# ADVANCED VOLUMETRIC CAPTURE AND PROCESSING

O. Schreer, I. Feldmann, T. Ebner, S. Renault, C. Weissig, D. Tatzelt, P. Kauff

Fraunhofer Heinrich Hertz Institute, Berlin, Germany

## ABSTRACT

Volumetric video is regarded worldwide as the next important development step in media production. Especially in the context of rapidly evolving Virtual and Augmented Reality markets, volumetric video is becoming a key technology. Fraunhofer HHI has developed a novel technology for volumetric video: 3D Human Body Reconstruction (3DHBR). The 3D Human Body Reconstruction 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 multi-camera and lighting system for full 360 degree acquisition. A cylindrical studio has been developed with a diameter of 6m and it consists of 32 20MPixel cameras and 120 LED panels that allow for arbitrary lit background. Hence, diffuse lighting and automatic keying is supported. The avoidance of green screen and provision of diffuse lighting offers best possible conditions for re-lighting of the dynamic 3D models afterwards at design stage of the 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 hours per minute 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 HHI.

## INTRODUCTION

Thanks to the availability of 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 see a lot of 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, 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 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 to create a dynamic 3D model of it. There are several companies worldwide offering volumetric capture, such as Microsoft with its Mixed Reality Capture Studio [1], 8i [2], Uncorporeal Systems [3] and 4D Views [4]. Compared to these approaches, the presented capture and processing system for volumetric video distinguishes in several key aspects, which will be explained in the next sections. Concerning multi-view video 3D reconstruction, several research groups work in this area. In [5], a spatio-temporal integration is presented for surface reconstruction refinement. The presented approach is based on 68 4Mpixel Cameras requiring approx. 20 min/frame processing time to achieve a 3M faces mesh. Robertini et al. present an approach focusing on surface detail refinement based on prior mesh by maximizing photo-temporal consistency [6]. Vlasic et al. [7] present a dynamic shape capture pipeline using eight 1k cameras and a complex dynamic lighting system that allow for controllable light and acquisition at 240 frames/sec. The high-quality processing requires 65 min/frame and a GPU based implementation with reduced quality achieves 15 min/frame processing time.

In the next section, the volumetric capture system is presented with its main feature of a combined capturing and lighting approach. After that, the underlying multi-view video processing workflow is presented. Finally, some results of the most recent productions are presented. The paper concludes with a summary.

## VOLUMETRIC CAPTURE

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, the construction drawing of the volumetric studio is presented.

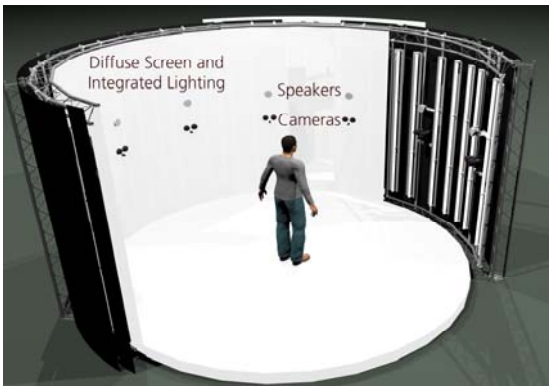


Figure 1 – Drawing of the capture and light stage

In addition, 120 LED panels are mounted outside of the truss system and a semi-transparent 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 virtual reality 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 multi-view 3D reconstruction and omits separate 3D sensors. The cameras are equipped with a high-quality sensor offering 20 MPixel resolution at 30 frames per second. 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 ultra-high resolution images. The overall ultra-high resolution video information from all cameras lead to a challenging amount of data, resulting in 1.6 TB per minute. In Fig. 2, a view inside the rotunda is shown, with an actor sitting in the centre.

An important aspect is the number and distribution of cameras. The objective was to find the best possible camera arrangement with the least possible number of cameras,

Figure 2 – View inside the rotunda during the first test production

Figure 3 – 32 camera views

whereas, at the same time, the largest possible capture volume with 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 multi-dimensional optimization problem.

## PROCESSING OF VOLUMETRIC VIDEO

### Pre-processing

In this section, the complete workflow for the processing of volumetric video is described and shown in the workflow diagram in Fig. 4. In the first step, a pre-processing of the multi-view input is performed. It consists of a colour matching to guarantee same colour for same parts of the object in all camera views. This has significant impact on stereo depth estimation, but even more important, it improves the overall

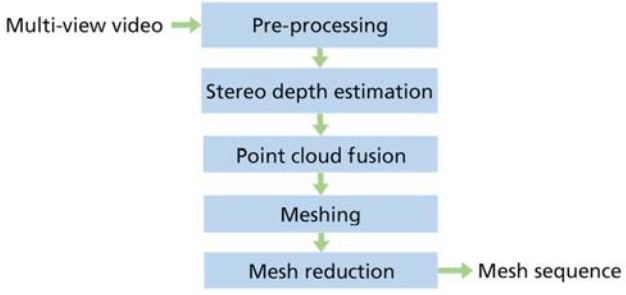


Figure 4 – Workflow diagram

texture quality in the point cloud fusion step and the final texturing of the 3D object. In addition, colour grading can be applied as well to match the colours of the object with artistic and creative expectations. E.g., colours of shirts can be further manipulated to get a different look. After colour matching and grading, the foreground object is segmented from background in order 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 the relevant 3D information from their viewing direction. A stereo video approach is applied that is based on the so-called IPSweep algorithm [8][9]. The presented 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.

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 in order 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 propagation of correct results to the neighborhood by keeping independence of the processing per pixel. This is significantly important for parallel processing on Graphics Processing Units (GPU), which has been heavily exploited in our case. Moreover, the GPU-centric implementation allows for an inherent sub-pixel 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 non-rectified stereo configurations;
- consideration of projective mapping via homography transfer;
- estimation on sub-pixel level by exploiting floating point texture memory access in graphics memory;
- fully parallelized processing per pixel through iterative depth propagation.

**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 [10]. In brief, all 3D points occluding any other depth maps are filtered out resulting in advanced foreground segmentation. Remaining artefacts have a bigger 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 10s of millions of 3D points.

### Meshing and mesh reduction

In order to match the high-resolution 3D point cloud with the performance limits of 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: In a first step, a screened Poisson Surface Reconstruction (SPSR) is applied [11]. 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. Secondly, 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 [12]. 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 in order to improve the quality of the simplified meshes. Another important aspect is the possibility to define the target resolution of meshes. Depending on the target device, a different mesh resolution is necessary in order to match with 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 UV coordinates for each vertex and create a texture of suitable size [13].

The final sequence of meshes can then be further manipulated in standardized post-production workflows, but also be rendered directly in virtual reality applications, created with 3D engines like Unity3D [14] or Unreal Engine [15].

## EXPERIMENTAL RESULTS

In this section, results from a recent 360-degree volumetric video production are presented. This production has been performed together with UFA GmbH for their VR Experience “A whole life”, which has been presented for six months at the Film Museum Berlin in the exhibition “UFA – The story of a brand”. Two actors have been captured separately in the new Volumetric Video Studio as described in Sec. II. 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. III. 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.

In Fig. 5 (left), a resulting depth map by one of the 16 stereo systems is shown. Beside 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 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 to be reduced down to 20k faces to render a sequence of about 40 sec. In Fig.7, an example for integration in an AR application 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 positioned on a horizontal plane, which 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.

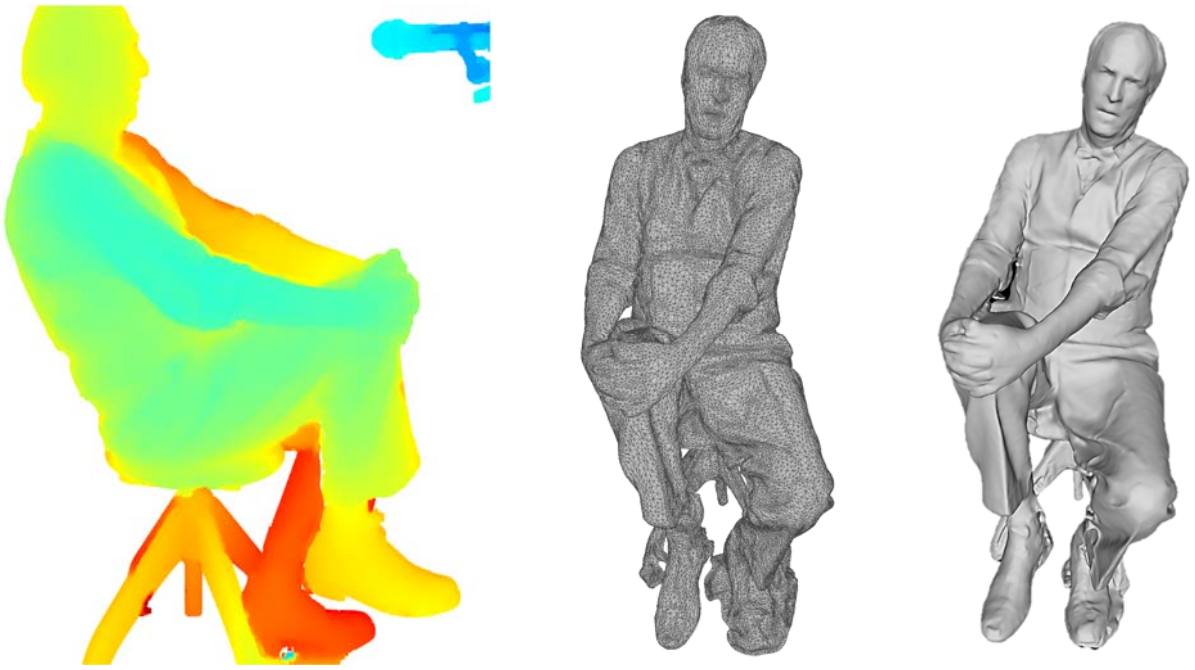

Figure 5 – Left) Example for a resulting depth map by one of the 16 stereo systems, middle) resulting fused mesh, right) rendered polygonal 3d model without texture

Figure 6 – Final meshes of different frames of a sequence


Figure 7 – Integration of mesh sequence on Google Pixel using the ARCore

## SUMMARY

A novel integrated capture and lighting system has been presented for the production of 360 degree volumetric video. Furthermore, the complete multi-view 3D processing chain has been explained that lead 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 in VR and AR applications offering highly realistic representations of human beings.

## ACKNOWLEDGMENT

We gratefully thank UFA GmbH for the provision of sample data that were created as part of the co-production of “A whole life” between Fraunhofer HHI and UFA.

## REFERENCES


[1] https://www.microsoft.com/en-us/mixed-reality/capture-studios
[2] https://8i.com/.
[3] http://uncorporeal.com/.
[4] 4D View Solutions, http://www.4dviews.com
[5] V. Leroy, J.-S. Franco, E. Boyer, “Multi-View Dynamic Shape Refinement Using Local Temporal Integration”. IEEE, International Conference on Computer Vision 2017, Oct 2017
[6] N. Robertini, D. Casas, E. de Aguiar and C. Theobalt, “Multi-view Performance Capture of Surface Details”. Int. Journal of Computer Vision (IJCV) 2017
[7] Vlasic, D., Peers, P., Baran, I., Debevec, P., Popovic, J., Rusinkiewicz, S., et al., “Dynamic shape capture using multi-view photometric stereo. ACM Transactions on Graphics, 28(5), 174.
[8] W. Waizenegger et al., “Real-time Patch Sweeping for High-Quality Depth Estimation in 3D Videoconferencing Applications“ SPIE Conf. on Real-Time Image and Video Processing, San Francisco, USA, (2011). DOI: 10.1117/12.872868
[9] 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 (ICIP 2016), Phoenix, Arizona, USA, September 25-28, 2016.
[10] S. Ebel, W. Waizenegger, M. Reinhardt, O. Schreer, I. Feldmann: Visibility-driven Patch Group Generation, IEEE Int. Conf. on 3D Imaging (IC3D), Liege, Belgium, December 2014, Best Paper Award.
[11] M. Kazhdan, H. Hoppe, “Screened Poisson Surface Reconstruction,” ACM Transactions on Graphics (TOG) 32, No. 3, (2013). DOI: 10.1145/2487228.2487237
[12] M. Garland, P. S. Heckbert, “Surface simplification using quadric error metrics,” SIGGRAPH '97, Proc. of the 24th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., New York, USA, 209-216 (1997) DOI: 10.1145/258734.258849
[13] T. Ebner, I. Feldmann, S., O. Schreer: 46-2: Distinguished Paper: Dynamic Real World Objects in Augmented and Virtual Reality Applications, SID Symposium Digest of Technical Papers, Los Angeles, USA, vol. 48, no. 1, pp. 673–676, May 2017, Distinguished Paper Award. DOI: 10.1002/sdtp.11726.
[14] https://www.unity3d.com/.
[15] https://www.unrealengine.com/.