame and texture images are 4069x4069. A sample frame of this dataset is provided in Fig. 5.

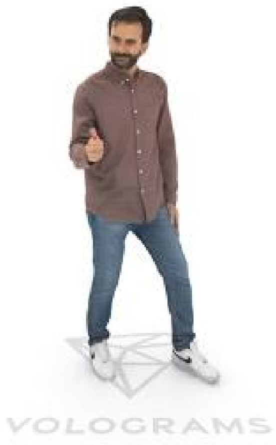

Figure 5. Sample frame of Rafa sequence.

Levi

Five second dancing sequence featuring Levi, an incredibly talented performer. Levi's moves are fast, dynamic and very complex, which poses a great challenge for 3D reconstruction algorithms. The models present very accurate details as fingers and facial features, even when motion blur could be an issue for the reconstruction process. Levi was captured in a 60 camera studio in California, US. Meshes are ~40k polys/frame and texture images are 4069x4069. A sample frame of this dataset is provided in Fig. 6.

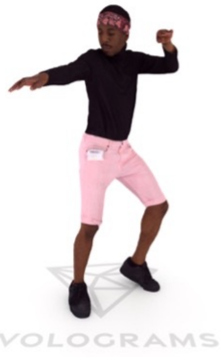

Figure 6. Sample frame of Levi sequence.

Sir Frederick

One minute monologue sequence featuring an actor performing as Sir Frederick Hamilton, from the Manorhamilton Castle in Leitrim, Ireland. This capture was done for an immersive cultural activation at the castle. Sir Frederick wears mediaeval clothing with some dark and shiny elements, which are typically challenging for 3D reconstruction. Sir Frederick speaks to the audience, and his facial features are very expressive. Sir Frederick was captured in a 12 camera studio captured at V-SENSE's 12 camera studio in Dublin, Ireland. Meshes are ~40k polys/frame and texture images are 4069x4069. A sample frame of this dataset is provided in Fig. 7.

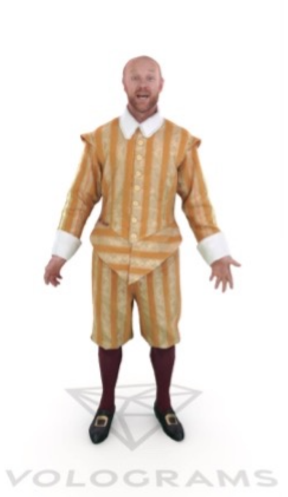

Figure 7. Sample frame of Sir Frederick sequence.

CONCLUSIONS

This paper introduces the Volograms & V-SENSE Volumetric Video Dataset which includes three volumetric video sequences that are created with a different application scenario in mind. Each of the sequences have different characteristics in terms of texture (e.g., clothing, skin colour) and movement (e.g., little movement to fast varying movement). Furthermore, the volumetric video sequences have different durations. These aspects make this dataset unique for its use in scientific studies and standardisation activities.

ACKNOWLEDGEMENTS

This work is part of the INVICTUS project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952147. It reflects only the authors' views and the Commission is not responsible for any use that may be made of the information it contains.

This publication has emanated partially from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number15/RP/27760.

REFERENCES

1. Li B., Heng L., Koser K., Pollefeys M., (2013), “A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern”, 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 2013, pp. 1301-1307, doi: 10.1109/IROS.2013.6696517.
2. Pagés, R., Berjón, D., Morán, F., & García, N. (2015, February). Seamless, Static Multi-Texturing of 3D Meshes. In Computer Graphics Forum (Vol. 34, No. 1, pp. 228-238)
3. Hartley, R., Zisserman, A., Multiple View Geometry in Computer Vision, 2000, Cambridge University Press, ISBN: 0521623049
4. Pagés, R., Amplianitis, K., Monaghan, D., Ondřej, J., & Smolić, A. (2018). Affordable content creation for free-viewpoint video and VR/AR applications. *Journal of Visual Communication and Image Representation*, *53*, 192-201.
5. Kazhdan, M., & Hoppe, H. (2013). Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, *32*(3), 1-13.
6. Zerman, E., Ozcinar, C., Gao, P., & Smolic, A. (2020, May). Textured mesh vs coloured point cloud: A subjective study for volumetric video compression. In 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX) (pp. 1-6). IEEE.
7. Alexiadis, D. S., Zarpalas, D., & Daras, P. (2012). Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras. IEEE Transactions on Multimedia, 15(2), 339-358.
8. Schreer, O., Feldmann, I., Renault, S., Zepp, M., Worchel, M., Eisert, P., & Kauff, P. (2019, September). Capture and 3D video processing of volumetric video. In 2019 IEEE International conference on image processing (ICIP) (pp. 4310-4314). IEEE.
9. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., ... & Sullivan, S. (2015). High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG), 34(4), 1-13.
10. d’Eon, E., Harrison, B., Myers, T., & Chou, P. A. (2017). 8i voxelized full bodies, version 2–A voxelized point cloud dataset. ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document m40059/M74006.
11. Zerman, E., Gao, P., Ozcinar, C., & Smolic, A. (2019). Subjective and objective quality assessment for volumetric video compression. Electronic Imaging, 2019(10), 323-1.
12. Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., ... & Izadi, S. (2016, October). Holoportation: Virtual 3d teleportation in real-time. In Proceedings of the 29th annual symposium on user interface software and technology (pp. 741-754).
13. Jansen, J., Subramanyam, S., Bouqueau, R., Cernigliaro, G., Cabré, M. M., Pérez, F., & Cesar, P. (2020, May). A pipeline for multiparty volumetric video conferencing: transmission of point clouds over low latency DASH. In Proceedings of the 11th ACM Multimedia Systems Conference (pp. 341-344).
14. V-SENSE. (2018, October 22). Volumetric Video of Trinity Provost Patrick Prendergast. VSENSE Trinity Provost Patrick Prendergast. Retrieved March 21, 2022, from https://v-sense.scss.tcd.ie/news/trinity_provost_prendergast/
15. O’Dwyer, N., Johnson, N., Bates, E., Pagés, R., Ondřej, J., Amplianitis, K., ... & Smolić, A. (2017, October). Virtual play in free-viewpoint video: Reinterpreting samuel beckett for virtual reality. In 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct) (pp. 262-267). IEEE.
16. O’Dwyer, N., Zerman, E., Young, G. W., Smolic, A., Dunne, S., & Shenton, H. (2021). Volumetric Video in Augmented Reality Applications for Museological Narratives: A user study for the Long Room in the Library of Trinity College Dublin. Journal on Computing and Cultural Heritage (JOCCH), 14(2), 1-20.
17. O’Dwyer, N., Young, G. W., & Smolic, A. (2022). XR Ulysses: addressing the disappointment of cancelled site-specific re-enactments of Joycean literary cultural heritage on Bloomsday. International Journal of Performance Arts and Digital Media, 1-19.
18. Arielle, L. G. and Smolic, A. “Bridging the Blue“, in The Art Exhibit at ICIDS 2019 Art Book: The Expression of Emotion in Humans and Technology, edited by Ryan Brown and Brian Salisbury, pp. 15-27, Carnegie Mellon University, Pittsburgh: ETC Press, 2020, ISBN: 9781716510809.
19. Start VR: Awake (Accessed: Jan 2020). https://startvr.co/project/awake/.