https://doi.org/10.2352/lim.2022.1.1.13

Volumetric Video Content Creation for Immersive XR Experiences

Aljosa Smolic[1,3]*, Konstantinos Amplianitis*[2,3]*, Matthew Moynihan*[3]*, Neill O'Dwyer*[3]*, Jan Ondrej*[2,3]*, Rafael Pagés*[2,3]*, Gareth W. Young*[3]*, Emin Zerman*[3]

[1]*Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland*
[2]*Volograms Limited, Guinness Enterprise Centre, Taylor's Lane, Dublin, Ireland*
[3]*V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland*

Abstract

Volumetric video (VV) is an emergent digital media that enables novel forms of interaction and immersion within eXtended Reality (XR) applications. VV supports 3D representation of real-world scenes and objects to be visualized from any viewpoint or viewing direction; an interaction paradigm that is commonly seen in computer games. This allows for instance to bring real people into XR. Based on this innovative media format, it is possible to design new forms of immersive and interactive experiences that can be visualized via head-mounted displays (HMDs) in virtual reality (VR) or augmented reality (AR). This paper highlights technology for VV content creation developed by the V-SENSE lab and the startup company Volograms. It further showcases a variety of creative experiments applying VV for immersive storytelling in XR.

Introduction

Volumetric video (VV) is a media format that represents 3D content as captured and reconstructed from the real world by cameras and other sensors, in a similar way as it is commonly known from computer graphics. With that it enables visualization of such content in full 6-degrees-of-freedom (6DoF). VV has seen interest from researchers in computer vision, computer graphics, multimedia and related fields over the last decades, often under other terms such as free viewpoint video (FVV), 3D video and others. Commercial application was, however, limited to few cases in special effects and game design.

Recent years have seen significantly growing interest in VV, including research, industry and standardization. This is driven on one hand by maturation of VV content creation technology, which has reached acceptable quality today for a variety of commercial applications. On the other hand, current interest for eXtended Reality (XR) also drives importance of VV, because VV allows to bring real people into XR.

In this paper, we outline content creation technology and workflows as developed by the V-SENSE lab of Trinity College Dublin and the startup company Volograms, which is commercializing VV content creation. We also present some aspects related to the content delivery pipeline of VV. Finally, we give an overview of creative experiments of V-SENSE and Volograms that develop immersive storytelling practice and study related user experience.

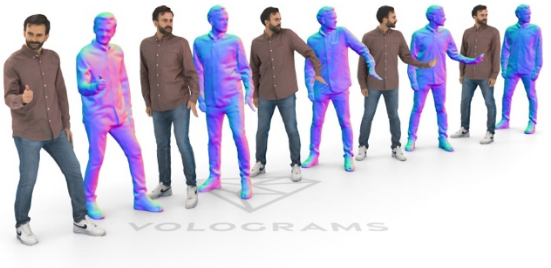


Figure 1. Example of 3D models (textured and non-textured) from different time instances of a volumetric video.

Volumetric Video Content Creation

Traditionally, VV content creation starts with synchronized multiview video capture in a specific studio as illustrated in Fig. 2. This shows the affordable setup in the V-SENSE lab in Dublin, which only uses 12 conventional cameras. Other studios can be equipped with up to hundred cameras and additional depth sensors [1]. Typically, the captured video and other data are then passed to a dedicated 3D reconstruction process.

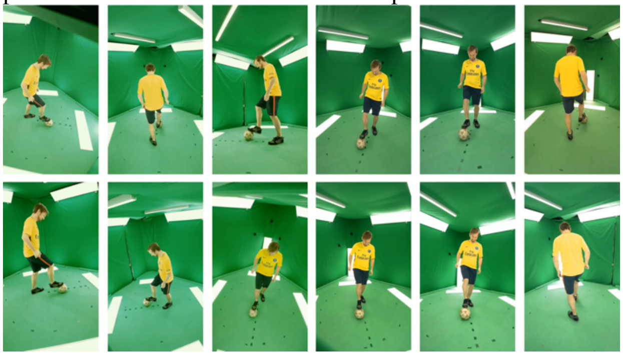


Figure 2. VV capture in the V-SENSE studio in Dublin, Ireland.

Classical Approaches

Classical approaches for VV content creation mostly rely either on structure-from-motion (SfM) type approaches, or on shape-from-silhouette (SfS) type approaches. While SfM relies on features and matching, and results in a dynamic 3D point cloud in the first place, SfS computes a volume populated by the object of interest in the first place. Both approaches have their advantages and drawbacks. Pagés et al. [2] present a system (Fig. 3) that combines advantages of both and is addresses affordable capture setups as the one in Fig. 2.

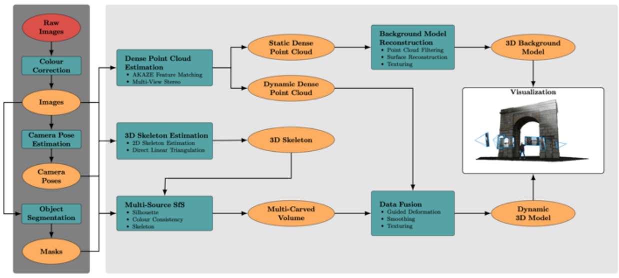


Figure 3. Overview of the affordable VV content creation pipeline as described in [2].

Typical SfM and SfS methods operate on a frame-by-frame basis, which, when applied to video, introduce temporal inconsistencies and flicker. Moreover, changing topology and connectivity makes editing cumbersome and compression inefficient. Thus, algorithms are required to enforce temporal coherence over VV. One approach to coherent, compressible VV content creation is the use of a template-based capture method [3,4]. However, these approaches are limited to the geometric constraints of the template i.e. clothing, number of connected components, etc. Another approach is to perform frame by frame registration [5,6]. This approach is generalizable and can be performed using keyframe strategies similar to that in video compression. Moynihan et al. [6] propose a system which automatically enforces temporal coherence end-to-end and also allows the user to sculpt the geometry during keyframes which can be tracked across the sequence.

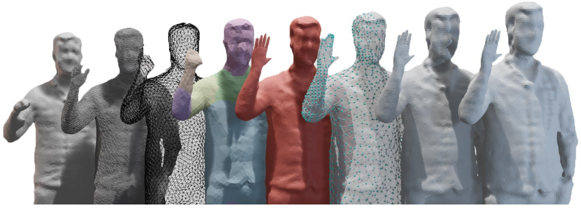


Figure 4. Mesh tracking and registration is typically applied for efficient representation and to enable editing of VV [6].

Deep Learning Approaches

Recently, powerful deep learning approaches have been presented for 3D geometry processing and reconstruction. First examples were able for instance to reconstruct 3D shape of an object from a certain class such as a chair from a single 2D image. 3D reconstruction of human faces from monocular images or video is another area that received a lot of attention. PIFu [7] is a method for single image 3D reconstruction of human bodies, which represents a milestone in this area.

Inspired by these developments, Volograms developed its own suite of AI powered algorithms to enable VV capture directly on a smartphone, bringing this technology to a whole new generation of creators. Volograms' AI back-end is able to generate a VV of a person from a single video casually captured in a non-controlled environment with a handheld smartphone, and from a single view point. To do so, the system first segments the user from the background semantically understanding different body parts and clothes; next, this information, together with the input image, is provided to a volume estimation network that outputs a textured 3D mesh representing the person at every frame of the video; lastly, using a temporal consistency algorithm as the ones mentioned above, the full sequence is compressed and encoded. The system is also able to calculate the underlying skeleton, which is a very useful creative tool. Fig. 5 shows some of the steps of this process.

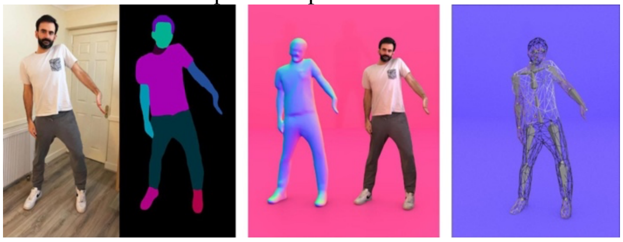


Figure 5. Different steps in the Volograms AI pipeline that use deep learning. Left: input image and semantic segmentation. Centre: volumetric reconstruction without and with texture. Right: underlying 3D skeleton.

Content Delivery Pipeline for Volumetric Video

Raw VV, i.e. time-varying 3D geometry with textures, results in huge data, e.g. hundreds of MB per minute of content. Therefore, efficient compression is crucial for efficient transmission and storage of such content. Standardization bodies like MPEG and JPEG are working on related solutions for compression and streaming. Further, quality assessment is crucial for evaluation and development of systems and components. This includes subjective methods as well as objective metrics, which are both very active research areas currently. Such research requires suitable test data. Volograms and V-SENSE provided several VV data sets to support the community as the one illustrated in Fig. 6 [8]. The full content delivery pipeline for VV is shown in Fig. 7.

Figure 6. Samples from volumetric video dataset by Volograms & V-SENSE [8].

There are 2 common ways to represent VV, as dynamic 3D point clouds or dynamic 3D meshes (Fig. 8). Both have advantages and draw backs. Point clouds can be seen as extension of video into the 3rd dimension, and many related processing and coding algorithms can be extended accordingly. Therefore, the video processing and coding community tends to prefer this representation. On the other hand, 3D meshes easily integrate with common graphics and pipelines, i.e., rendering. Therefore, the computer graphics community but also the VV industry tends to prefer this representation.

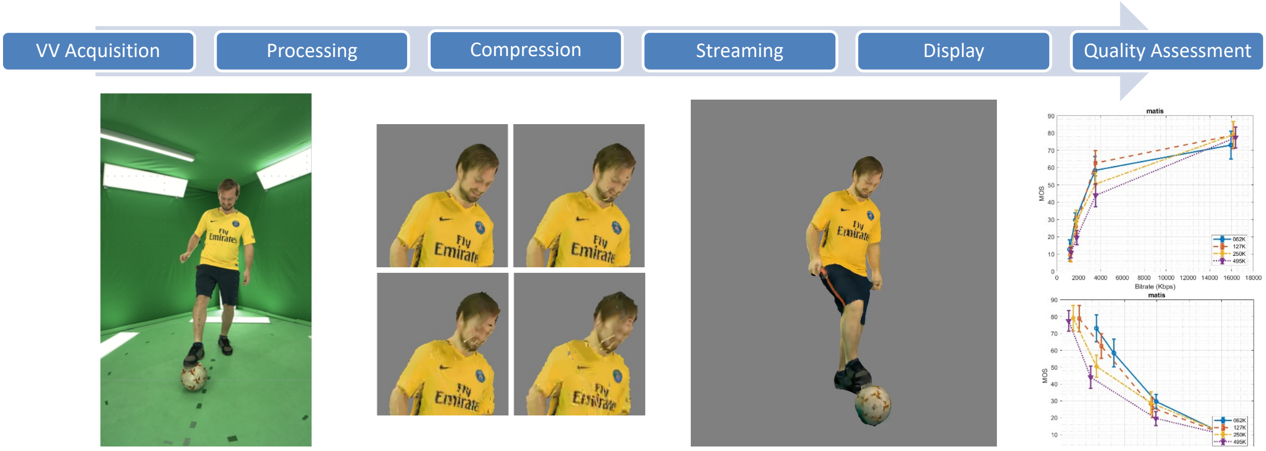


Figure 7. VV content delivery pipeline.

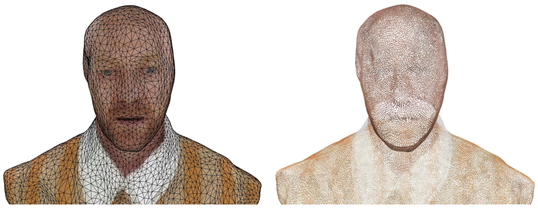
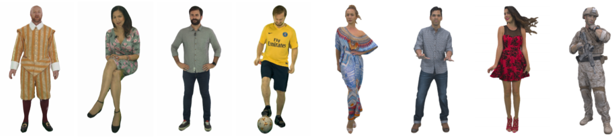


Figure 8. Volumetric video represented as 3D mesh (left) and 3D point cloud (right) [10].

It remains unclear though, which the pros and cons of both representations are, thus this is a very interesting research question, first formulated in [10]. This paper also presents a database (Fig. 9, vsenseVVDB2) that provides the same VV content in point cloud as well as in mesh format, and with that allows for comparative research. Initial results are presented in the paper as well.



(a) AxeGuy [v:25K / p:405K] (b) LubnaFriends [v:25K / p:402K] (c) Rafa2 [v:25K / p:406K] (d) Matis [v:25K / p:406K] (e) Longdress [p:765K] (f) Loot [p:784K] (g) Redandblack [p:729K] (h) Soldier [p:1.06M]

Figure 9. Samples from volumetric video dataset vsenseVVDB2 [REF].

Examples of XR Projects with Volumetric Video

As technology for volumetric video content creation, delivery and visualization is maturing, the novel media format becomes increasingly interesting for novel forms of immersive and interactive applications. However, creative aspects of such applications as well as related human factors are still relatively unresolved. To address this, we engaged in a number of creative experiments to implement, showcase, and evaluate XR storytelling with volumetric video, with a focus to develop novel forms of immersive storytelling and to study related user experience. In the following section, we give an overview of some creative experiences produced by V-SENSE. More details are available online [11].

MR Play – After Samuel Beckett

Virtual Play is a reimagining of Samuel Beckett's theatrical text, Play (1963) for virtual reality (VR). By making the user a key figure in how the story unfolds, the project explores the redefined relationship between author and audience, made possible by digital interactive technologies. In the story three characters are doomed to repeat the sorry tale of their love triangle, into infinity. Play was chosen because it specifically engages the questions of dialogue and interactivity. The sequence of the actors speaking is determined by a moving spotlight, which Beckett calls the 'inquisitor' [12] "they speak when the light is on them, and fall silent when the light is off" [13]. "Play is a game of interaction between the light operator and the actor, mediated by light technology" [14]. In the theatre, the audience passively observe the interaction; but, "in our VR version we acknowledge the role of the user as active; we recognise new opportunities for narrative and give the power of activation over to the end user, whose gaze becomes the spotlight. The user thus embodies the 'interrogator' and is empowered to independently discover the story, merely by looking at the characters" [15]. The user is placed in the centre and is surrounded by the three characters in urns, which are spaced far enough apart to allow the user to experience a natural sensation of movement, whilst exploring the three monologues. Six degrees of freedom is afforded by the ability to move around the virtual environment, and the experience is one of active immersion as opposed to passive observation. Virtual Play questions the essence of the performance spectacle in digital culture. The project aims to investigate how narrative, perception, communication, and embodiment have been altered through contemporary media, and asks how they might operate in the future.

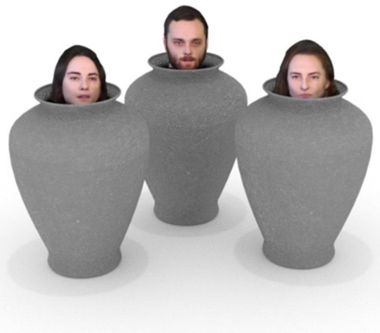


Figure 10. VV elements in MR Play.

Jonathan Swift: An extended reality VV application for Trinity Library's Long Room

XR technologies are quickly establishing themselves as commonplace platforms for presenting objects of historical, scientific, artistic, and cultural interest to the public. The Jonathan Swift project exemplars the creative potential of VV used in the museological context of cultural heritage (CH) [16]. In this applied use-case scenario, we present a VV actor as an AR character that attempts to break the mold by employing a humorous and playful mode of communication [17]. Our bespoke AR experience harnesses VV to create a digital tour guide that playfully embellishes the museological experience of museum visitors. This project discusses the appeal, interest, and ease of using this ludic storytelling strategy mediated via AR technology in a CH context [18].

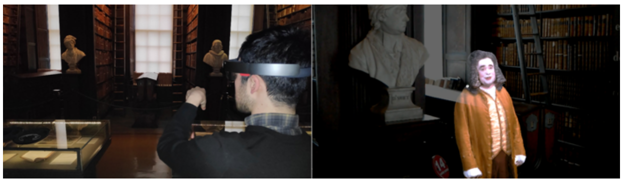


Figure 11. Example of an AR enhanced museum experience with VV.

Bridging the Blue

Bridging the Blue (BtB) is an immersive experience by Lubna Gem Arielle that explores the potential of VR as an empathy machine [19]. It applies principles from game design and non-linear storytelling. The user enters a virtual world via a portal that is designed as an island, where interactive symbols are distributed, which allow the user to enter one of 7 different scenes. Each scene is a separate VR environment where the user meets the protagonist as a VV, and as they proceed, the narrative is explored in a non-linear and interactive way. BtB aims to create empathy for clinical depression, and qualitative feedback from audiences indicates that the immersive format with a real person is well suited to achieve this.

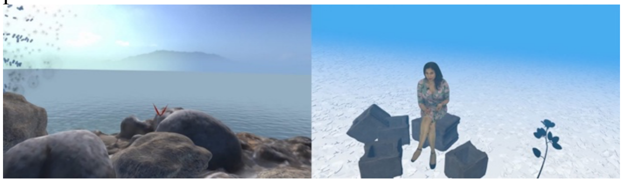


Figure 12. VR with VV as empathy machine in Bridging the Blue.

Image Technology Echoes

The immersive experimental fiction *Image Technology Echoes* explores paradigms of the objective and subjective world, the inner self, perception and embodiment through VR. We meet the 2 protagonists in a gallery from a 3rd person perspective but can also dive in each's inner self. The two VV characters were captured separately and combined in post-production. The work received the prestigious VR Art Prize, awarded by Deutsche Kreditbank (DKB) [20].

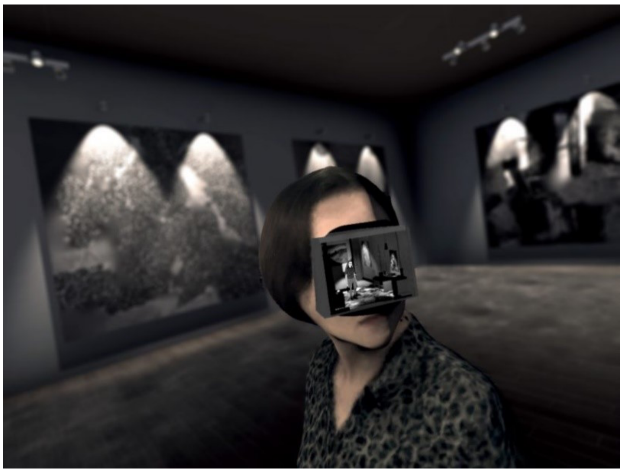


Figure 13. Impression from Image Technology Echoes, an award-winning VR experience with VV.

XR Ulysses

XR Ulysses is a creative project investigating possibilities for live performance using three-dimensional, volumetric video filming techniques in conjunction with XR technologies, including VR and AR. XR Ulysses is part of a series of innovative performance experiments hybridizing theatre and XR and "investigating questions around preservation, access, reactivation, and the transmission of dramatic/literary heritage in the twenty-first century" [21]. In the VR version audiences are invited to don a head-mounted display and immerse themselves in virtual simulations of scenes from James Joyce's classic work of modern literature, Ulysses [22], or, in the AR version they can go to the physical locations and engage characters using smartphones or tablets. The represented scenes include Buck Mulligan's famous speech about death from the opening scene on the top of the Martello Tower at Sandycove (from Episode 1, 'Telemachus'), and Stephen Dedalus' satirical "Parisian Parleyvoo" in Bella Cohen's establishment, (from Episode 15, 'Circe'). The major stand of enquiry underpinning this creative experiment consists of better understanding how VV and XR technologies can be combined with theatre practices to augment site specific performances [23].

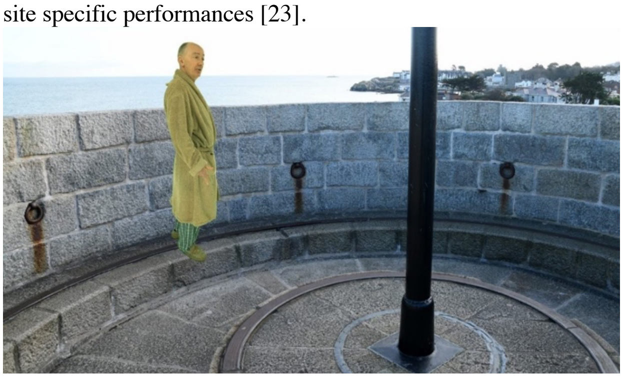


Figure 14. Impression from XR Ulysses.

XR Music Videos

Contemporary VV productions offer an array of entertainment value to existing and emergent 3D media practices. Our initial user-focused XR Music Video research showed how users responded to volumetric representations of music performance in VR [24]. This investigation was the starting point for a more sophisticated, interactive music video featuring the Irish rock band New Pagans that applied a user-centered design approach. Beyond its innate entertainment potential, we explored how VV music videos are captured, edited, and accessed for live performance reproductions. By approaching VV productions in this way, objects of attention within an immersive virtual environment were presented so that audiences could move around and interact and engage with creative materials, making the experience fundamentally different, unique, and rewarding.

Figure 15. A concept image for New Pagan's Lily Yeates VR music video.

Conclusion

VV as an emergent format of 3D digital media receives increasing attention among researchers, creatives, and audiences. It enables novel forms of 6DoF applications and open new business opportunities. Classical content creation is mature enough, while still relatively complex and thus expensive. Deep learning solutions open the way to user generated VV content and other more widely affordable use case scenarios. In particular, novel neural representations such as Neural Radiance Fields (NeRF) hold promise for further improvement, as recently shown for VV [25]. These can use multiview input and create impressive results, while the computational complexity is still very high. Bridging the gap to the animation world, i.e., making VV editable and animatable is an area of further research. Also, the content delivery pipeline is an area of research opportunities, e.g., regarding coding, streaming, related standards, as well as quality assessment and related metrics. While prototypes have been demonstrated, real-time processing for applications like holographic telecommunication is still in early stages of development.

Acknowledgements

This publication has emanated partially from research conducted with the financial support of Science Foundation Ireland (SFI) under the GrantNumber15/RP/27760.

Some of 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.

References

[1] Collet A, Chuang M, Sweeney P, Gillett D, Evseev D, Calabrese D, Hoppe H, Kirk A, Sullivan S. High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG). 2015 Jul 27;34(4):1-3.

[2] Pagés R, Amplianitis K, Monaghan D, Ondřej J, Smolić A. Affordable content creation for free-viewpoint video and VR/AR applications. Journal of Visual Communication and Image Representation. 2018 May 1;53:192-201.

[3] Habermann M, Xu W, Zollhoefer M, Pons-Moll G, Theobalt C. Livecap: Real-time human performance capture from monocular video. ACM Transactions On Graphics (TOG). 2019 Mar 13;38(2):1-7.

[4] Xu W, Chatterjee A, Zollhöfer M, Rhodin H, Mehta D, Seidel HP, Theobalt C. Monoperfcap: Human performance capture from monocular video. ACM Transactions on Graphics (ToG). 2018 May 21;37(2):1-5.

[5] Li H, Sumner RW, Pauly M. Global correspondence optimization for non-rigid registration of depth scans. InComputer graphics forum 2008 Jul (Vol. 27, No. 5, pp. 1421-1430). Oxford, UK: Blackwell Publishing Ltd.

[6] Moynihan M, Ruano S, Smolic A. Autonomous tracking for volumetric video sequences. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2021 (pp. 1660-1669).

[7] Saito S, Huang Z, Natsume R, Morishima S, Kanazawa A, Li H. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. InProceedings of the IEEE/CVF International Conference on Computer Vision 2019 (pp. 2304-2314).

[8] Pagés, Rafael, Konstantinos Amplianitis, Jan Ondrej, Emin Zerman, and Aljosa Smolic. "Volograms & V-SENSE Volumetric Video Dataset." ISO/IEC JTC1/SC29/WG07 MPEG2021/m56767 (2021) DOI:10.13140/RG.2.2.24235.31529/1.

[10] Zerman, Emin, Cagri Ozcinar, Pan Gao, and Aljosa Smolic. "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, 2020.

[11] https://v-sense.scss.tcd.ie/

[12] Beckett, Samuel. 2006. 'Play'. In The Complete Dramatic Works of Samuel Beckett, 305–20. London: Faber & Faber.

[13] O'Dwyer, Néill, Gareth W. Young, Nicholas Johnson, Emin Zerman, and Aljosa Smolic. 2020. 'Mixed Reality and Volumetric Video in Cultural Heritage: Expert Opinions on Augmented and Virtual Reality'. In Culture and Computing, 195–214. Springer, Cham. https://doi.org/10.1007/978-3-030-50267-6_16.

[14] O'Dwyer, Néill, Nicholas Johnson, Enda Bates, Raphael Pagés, Jan Ondřej, Konstantinos Amplianitis, David Monaghan, and Aljosa Smolić. 2017. 'Virtual Play in Free-Viewpoint Video: Reinterpreting Samuel Beckett for Virtual Reality'. In 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), 262–67. https://doi.org/10.1109/ISMAR-Adjunct.2017.87.

[15] O'Dwyer, Néill, Nicholas Johnson, Rafael Pagés, Jan Ondřej, Konstantinos Amplianitis, Enda Bates, David Monaghan, and Aljoša Smolić. 2018. 'Beckett in VR: Exploring Narrative Using Free Viewpoint Video'. In ACM SIGGRAPH 2018 Posters on - SIGGRAPH '18, 1–2. Vancouver, British Columbia, Canada: ACM Press. https://doi.org/10.1145/3230744.3230774.

[16] O'Dwyer, N., Ondřej, J., Amplianitis, K., & Smolić, A. (2018). Jonathan Swift: Augmented reality application for Trinity library's long room. In International Conference on Interactive Digital Storytelling (pp. 348-351). Springer, Cham.

[17] Zerman, E., O'Dwyer, N., Young, G. W., & Smolic, A. (2020). A case study on the use of volumetric video in augmented reality for cultural heritage. In Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (pp. 1-5).

[18] 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.

[19] 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.

[20] https://v-sense.scss.tcd.ie/creative-experiments/image-technology-echoes/

[21] O'Dwyer, Neill, Gareth W. Young, Aljosa Smolic, Matthew Moynihan, and Paul O'Hanrahan. 2021. 'Mixed Reality Ulysses'. In SIGGRAPH Asia 2021 Art Gallery, 1. SA '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3476123.3487880.

[22] Joyce, James. 1922. Ulysses. Paris: Shakespeare and Company.

[23] O'Dwyer, Néill, Gareth W. Young, and Aljosa Smolic. 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 0 (0): 1–19. https://doi.org/10.1080/14794713.2022.2031801.

[24] Young, G. W., O'Dwyer, N., Moynihan, M., & Smolic, A. (2022). Audience Experiences of a Volumetric Virtual Reality Music Video. In 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 775-781). IEEE.

[25] Weng, Chung-Yi and Curless, Brian and Srinivasan, Pratul P. and Barron, Jonathan T. and Kemelmacher-Shlizerman, Ira, HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video, CVPR 2022.