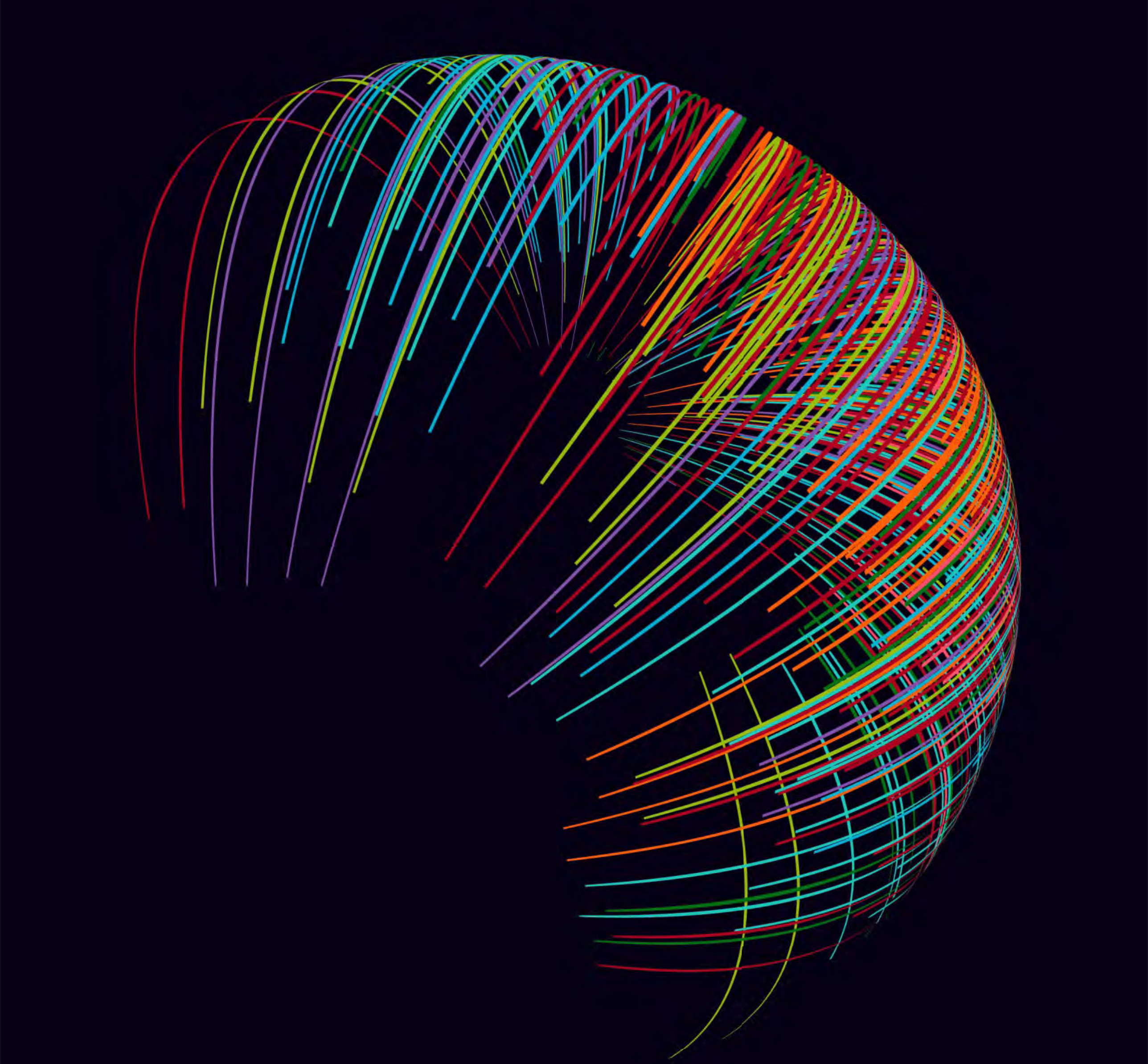
SMPTE

2021

MOTION IMAGING JOURNAL

Covering Emerging Technologies for the Global Media Community



Advanced Volumetric Capture and Processing

By Oliver Schreer, Ingo Feldmann, Thomas Ebner, Sylvain Renault, Christian Weissig, Danny Tatzelt, and Peter Kauff

Introdução:

Eu gosto muito quando novidades dos laboratórios chegam até nós, pobres mortais, nos dando aquele gostinho de ficção científica. E este artigo é um verdadeiro banquete mostrando o futuro que está por vir. O vídeo volumétrico é considerado mundialmente como o passo mais importante no desenvolvimento na produção de mídia, especialmente no contexto dos mercados de Realidade Virtual (VR) e Realidade Aumentada (AR), onde a volumetria do vídeo está tornando-se uma tecnologia-chave. O Fraunhofer Heinrich Hertz Institute (HHI) desenvolveu uma nova tecnologia para volumetria de vídeo, intitulada: *3D Human Body Reconstruction* (3DHBR), que captura pessoas reais e cria modelos 3D dinâmicos que se movem naturalmente, e que podem então ser observados de pontos de vista arbitrários em realidade virtual ou aumentada. O artigo apresenta de forma completa o processo 3D multiview, que traz a alta qualidade de sequência de malhas em termos de detalhes geométricos e de textura. Boa Leitura e bem-vindos ao futuro!

Por: Tom Jones Moreira

Abstract

Volumetric video is regarded worldwide as the next important development in media production, especially in the context of rapidly evolving virtual and augmented reality markets where volumetric video is becoming a key technology. Fraunhofer Heinrich Hertz Institute (HHI) has developed a novel technology for volumetric video. The 3D Human Body Reconstruction (3DHBR) technology captures real persons with our novel volumetric capture system and creates naturally moving dynamic 3D models, which can then be observed from arbitrary viewpoints in a virtual or augmented reality scene. The capture system consists of an integrated multicamera and lighting system for a full 360° acquisition. A cylindrical studio has been developed with a diameter of 6 m and consists of 32 20-MPixel cameras and 120 light-emitting-diode (LED) panels that allow for an arbitrary lit background. Hence, diffuse lighting and automatic keying are supported. The avoidance of green screen and provision of diffuse lighting offers the best possible conditions for relighting of the dynamic 3D models afterward at the design stage of the virtual reality (VR) experience. In contrast to classical character animation, facial expressions and moving clothes are reconstructed at high geometrical detail and texture quality. The complete workflow is fully automatic, requires about 12 hr/min of mesh sequence, and provides a high level of quality for immediate integration in virtual scenes. Meanwhile, a second, professional studio has been built up on the film campus of Potsdam Babelsberg. This studio is operated by VoluCap GmbH, a joint venture between Studio Babelsberg, ARRI, UFA, Interlake, and Fraunhofer Heinrich Hertz Institute (HHI).

The complete workflow is fully automatic, requires about 12 hr/min of mesh sequence, and provides a high level of quality for immediate integration in virtual scenes.

Keywords

Augmented reality (AR), computer vision, mesh, virtual reality (VR), volumetric video

Introduction

Thanks to the availability of new head-mounted displays (HMDs) for virtual reality (VR), such as Oculus Rift and HTC Vive, the creation of fully immersive environments has gained a tremendous push. In addition, new augmented reality (AR) glasses and mobile devices reach the market that allows for novel mixed reality experiences. With the ARKit by Apple and ARCore for Android, mobile devices are capable of registering their environment and putting computer-generated imagery (CGI) objects at fixed positions in viewing space. Besides the entertainment industry, many other application domains see a lot of potential for immersive experiences based on virtual and AR. 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 recently been destroyed, can be experienced again. Finally, yet importantly, therapy and rehabilitation are other important applications, where VR and AR may offer completely new approaches.

For all these application domains and new types of immersive experiences, 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 of the detailed motions of an actor, especially facial expressions and the motion of clothing. This can be achieved with a new technology called *volumetric video*. The main idea is to capture an actor with multiple cameras from all directions and create a dynamic 3D model. There are several companies worldwide offering volumetric capture, such as Microsofts Mixed Reality Capture Studio, [1] 8i,[2] and 4D Views.[3] Compared to these approaches, the presented capture and processing system for volumetric video distinguishes itself in several key aspects, which will be explained in the following sections. Concerning multiview reconstruction, several research groups work in this area. In Reference 4, a spatiotemporal integration is presented for surface reconstruction refinement. The presented approach is based on 68 4MPixel cameras requiring approximately 20 min/frame processing time to achieve

a 3M faces mesh. Robertini et al. present an approach focusing on surface detail refinement based on a prior mesh by maximizing photo-temporal consistency. Vlasic et al.[6] present a dynamic shape capture pipeline using eight 1k cameras and a complex dynamic lighting system that allows for controllable light and acquisition at 240 frames/s. The high-quality processing requires 65 min/frame and a graphics processing unit (GPU)-based implementation with reduced quality achieves 15 min/frame processing time.

Volumetric Capture

A novel integrated multicamera and lighting system for a full 360° acquisition of persons has been developed. It consists of a metal truss system forming a cylinder of 6 m diameter and 4 m 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° volumetric video. In **Fig. 1**, the construction drawing of the volumetric studio is presented.

In addition, 120 light-emitting-diode (LED) panels are mounted outside of the truss system and a semitransparent 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 the best possible conditions for relighting of the dynamic 3D models afterward at the design stage of the VR experience. This combination of integrated lighting and background is unique. All other currently existing volumetric video studios rely on green screen and directed light from discrete directions.

The system relies completely on a vision-based stereo approach for multiview 3D reconstruction and omits separate 3D sensors. The cameras are equipped with a high-quality sensor offering a 20MPixel resolution at 30 frames/s. This is another key difference compared to other existing volumetric video capture systems as this approach benefits from experience in photogrammetry, where high-quality 3D reconstruction can be achieved using ultrahigh-resolution images. The overall ultrahigh-resolution video information from all cameras

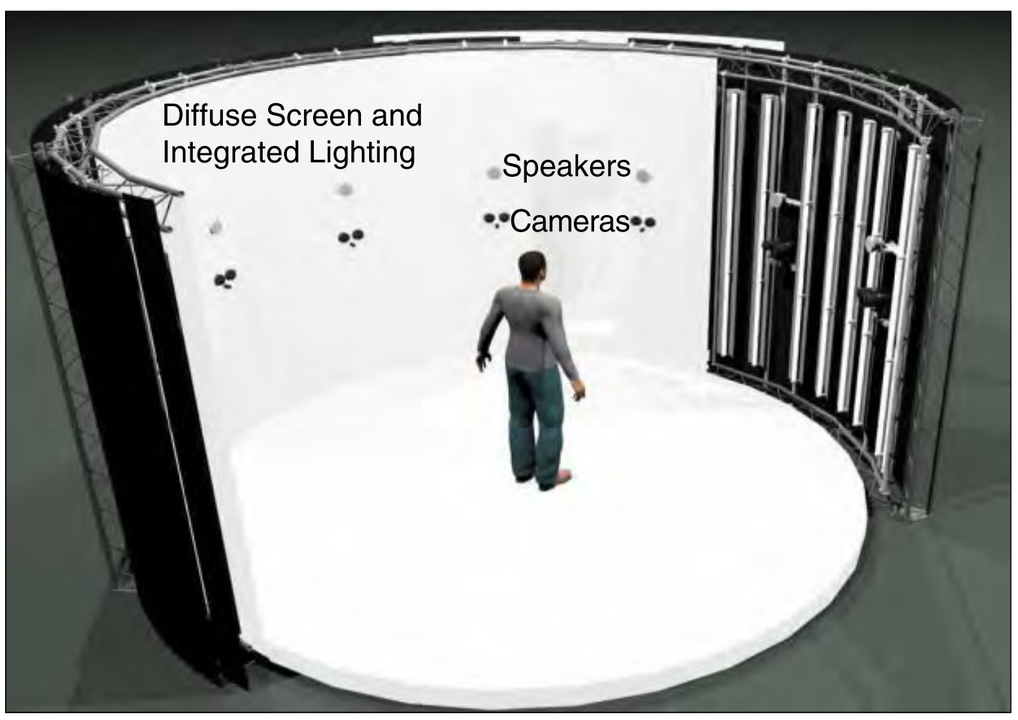


FIGURE 1. Drawing of the capture and light stage.

FIGURE 2. View inside the rotunda during the first test production.

leads to a challenging amount of data, resulting in 1.6 Tbytes/min. In **Fig. 2**, a view inside the rotunda is shown, with an actor sitting in the center.

An important aspect is the number and the distribution of cameras. The objective was to find the best possible camera arrangement with the least possible number of cameras, and, at the same time, the largest possible capture volume with a minimum amount of occlusions had to be achieved. In **Fig. 3**, a sample view of all the 32 cameras is presented that represents our solution for the multidimensional optimization problem.

Processing of Volumetric Video

Preprocessing

In this section, the complete workflow for the processing of volumetric video is described and shown in the workflow diagram shown in **Fig. 4**. In the first step, a preprocessing of the multiview input is performed. It consists of a color matching to guarantee the same color for the same parts of the object in all camera views. This has a significant impact on stereo depth estimation, but even more importantly, it improves the overall texture quality in the point cloud fusion step and 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. For example, colors of shirts can be further manipulated to get a different look. After color matching and grading, the foreground object is segmented from the background to reduce the amount of data to be processed. The segmentation approach is a combination of difference and depth keying.

Stereo Depth Estimation

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 relevant 3D information from their viewing direction. A stereo video approach is applied that is based on the so-called IPSweep algorithm.[7,8] The presented stereo processing approach

FIGURE 3. 32 camera views.

consists of an iterative algorithmic structure that compares projections of 3D patches from left to right image using point transfer via homography mapping.

In contrast to many other approaches that evaluate a fixed disparity range, a set of spatial candidates and a statistically guided update for comparison is used in this algorithm, which significantly speeds up correspondence search. Moreover, the selection of candidates is performed along the optical ray defined by the depth of the candidate related to the first camera. Once all candidates are evaluated for a given similarity measure, the best candidate is selected as final depth candidate. Compared to standard block-matching approaches, spatial 3D patches are projected from left to the right image to cope with perspective distortions. This process is applied for a left-to-right and right-to-left estimation. After that, a consistency check is performed between both depth maps, and a consistency map is produced. This consistency map is used to hinder their propagation in the next iteration and to penalize their selection. The iterative structure of the algorithm allows for the propagation of correct results to the neighborhood by keeping the independence of the processing per pixel. This is significantly important for parallel processing on GPUs, which has been heavily exploited in our case. Moreover, the GPU-centric implementation allows for an inherent subpixel processing due to texture lookups. To summarize, the GPU-based design of the iterative patch sweep has the following important properties:

- correspondence analysis performed on nonrectified stereo configurations
- consideration of projective mapping via homography transfer
- estimation on subpixel level by exploiting floating point texture memory access in graphics memory
- fully parallelized processing per pixel through iterative depth propagation.

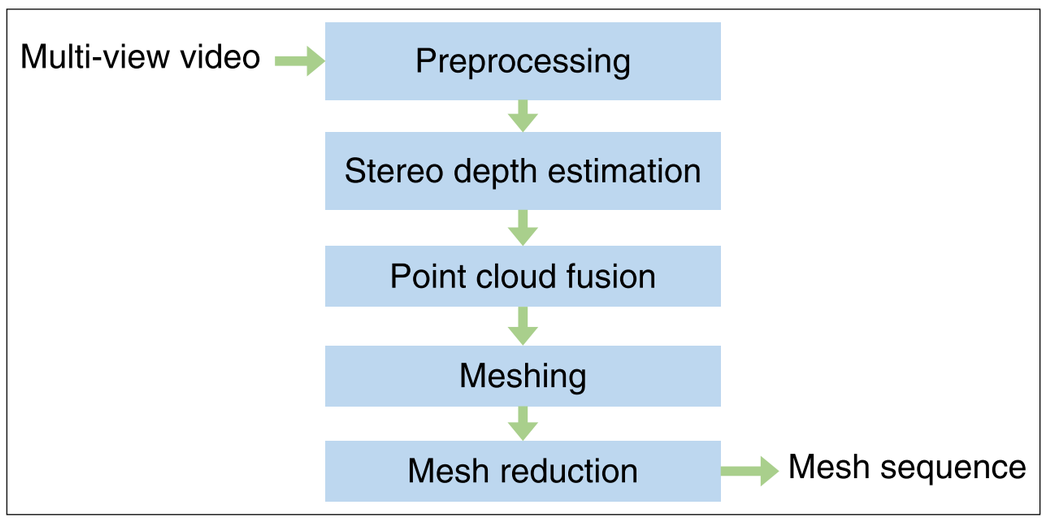


FIGURE 4. Workflow diagram.

Point Cloud Fusion

As an additional result of stereo processing, 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.[9] In brief, all 3D points occluding any other depth maps are filtered out resulting in advanced foreground segmentation. Remaining artifacts have a larger distance to the object to be reconstructed and, as a result, they do not occlude any other depth maps. The efficiency of this approach is 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. Due to the high resolution of original images, the resulting 3D point cloud per frame is in a range of several tens of millions of 3D points.

Meshing and Mesh Reduction

To match the high-resolution 3D point cloud with the performance limits of the state-of-the-art render engines, the 3D point cloud needs to be simplified and converted to a single consistent mesh. Therefore, a geometry simplification is performed that involves two parts. First, a screened Poisson Surface Reconstruction (SPSR) is applied.[10] SPSR efficiently meshes the oriented points calculated by our patch fusion and initially reduces the geometric complexity to a significant extent. In addition, this step generates a watertight mesh. Holes that remained in the surface after the reconstruction due to complete occlusion or data imperfections are closed. Second, the resulting mesh is elementally trimmed and cleaned based on the sampling density values of each vertex obtained by SPSR. In contrast to the common approaches introduced earlier, we do not require an extensive intersection of the resulting surface with the visual hull. Outliers and artifacts are already reliably removed by our patch fusion.

Subsequently, the triangulated surface is simplified even further to a dedicated number of triangles by iterative contraction of edges based on Quadric Error Metrics.[11] Thus, detailed areas of the surface are represented by more triangles than simple regions.

During this stage, we ensure the preservation of mesh topology and boundaries to improve the quality of the simplified meshes. Another important aspect is the possibility of defining the target resolution of meshes. Depending on the target device, a different mesh resolution is necessary to match the rendering and memory capabilities. For a desktop application using Oculus Rift or HTC Vive, a mesh size of 70K faces is appropriate. However, mobile devices such as Google Pixel can render mesh sequences of 20K faces fluently. To recover details lost during simplification, we compute ultraviolet coordinates for each vertex and create a texture of suitable size.[12]

The final sequence of meshes can then be further manipulated in standardized postproduction workflows but also be rendered directly in VR applications, created with 3D engines like Unity3D[13] or Unreal Engine.[14]

Experimental Results

In this section, results from a recent 360° volumetric video production are presented. This production has been performed together with UFA GmbH for their VR experience Entire life, which has been presented for six months at the Film Museum Berlin in the exhibition UFAThe story of a brand. Two actors have been captured separately in the new Volumetric Video Studio as described in the Volumetric Capture section. The resulting dynamic 3D models have then been integrated into a joint scene performing a dialogue. The separate capture led to some challenges for the actors, as they had to speak during the other performers breaks. The raw data consisted of 25 Tbytes, which have then been processed with the 3D workflow presented in the Processing of Volumetric Video section. The processing is performed on a local cloud system resulting in a final sequence of texturized meshes. The overall processing is about 28 frames/s. The resulting quality of the meshes is achieved fully automatically without manual postprocessing of individual meshes.

In **Fig. 5** (left), a resulting depth map by one of the 16 stereo systems is shown. Besides the depth, we also compute the normal of each 3D point, which is then used during our fusion approach as described in the previous section. The fusion approach leads to a very detailed point cloud of about 20M 3D points. The challenge is now to further reduce the mesh complexity being able to render the sequence of meshes in the render engine of the target device. Currently, Unity 3D is used to render the sequence of meshes. In the near future, dynamic rendering in Unreal will be supported as well. The final complexity of the meshes depends on the rendering capabilities of the target device. For HTC Vive or Oculus Rift, a graphics workstation is used, and, for these devices, a mesh resolution of 70K faces is an appropriate compromise in terms of the level of geometric detail and performance. The result of the mesh fusion and simplification process is shown in **Fig. 5** (middle) and (right). In **Fig. 6**, several final texturized meshes are presented. For AR applications, the resulting mesh needs

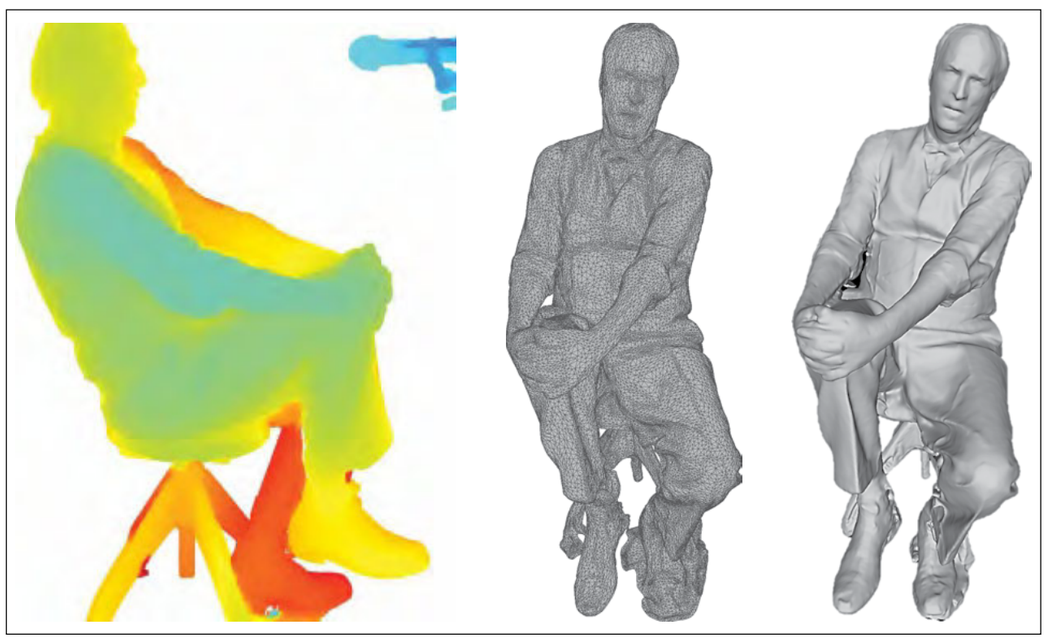


FIGURE 5. Example for a resulting depth map by one of the 16 stereo systems (left), resulting fused mesh (middle), and rendered polygonal 3D model without texture (right).

FIGURE 6. Final meshes of different frames of a sequence.

to be reduced to 20K faces to render a sequence of about 40 sec. In **Fig. 7**, an example of integration in an AR application is shown. A Google Pixel smartphone is used together with the provided ARCore to register the device with the environment. The integrated mesh is positioned on a horizontal plane that is registered by the device with the floor plane of the real-world environment. The dynamic mesh can then be rotated interactively and viewed from any direction. Due to the dynamic registration of the floor plane, the user can virtually walk around the volumetric model in the AR device.

Summary

A novel integrated capture and lighting system has been presented for the production of 360° volumetric video. Furthermore, the complete multiview 3D processing chain has been explained that leads to high-quality sequence of meshes in terms of geometrical detail and texture quality. The overall processing time is rather low compared to other approaches. The main reasons for this are efficient algorithmic workflow using stereo processing and smart fusion of 3D information, a parallel algorithmic structure, and exploitation of GPU capabilities. The final meshes can then be integrated into VR and AR applications offering highly realistic representations of human beings.

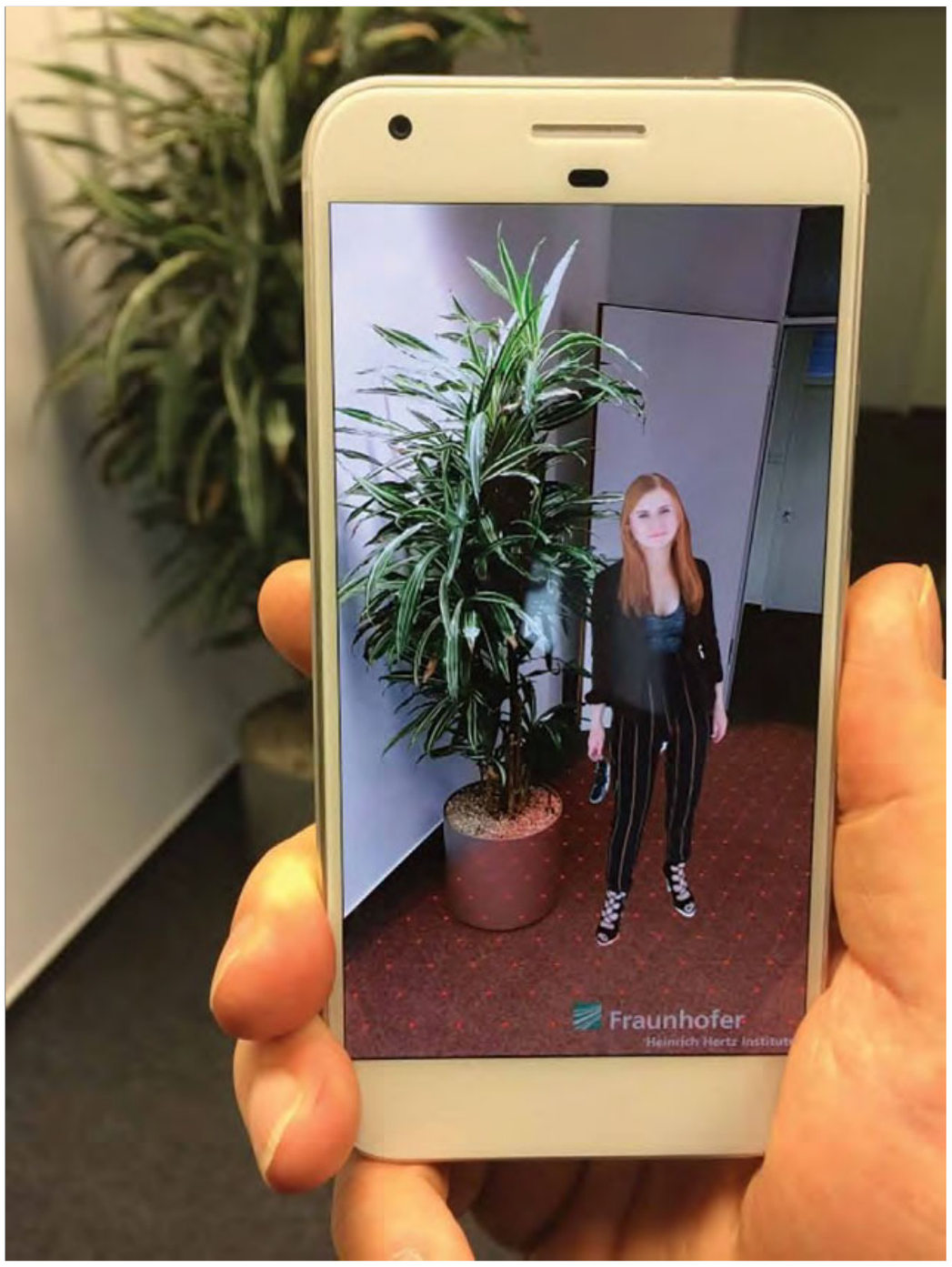


FIGURE 7. Integration of mesh sequence on Google Pixel using the ARCore.

Acknowledgments

We gratefully thank UFA GmbH for the provision of sample data that were created as part of the co-production of Entire life between Fraunhofer Heinrich Hertz Institute (HHI) and UFA.

References

1. Microsoft Mixed-Reality Studio, 2019. [Online]. Available: https://www.microsoft.com/en-us/mixed-reality/capture-studios
2. 8i volumetric Video Studio, 2019. [Online]. Available: https://8i.com/
3. 4D View Solutions. [Online]. Available: http://www.4dviews.com
4. V. Leroy, J.-S. Franco, and E. Boyer, Multi-View Dynamic Shape Refinement Using Local Temporal Integration, *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, Oct. 2017.
5. N. Robertini, D. Casas, E. de Aguiar, and C. Theobalt, Multi-View Performance Capture of Surface Details, *Int. J. Comput. Vision*, Online, Jan. 2017.
6. D. Vlasic et al., Dynamic Shape Capture Using Multi-View Photometric Stereo, *ACM Trans. Graph.*, 28(5):174, Dec. 2009.
7. W. Waizenegger et al., Real-Time Patch Sweeping for High-Quality Depth Estimation in 3D Videoconferencing Applications, *Proc. SPIE Conference on Real-Time Image and Video Processing*, San Francisco, CA, 2011.
8. W. Waizenegger, I. Feldmann, O. Schreer, P. Kauff, and P. Eisert, Real-Time 3D Body Reconstruction for Immersive TV, presented at the 23rd Int. Conf. Image Processing, Pheonix, AZ, Sept. 2016.
9. S. Ebel, W. Waizenegger, M. Reinhardt, O. Schreer, and I. Feldmann, Visibility-Driven Patch Group Generation, presented at the IEEE Int. Conf. 3D Imaging, Liege, Belgium, Dec. 2014.

10. M. Kazhdan and H. Hoppe, Screened Poisson Surface Reconstruction, *ACM Trans. Graph.*, 32(3), June 2013.
11. M. Garland and P. S. Heckbert, Surface Simplification Using Quadric Error Metrics, *Proc. 24th Ann. Conf. on Comp. Graph. and Inter. Techniques SIGGRAPH 97* , pp. 209216, New York, NY, 1997.
12. T. Ebner, I. Feldmann, and S. O. Schreer. 46-2: Distin - guished Paper: Dynamic Real World Objects in Augmented and Virtual Reality Applications, *SID Symposium Digest of Technical Papers*, 48 (1): 673 676, May 2017.
13. Unity Company Website, 2019. [Online]. Available: https://unity.com/
14. Unreal Engine, 2019. [Online]. Available: https://www.unrealengine.com/en-US/

About the Authors

Oliver Schreer is head of the Immersive Media and Communication Group at the Vision and Imaging Technologies Depart ment, Fraunhofer HHI, Berlin, Germany. His main research fields are realtime 3D video processing, future immersive media services exploiting augmented reality (AR) and virtual reality (VR) technologies, 3D video communication, human motion analysis, and touchless interaction. Since 2001, he has been an adjunct professor with the Faculty of Electrical Engineering and Computer Science, Technical University of Berlin He provided a complementary set of lectures titled Stereo Image Processing and View Synthesis. In June 2006, he received a Habilitation degree in the field of computer vision/video communication from the Faculty of Electrical Engineering and Computer Science, Technical University Berlin. Since November 2006, he has been an associate professor (Privatdozent) in the same faculty in the Computer Vision and Remote Sensing Group. He has published more than 100 papers in journals and conferences and edited two books at Wiley & Sons.

Ingo Feldmann is head of the Immersive Media and Communication Group at the Vision and Imaging Technologies Department, Fraunhofer HHI, Berlin, Germany. He received a Dipl.-Ing. degree in electrical engineering from the Technical University of Berlin in 2000. Since September 2000, he has been with Fraunhofer HHI, where he was engaged in various research activities in the field of AR / VR technologies, volumetric scene reconstruction and modeling, digital cinema, multiview projector-camera systems, real - time 3D video conferencing, and future immersive media services. He was involved in several German and European research projects, such as ATTEST, VIRTUE, ITI, TSDK, Prime, Rushes, 3DPresence, FascinatE, SCENE, ActionTV, and BRIDGET. Later, he participated in the MPEG 3DAV Ad Hoc Group. He has been a reviewer for major computer vision conferences and has published more than 50 papers.

Thomas Ebner joined Fraunhofer HHI in 2008, where he is a- sci entist and software developer in the Immersive Media and Communication Group of the Vision and Imaging Technologies Department. As a project manager for volumetric video for VR/AR, he took a leading role in integrating dynamic 3D reconstructions of real persons into immersive experiences and expanding the groups X Reality activities. Since 2018, he has also been the CTO of Volucap, a volumetric capture studio in Potsdam-Babelsberg. He holds a diploma degree in media and computer science from the Technical University of Dresden. He was with the Division of Computer Engineering, Digital Dongseo University Busan, Busan, South Korea, from 2005 to 2006.

Sylvain Renault received a Dipl.-Ing. degree in computer science from the Technical University of Berlin in 1997. He currently works at Fraunhofer HHI in the Interactive Media and Human Fac - tors Department in the research field of computer graphics and novel human-machine interactions. He implemented several 3D realtime frameworks and developed many gesture-based multimodal 3D applications. Since 2015, he has been working in the Vision and Imaging Technologies Department, Fraunhofer HHI, in the research group of Immersive Media and Communication. His current objectives focus on novel concepts and scenarios for user-centered visualization techniques, virtual and augmented reality applications, and the development of human body reconstruction modules. He is also responsible for the implementation of GPU-based stereo drivers for various autostereoscopic and panoramic displays. He realized many prototypes and individual solutions for national and European research projects and for companies in the construction, automotive, medical, cultural-heritage, teleconference, and info tainment fields.

Christian Weissig is head of the Capture and Display Systems Group at the Vision & Imaging Technology Department, Fraunhofer HHI. He received a Dipl.-Ing. degree in information technologyand mechanical engineering from the Technical University of Berlin in 1998.

After working as a visiting scientist, he joined the HHI in 1999. Since then, he has been involved in numerous German and European projects related to novel multimedia production technologies and services in the area of UHD and 3D. He is a member of the Film Technology experts group of the German Society for Information Technology (ITG). Within the HHI, he is responsible for ultrahigh-resolution video panorama solutions, in particular, UHD acquisition and presentation systems.

Danny Tatzelt works as a free lance cinematographer, editor, and director, based in Berlin, Germany. He received a degree in cinematography from the University for Film and Television Konrad Wolf Potsdam-Babelsberg in 2014 with the documentary VIAJAR, and completed a masters schol ars program in 2016 with a 360° film Berlin. He received the Studienstiftung des deutschen Volkes scholarship. In addition to classic film works, he also specializes in VR and AR film productions. He designed the Volumetric Capture Studio in close collaboration with the Fraunhofer HHI, which resulted in the founding of Volucap GmbH, the first commercial volumetric video studio in Europe.

Peter Kauff is one of the two heads of the Vision and Imaging Technologies Department, Fraunhofer HHI, Berlin, Germany. He is also co-head of the Capture and Display Systems Group in his department and a co-founder of the Tomorrows Immersive Media Experience Laboratory (TiME Lab) at Fraunhofer HHI. He received a Dipl.-Ing. degree in electrical engineering and telecommunication from the Technical University of Aachen, Aachen, Germany, in 1984. He has been with Fraunhofer HHI since then. He was involved in numerous German and European projects related to advanced 3D cinema, 3D television, telepresence, and immersive media. Currently, his department consists of three research groups with about 30 researchers who manage up to 40 research and development (R&D) projects and industrial contracts. Main topics of the ongoing R&D work are 3D image and video analysis, immersive media, multicamera for 3D capturing, and immersive 2D/3D multiprojection, as well as augmented, virtual, and mixed reality.