User Behaviour Analysis of Volumetric Video in Augmented Reality

Mr. Arunkumar Joshi[1], Mr. Vikram Shirol[2], Ms. Aishwarya K[3]
Smt Kamala and Shri Venkappa M. Agadi College of Engineering and Technology, Laxmeshwar, India

Abstract: *Volumetric video is regarded worldwide as the next important development step in the field of media production. Especially in the context of the extremely rapid development of the Virtual Reality (VR) and Augmented Reality (AR) markets, volumetric video is becoming a key technology. In this, a new capture and processing system for volumetric video is presented, called 3D Human Body Reconstruction (3DHBR). The system is based on 16 stereo pairs of high-resolution cameras capturing a moving person in 360degree. A novel stereo approach provides depth information from all perspectives, which is then fused to a single consistent 3D point cloud. A meshing and mesh reduction algorithm finally produces a sequence of meshes that can be integrated into common render engines. Given that, an integration of realistic dynamic 3D reconstructions of moving persons in VR and AR applications is possible.*

Keywords: Virtual Reality

I. INTRODUCTION

Thanks to new head mounted displays (HMD) for virtual reality, such as Oculus Rift and HTC Vive, the creation of fully immersive environments has gained a tremendous push.

In addition, new augmented reality glasses and mobile devices reach the market that allow for novel mixed reality experiences. With the ARKit by Apple and ARCore for Android, mobile devices are capable of registering their environment and put CGI objects at fixed positions in viewing space. Beside the entertainment industry, many other application domains have potential for immersive experiences based on virtual and augmented reality.

In the industry sector, virtual prototyping, planning, and e-learning benefit significantly from this technology. VR and AR experiences in architecture, construction, chemistry, environmental studies, energy and edutainment offer new applications. Cultural heritage sites, which have been destroyed recently, can be experienced again. Finally yet importantly, therapy and rehabilitation are other important applications. For all these application domains, a realistic and lively representation of human beings is desired.

However, current character animation techniques do not offer the necessary level of realism. The motion capture process is time consuming and cannot represent all detailed motions of an actor, especially facial expressions and the motion of clothes. This can be achieved with Volumetric Video. The main idea is to capture an actor with multiple cameras from all directions and to create a dynamic 3D model.

II. WHAT IS ARGUMENTED REALITY?

The process of superimposing digitally rendered images onto our real-world surroundings, giving a sense of an illusion or virtual reality. Recent developments have made this technology accessible using a Smartphone.

2.1 What is Volumetric Vedio?

Volumetric video (VV) is a new technique used to generate content for augmented reality (AR) and virtual reality (VR) applications. The VV algorithms build a dynamic 3D representation using real-life video from cameras surrounding the 3D object or scene. The generated VVs are represented as point clouds or 3D textured mesh sequences and can be looked at from any viewpoint.

The VVs are used in different marker-based or markerless AR applications.

In almost all cases, these applications use VVs that are stored in the device; however, there is a growing interest in VV compression and adaptive streaming, as real-time streaming is necessary for some applications, e.g., telepresence and remotecollaboration. Understanding how users interact with the VV in AR or VR is critical for optimising compression

DOI: 10.48175/568

and adaptive streaming methods, such as viewport prediction or rate-distortion or rate-utility estimations based on users' preferred distance. User behaviour has been studied and found critical in viewport prediction for another immersive video technology: 360-degree video. For volumetric media, there are only a handful of studies that focus on understanding user interaction, AR viewport prediction, and navigation.A novel integrated multi-camera and lighting system for full 360-degree acquisition of persons has been developed. It consists of a metal truss system forming a cylinder of 6m diameter and 4m height. On this system, 32 cameras are arranged in 16 stereo pairs and equally distributed at the cylindrical plane in order to capture full 360-degree volumetric video.In Fig.1, left, the construction drawing of the studio is presented. 120 Kino Flo LED panels are tissue is covering the inside to provide diffuse lighting from any direction and automatic keying. The avoidance of green screen and provision of diffuse lighting from all directions offers best possible conditions for re-lighting of the dynamic 3D models afterwards at design stage of the VR experience. This combination of integrated lighting and background is unique. All other currently existing volumetric video studios use green screen and directed light from discrete directions.

The system completely relies on a vision-based stereo approach for multi-view 3D reconstruction and omits separate active 3D sensors. The cameras are equipped with a high-quality 20 MPixel sensor at 30 frames per second. This is another key difference compared to other existing capture systems. The overall ultra-high resolution video information from all cameras leads to a challenging amount of data, resulting in 1.6 TB per minute.The avoidance of green screen and provision of diffuse lighting from all directions offers best possible conditions for re-lighting of the dynamic 3D models afterwards at design stage of the VR experience. The overall ultra-high resolution video information from all cameras leads to a challenging amount of data, resulting in 1.6 TB per minute.

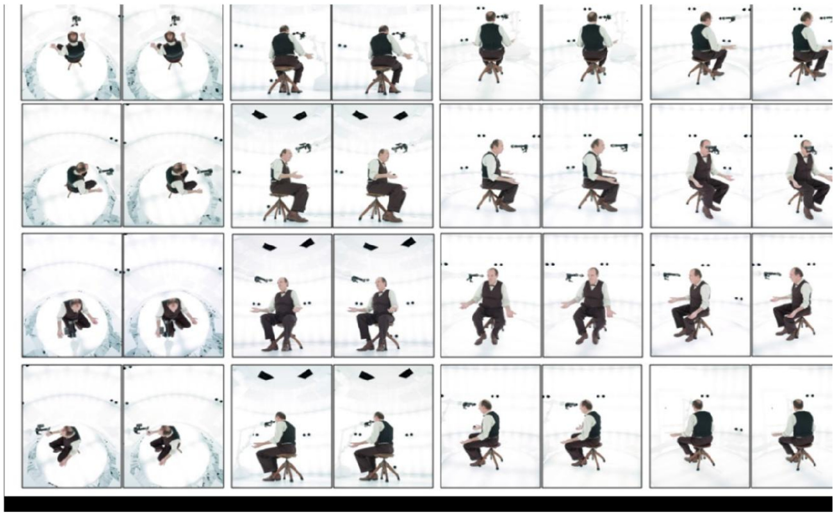

Fig 1 a sample view of all 32 cameras

A sample view of all 32 cameras is presented that represents our solution for the multi-dimensional optimization problem. Four pairs are mounted on the ceiling and on the bottom, while eight pairs are distributed equally at middle height in the cylinder.

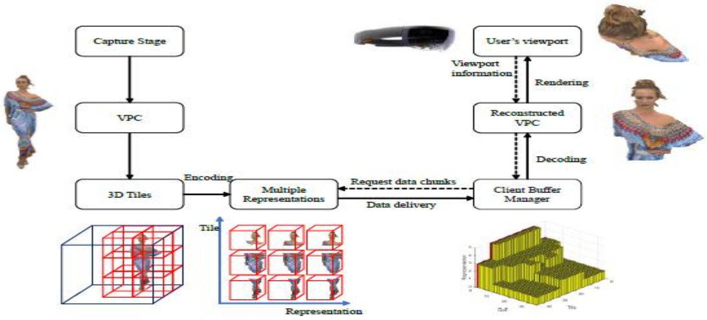

Fig 2 Volumetric Media Streaming System

An overview of our system for streaming volumetric media is shown in Fig.2 . On the left, a web server stores the volumetric media objects, or holograms, for streaming. Each object is represented as a sequence of voxelized point cloud (VPC) [19] frames, the frames are grouped into groups of frames (GOFs), and the sequence of GOFs is divided temporally into segments. Each segment is independently compressed to a small set of representations, each representation at a different bitrate. The bitrates are constant across all segments. Each GOF is divided spatially into tiles, which are independently coded. A media presentation description (MPD), or manifest, accompanies each object to describe the collection of representations, while a segment index accompanies each segment to index the collection of tiles within the segment. On the right, the client buffer manager (CBM) in an HTTP client downloads the manifest, downloads the segment index for each segment in the window, estimates the throughput, estimates the user position, calculates the optimal collection of 3D tiles that will maximize utility for a given rate constraint, requests the tiles, downloads them into the buffer, advances the window, releases the tiles that fall out of the window for rendering, decoding, and presentation by the application, and repeats.

2.2 Processing of Volumetric Video

The complete volumetric video workflow is described, consisting of pre-processing, stereo depth estimation, point-cloud fusion, meshing and mesh reduction.

A. Pre-Processing

In the first step, a pre-processing of the multi-view input is performed. It consists of a color matching to guarantee consistent colors in all camera views.This has significant impact on stereo depth estimation, but even more important, it improves the overall texture during the final texturing of the 3D object.In addition, color grading can be applied as well to match the colors of the object with artistic and creative expectations. E.g. colors of shirts can be further manipulated to get a different look.The segmentation approach is a combination of difference and depth keying supported by the active background lighting.

The next step is stereo depth estimation. As mentioned in the previous section, the cameras are arranged in stereo pairs that are equally distributed in the cylinder. These stereo base systems offer the relevant 3D information from their viewing direction.A stereo video approach is applied that is based on the IPSweep algorithm.This stereo processing approach consists of an iterative algorithmic structure that compares projections of 3D patches from left to right image using point transfer via homography mapping as defined in Equ. (1) and (2) and illustrated in Fig.3.Since a fully calibrated camera system is available, the point transfer between projections of a plane in space in two images can be defined via the homography transformation as defined in Equ. (2).The matrices **K**, **R** and vector **t** relate to the intrinsic camera matrix, the rotation and the translation between camera **C1** and **C2**.

The position of the related 3D patch along the optical ray can be defined via parameter ZP and thecorresponding orientation is defined via the patch normal ***n***.

B. Point Cloud Fusion

For each 2D depth map, initial patches of neighbored 2D points can be calculated straight away including normal information for each 3D point.The resulting 3D information from all stereo pairs is then fused with a visibility-driven patch-group generation algorithm. In brief, all 3D points occluding any other depth map are filtered out resulting in an advanced foreground segmentation. given through the application of fusion rules that are based on an optimized visibility driven outlier removal, and the fusion taking place in both, the 2D image domain as well as the 3D point cloud domain.

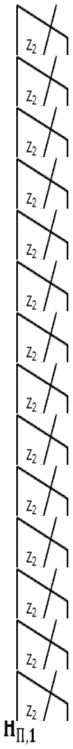

Fig 3: Design of point cloud fusion

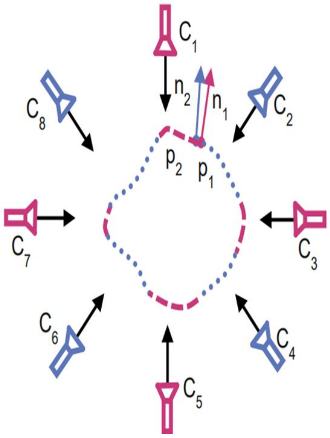

Fig.4: Selection of appropriate cameras by comparing camera orientation and patch normal.

- 3D points survive the fusion process, which are facing the related camera with least angle deviation.
- Due to the high-resolution original images, the resulting 3D point cloud per frame is in a range of several 10s of millions of 3D points.
- In order to match with common render engines, the 3D point cloud needs to be converted to a single consistent mesh.

III. MESHING AND MESH REDUCTION

A geometry simplification is performed that involves two parts.In a first step, a screened Poisson Surface Reconstruction (SPSR) is applied. SPSR efficiently meshes the oriented points calculated by our patch fusion and initially reduces the geometric complexity to a significant extent. Secondly, the resulting mesh is elementally trimmed and cleaned based on the sampling density values of each vertex obtained by SPSR. We ensure the preservation of mesh topology and boundaries in order to improve the quality of the simplified meshes. Depending on the target device, a different mesh resolution is necessary in order to match with the rendering and memory capabilities.

IV. EXPERIMENTAL RESULT

In this section, results from the 360-degree volumetric video production together with UFA GmbH are presented. The volumetric assets are part of the VR Experience “A whole life”, which is continuously presented at the Film Museum Berlin. Two actors have been captured separately in the new Volumetric Video Studio. The resulting dynamic 3D models have then been integrated in a joint scene performing a dialogue. The separate capture led to some challenges for the actors as they had to speak during the other performer’s breaks. The raw data consisted of 25 TB, which have then been processed with the 3D workflow presented in Sec.3.

The processing is performed on a local cloud system resulting in a final sequence of texturized meshes. The overall processing is about 50 sec/frame. The resulting quality of the meshes is achieved fully automatically without manual post-processing of individual meshes.

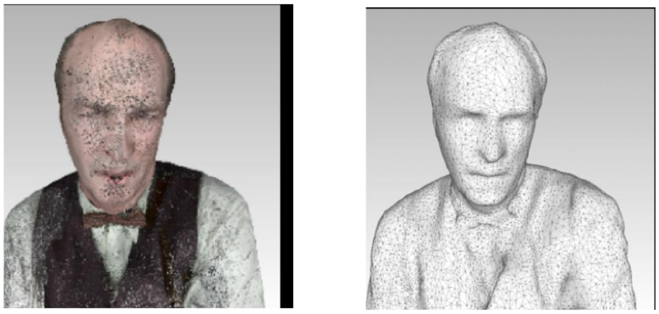

Fig 5: Example for a resulting depth map by one of the 16 stereo systems (top) and close-up of fused point cloud with 20M 3D points (middle) and simplified mesh with 70k faces (bottom)

In Fig.5 (middle), a closeup of the resulting point cloud is shown. The fusion approach leads to a very detailed point cloud of about 20M 3D points. The resulting dynamic 3D models have then been integrated in a joint scene performing a dialogue. The separate capture led to some challenges for the actors as they had to speak during the other performer’s breaks. The raw data consisted of 25 TB, which have then been processed with the 3D workflow presented in Sec.3.

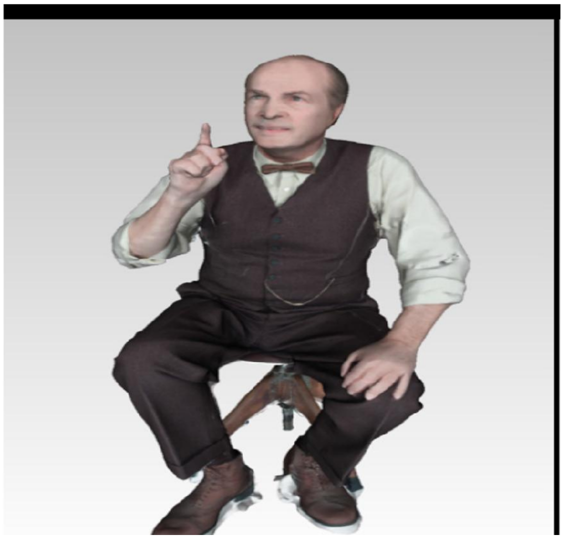

Fig.6: Final meshes of a sequence.

In Fig.4.2, several final texturized meshes are presented. For AR applications, the resulting mesh needs to be reduced down to 20k faces to render a sequence of about 40 sec.

In Fig.7, an example for integration in AR is shown. A Google Pixel smart phone is used together with the provided ARCore to register the device with the environment. The integrated mesh is then rendered on a horizontal plane that has been detected by the device. The dynamic mesh can then be rotated interactively and viewed from any direction.
such as information in the manifest to support the notion of a volume and its coordinate system relative to the user, as well as tiles and segment indexes for the tiles. Simulation results show that even in the absence of tiles and rate-utility optimization, our algorithm has both higher throughput and fewer rebuffering events than existing algorithms, in both stable and variable network conditions. Simulation results also show that our algorithm handles user interaction for volumetric video as expected. The integrated mesh is then rendered on a horizontal plane that has been detected by the device. The dynamic mesh can then be rotated interactively and viewed from any direction.

V. CONCLUSION

Virtual and augmented reality have contributed to the planning of maxillofacial procedures and surgery training. Few articles highlighted the importance of this technology in improving the quality of patients' care. There are limited prospective randomized studies comparing the impact of virtual reality with the standard methods in delivering oral surgery education.

A novel integrated capture and lighting system has been presented for the production of 360degree volumetric video. Furthermore, the complete multi-view 3D processing chain has been explained that leads to high quality sequences of meshes in terms of geometrical detail and texture quality.

The overall processing time is rather low compared to other approaches. This is achieved by an efficient algorithmic workflow using stereo processing and smart fusion of 3D information as well as by a parallel algorithmic structure and exploitation of GPU capabilities.

The final meshes can be integrated in VR and AR applications offering highly realistic representations of human beings. The main new aspect of streaming volumetric content for VR/AR applications, compared to streaming video (or even spherical video), is the high amount of user interactivity. This requires the streaming to be not only network-adaptive but also user adaptive. Thus, new approaches are needed to minimize both network bandwidth and perceived latency due to user interactions. Our solution to volumetric streaming is consistent with modern HTTP streaming and requires only minor modifications to, for example, DASH, such as information in the manifest to support the notion of a volume and its coordinate system relative to the user, as well as tiles and segment indexes for the tiles. Simulation results show that even in the absence of tiles and rate-utility optimization, our algorithm has both higher throughput and fewer rebuffering events than existing algorithms, in both stable and variable network conditions. Simulation results also show that our algorithm handles user interaction for volumetric video as expected.

REFERENCES

[1]. https://www.microsoft.com/en-us/mixed-reality/capturestudios

[2]. https://8i.com/

[3]. http://uncorporeal.com/

[4]. http://www.4dviews.com

[5]. A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, S. Sullivan, “High-quality streamable free-viewpoint video”.

[6]. V. Leroy, J.-S. Franco, E. Boyer, “Multi-View Dynamic ShapeRefinement Using Local Temporal Integration”. IEEE,International Conference on Computer Vision 2017, Oct 2017

[7]. N. Robertini, D. Casas, E. de Aguiar and C. Theobalt, “MultiviewPerformance Capture of Surface Details”. Int. Journal ofComputer Vision (IJCV) 2017

[8]. D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popovic, S.Rusinkiewicz, “Dynamic shape capture using multi-viewphotometric stereo. ACM Transactions on Graphics, 28(5),174.

[9]. W. Waizenegger, I Feldmann, O. Schreer, “Real-time PatchSweeping for High-Quality Depth Estimation in 3DVideoconferencing Applications,“SPIE Conf. on Real-TimeImage and Video Processing, San Francisco, USA, (2011).DOI: 10.1117/12.872868

[10]. W. Waizenegger, I. Feldmann, O. Schreer, P. Kauff, P. Eisert:Real-time 3D Body Reconstruction for Immersive TV, Proc.23rd Int. Conf. on Image Processing