EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2020)
F.-J. Melero and M. Spagnuolo (Editors)

Short Paper

Ernst Grube: A Contemporary Witness and His Memories Preserved with Volumetric Video

M. Worchel[1], M. Zepp[1], W. Hu[1], O. Schreer[1], I. Feldmann[1] and P. Eisert[1,2]

[1]Fraunhofer Heinrich Hertz Institute, Germany
[2]Humboldt University of Berlin, Germany

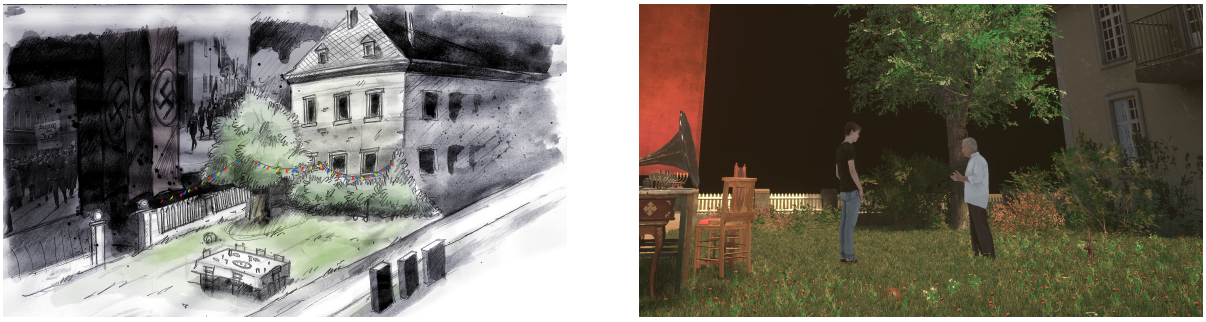


Figure 1: *VR documentary "Ernst Grube – The Legacy". Left: Concept art of one historical site. Right: Virtually recreated scene with volumetric reconstructions of the Holocaust survivor Ernst Grube and an interviewer.*

Abstract

"Ernst Grube – The Legacy" is an immersive Virtual Reality documentary about the life of Ernst Grube, one of the last German Holocaust survivors. From interviews conducted inside a volumetric capture studio, dynamic full-body reconstructions of both, the contemporary witness and its interviewer, are recovered. The documentary places them in virtual recreations of historical sites and viewers experience the interviews with unconstrained motion. As a step towards the documentary's production, prior work presents reconstruction results for one interview. However, the quality is unsatisfying and does not meet the requirements of the historical context. In this paper, we take the next step and revise the used volumetric reconstruction pipeline. We show that our improvements to depth estimation and a new depth map fusion method lead to a more robust reconstruction process and that our revised pipeline produces high-quality volumetric assets. By integrating one of our assets into a virtual scene, we provide a first impression of the documentary's look and the convincing appearance of protagonists in the virtual environment.

CCS Concepts

• ***Human-centered computing*** → ***Virtual reality;*** • ***Computing methodologies*** → ***Reconstruction;*** *Rendering;*

1. Introduction

75 years ago, the most violent episode in European history ended with the defeat of Nazi Germany. With every passing year, fewer survivors of this dark period remain. Preserving their memories, stories and impressions is one of the most important cultural heritage missions today. While photographs, audio recordings or videos can capture the historical content, they fail to accurately depict a witness *as a whole*. Re-experiencing their presence or even standing next to them during the narration could drastically increase the depth of their stories. Recent advances in computer vision and media presentation allow us to close this gap: volumetric capture systems can acquire a dynamic fully-body human reconstruction. This *volumetric video* of a person can be directly integrated into (interactive) Virtual Reality (VR) or Augmented Reality (AR) experiences in which a user is immersed through a head-mounted display.

The VR documentary "Ernst Grube – The Legacy" [FGG*20] (see Figure 1) uses volumetric video to retell the stories of Ernst Grube, one of the last German survivors of the Holocaust. It consists of six interviews, in which the Jewish contemporary witness reflects on his life under the Nazi regime, from his youth to his imprisonment in the Theresienstadt concentration camp, as well as his

DOI: 10.2312/gch.20201300

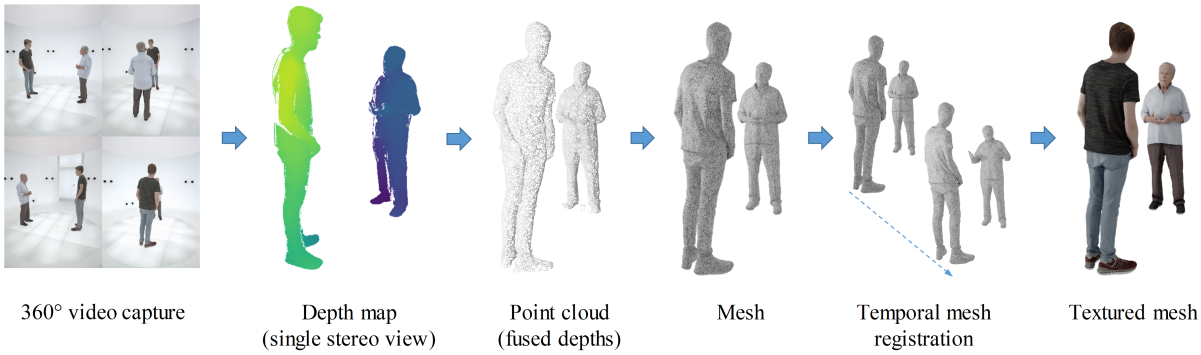


Figure 2: *Our reconstruction pipeline based on 3DHBR [SFR*19]. We improve stereo depth estimation and implement a new depth fusion.*

life after the Second World War. The content was recorded using the 3D Human Body Reconstruction (3DHBR) system [SFR*19] which allows creating dynamic volumetric reconstructions of both, Ernst Grube and the interviewer. Each of the documentary's six episodes places the protagonists in an authentic virtual recreation of a different historical site. Viewers experience the film with unconstrained motion, i.e., they can move *freely* in the VR environment.

Working towards the production of the documentary, Feiler et al. [FGG*20] present reconstruction results for one interview. However, the asset quality is unsatisfying and the captured data reveals several weaknesses of the 3DHBR pipeline: the two-person scenario and its occlusions are not robustly handled by the depth estimation. Neither is noise in the foreground masks that results from lighting errors in the footage. The interview setup creates large distances between the protagonists and some of the cameras. This can increase the general noise in estimated depth maps which is in turn not robustly handled by depth map fusion. We believe that faithful reconstructions are key to the success and immersion of the Ernst Grube VR documentary. Therefore, we take the next step towards production by revising the 3DHBR pipeline and proposing improvements to it that eliminate the weaknesses above. Our revised pipeline is general and thus applicable to other productions. Our contributions are:

- Revising the volumetric reconstruction pipeline of [SFR*19] in the context of the VR documentary "Ernst Grube – The Legacy"
  - Improved stereo depth estimation that better handles multiple persons, occlusions and foreground segmentation errors
  - A new depth map fusion algorithm that is more robust to noise in depth estimation
- Providing a first impression of the documentary's look by integrating a high-quality asset from our revised pipeline into the designated virtual scene

2. Related work

We focus on work that is closely related to the capture and processing setup used for the Ernst Grube documentary as well as immersive cultural heritage projects with Holocaust survivors.

2.1. Volumetric capture

Volumetric capture systems are end-to-end solutions that output a dynamic 3D reconstruction of the captured content. These systems have not only been commercialized in recent years [4D20, 8i20, Vol20] but also seen increased research interest. The early work of [SH07] reconstructs dynamic surfaces from only 8 RGB video streams. Similarly, [VPB*09] use 8 RGB cameras but also recover normal information from high-frequency lighting patterns. 3DHBR [SFR*19] is a system with 32 RGB cameras (16 stereo pairs) that fully integrates lighting and capture hardware in one cylindrical studio. [CCS*15] illuminate theirs with extra infrared light and use a multimodal fusion to combine data from 106 RGB and infrared cameras. [GLD*19] present "The Relightables", a system with 90 RGB and infrared cameras that not only recovers geometry but also material reflectance. Other work focuses more on processing than capturing [LFB17, RCDAT17]. Due to the heterogeneity of end-to-end systems and tight coupling of hardware and software, it is usually not straightforward and in some cases even impossible to apply other processing workflows directly to the Ernst Grube data. However, parts of different reconstruction pipelines often have similar purposes (e.g., depth estimation), so we can integrate ideas from other systems when improving 3DHBR. For example, [LFB17] inspired us to use truncated signed distance functions (TSDF) [CL96] in our depth map fusion.

2.2. Experiencing witnesses of the Holocaust

The "LediZ" project [BG20, LMU20] creates holographic testimonies of German Holocaust survivors. A viewer experiences the stereoscopically filmed witnesses using 3D glasses and can vocally query a pool of pre-recorded answers. While "New Dimensions in Testimony" [USC20], a similar project, uses volumetric capture, their display-based presentation can only cover very limited viewpoints. The interactive VR experience "Journey Through the Camps" [Sti18, VCTB18] lets users explore and re-experience Holocaust sites by virtually recreating parts of concentration camps. It uses a realistic soundscape and vocal testimonies of real survivors to immerse users into the narration. In contrast to these works, the Ernst Grube VR documentary features a *full* volumetric reconstruction as well as voice recordings of a Holocaust survivor and can be experienced from *arbitrary* viewpoints.

3. Volumetric reconstruction pipeline

In this section, we introduce our revised version of the 3DHBR pipeline [SFR*19]. We first give a basic overview and then present specific improvements. Prior to reconstruction, the 32 cameras are calibrated, their images are radiometrically matched and foreground masks are extracted with an algorithm similar to [HHD99]. Reconstruction itself can be roughly divided into five sequentially executed stages (see Figure 2). Stereo depth estimation computes 16 depth maps (one for each camera pair), which are then fused into an oriented point cloud. The meshing stage extracts a surface using the Screened Poisson Surface Reconstruction [KH13] and smooths as well as simplifies the geometry depending on the desired level of detail. Like the current 3DHBR system, we use recent improvements from [DSF*19]: the full sequence of meshes is temporally registered with the method of [MHE19] and after that, individual meshes are colored by projecting the camera images onto them. Apart from registration, all stages are executed frame-by-frame.

3.1. Stereo depth estimation

Like [SFR*19], we use a stereo depth estimation method based on the Iterative Patch Sweep [WFS11, WFS*16]. It is conceptually similar to PatchMatch Stereo [BRR11] but operates on depths instead of disparities. More precisely, for each foreground pixel in the reference image of a stereo pair, it finds a depth and normal vector with an iterative optimization that considers both temporal and spatial neighbors. At the start of the sequence or at sudden visibility changes, the approach requires a strong initial guess as there is no reliable temporal or spatial information. Schreer et al. [SFR*19] find one reference depth by triangulating the apparent center of mass of the largest foreground object and randomly distribute the initial pixel depths around it. This approach requires a single object, small occlusions and clean foreground masks. The first two assumptions do not hold for the two-person interview setup of the documentary. Additionally, lighting errors in the captured video footage cause the foreground segmentation to fail occasionally, which results in noisy masks. We propose an initialization that is more robust in this context by not making the assumptions above. First, we downscale both stereo images and match feature points between them with the detector and descriptor from [ZREK11]. We then triangulate these points to obtain their depths. Finally, each pixel in the image is initialized with the randomly perturbed depth of its closest feature point and a normal facing the camera.

3.2. Depth map fusion

Schreer et al. [SFR*19] use the method of [EWR*14] for depth map fusion. It builds an oriented point cloud by *selecting* patches from different depth maps, assuming little noise in the input. The interview setup places the protagonists distant from the cameras that see their front sides. Thus, noisy depth estimates for important regions like faces become more likely but are not robustly handled by the selection-based fusion. Instead of relying on error-free inputs, we acknowledge the noise inherent to estimated depths and propose a fusion algorithm based on truncated signed distance functions and point cloud cleaning.

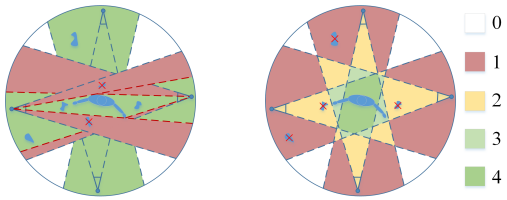


Figure 3: *Soft hull cleaning as seen from the top. The actor in the middle is the foreground object. Left: We remove points that fail the foreground test in one of the cameras. Right: We then remove points from subvolumes visible only in few cameras.*

3.2.1. From depth maps to oriented point clouds

We join depth maps into a single 3D representation, a truncated signed distance function [CL96], with an algorithm based on [NIH*11]. We first build a discretized 3D volume and project all depths maps into it. Using constraints from the projected depths, we then compute the distance to potential surfaces for nearby voxels. A point cloud is extracted by finding zero crossings in the grid of this discretized TSDF. Since all depths are integrated by weighted averaging, Gaussian noise is reduced along the surface. We extend the algorithm above with normal propagation. More precisely, we also average the estimated per-pixel normals. The resulting point cloud normals do not strictly adhere to the surface approximated by the TSDF but are biased towards the average direction of the points' source cameras. This property is exploited in a later cleaning stage (see Section 3.2.3).

3.2.2. Soft hull cleaning

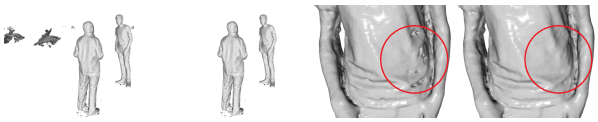
Although the TSDF approximates surfaces reasonably well, the resulting point cloud can still contain points not belonging to the protagonists. To tackle this issue, we introduce a two-step method that uses visual hull-like constraints and a "soft" capture volume clipping (see Figure 3). First, we project each point onto each camera image plane. Only if a point falls within the image borders and is not covered by the foreground mask, we remove it. Retaining points outside of the frustum is motivated by partial scene coverage (i.e., not every camera sees every part of the studio). In the second step, we divide the scene into multiple subvolumes depending on the number of cameras covering them. We then remove points from volumes that are only visible in a few cameras.

3.2.3. Occlusion cleaning

While the soft hull cleaning removes point cloud artifacts outside of reconstructed objects, it cannot remove those inside. We found that noisy depths can also result in *internal surfaces*: groups of points beneath the real surface that follow its local shape. These errors can manifest as bump-like artifacts in the resulting mesh. Note, that normal propagation ensures similar orientation of real surfaces and internal ones beneath. Our idea is to exploit this property and detect the latter based on point occlusions. More precisely, we determine how much a point is occluded by neighboring points in front of it. The closer the neighbors and the similar their orientations are to that of the point, the stronger the occlusion. We then use a threshold to classify strongly occluded points and remove them.

Figure 4: *Depth estimation. Left: [SFR*19]. Right: Ours. We require fewer frames from the start of the sequence to convergence.*

(a) Oriented point clouds (gray) next to their colored meshes.



(b) Soft hull cleaning (c) Occlusion cleaning

Figure 5: *Fusion comparison and effects of cleaning. For (a), left is [SFR*19] and right is ours. For (b) and (c), left is without and right is with.*

4. Experimental results

We show the effectiveness of our changes to the reconstruction pipeline by comparing results of the individual stages as well as a final volumetric asset. Figure 4 shows depth estimation results for a challenging view suffering from strong occlusions. The method by [SFR*19] requires up to 30 frames from the start of the sequence to reasonable depth estimates for both protagonists, while ours only requires one. By using feature points, our initialization is independent of (noisy) foreground masks or the number of objects. As sudden visibility changes pose a similar problem, the results suggest that our method can handle them equally well. While homogeneous objects could be challenging, clothed humans usually provide enough texture to detect at least some features. Figure 5 shows a comparison of fusion methods as well as the contribution of each cleaning stage. The point clouds produced by the selection-based approach [SFR*19] contain more artifacts and surface noise than ours. Although we use averaging, our meshes not only recover the general shape but also details more accurately (e.g., see nose or lips). Soft hull cleaning mainly helps here to remove remains of reconstructed cameras by volume clipping and occlusion cleaning eliminates the bump-like mesh artifacts caused by internal surfaces. The clutter near the feet cannot be removed by visual hull-like constraints because the foreground masks cover it (also see Figure 4).

Figure 6 shows the volumetric asset presented by Feiler et al. [FGG*20] next to ours. Note, that we also tuned mesh smoothing and coloring parameters, so the results are not one-to-one comparable. However, it shows that our revised pipeline is able to produce high-quality assets with faithful volumetric reconstructions of the protagonists. Figures 1 and 7 show the asset integrated into its designated scene: the virtually recreated garden of the children's home in which Ernst Grube spent parts of his childhood. This is the first visual impression of the documentary's episode that will

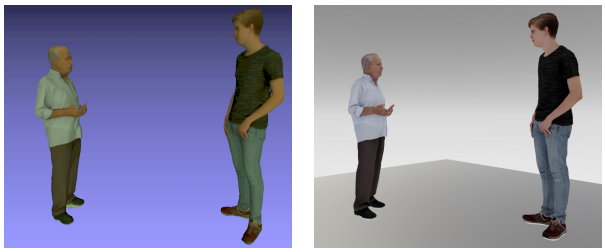


Figure 6: *Dynamic volumetric asset produced from the same data. Left: [FGG*20] using the pipeline by [SFR*19]. Right: Ours.*

be part of a proof-of-concept experience presented at the memorial site Sachsenhausen in Germany. We think that the virtual protagonists not only convincingly portray their real counterparts but also blend well with the surroundings. Having the ability to produce these high-quality assets will be an integral part of the further development of the documentary and a key component to immersion when viewers stand right next to the contemporary witness in VR.

Figure 7: *Protagonists integrated into the virtual scene.*

5. Conclusion

Experiencing stories of Holocaust survivors by virtually joining them on a journey through time is possible with today's technology. The VR documentary "Ernst Grube – The Legacy" features full volumetric reconstructions of interviews with the Holocaust survivor Ernst Grube. We identified weaknesses in the handling of multi-person scenarios and noisy intermediate results in the used 3DHBR processing pipeline, which previously led to insufficient reconstruction quality. We equipped the depth estimation with a more robust feature point-based initialization and replaced depth map fusion with a new method using truncated signed distance functions and point cloud cleaning. Apart from producing high-quality volumetric assets, our revised pipeline improves the robustness and quality of intermediate results. We presented an interview asset produced by our pipeline and integrated it into its designated virtual scene. This preliminary result provides a first impression of the documentary's look and shows that our high-quality assets can help to produce a convincing and immersive experience.

6. Acknowledgments

We would like to thank Frank Govaere and his colleagues from UFA GmbH for the fruitful collaboration in this joint project.

References

[4D 20] 4D VIEW SOLUTIONS SAS: Holosys. https://www.4dviews.com/holosys-volumetric-video-system, 2020. Accessed: 18.10.2020. 2

[8i20] 8I: Hologram capture systems. https://www.8i.com/, 2020. Accessed: 18.10.2020. 2

[BG20] BALLIS A., GLOE M.: *Interactive 3D Testimonies of Holocaust Survivors in German language*. Springer Fachmedien Wiesbaden, Wiesbaden, 2020, pp. 343–368. doi:10.1007/978-3-658-24207-7_21. 2

[BRR11] BLEYER M., RHEMANN C., ROTHER C.: Patchmatch stereo - stereo matching with slanted support windows. In *BMVC* (January 2011). 3

[CCS*15] 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 Trans. Graph. 34*, 4 (July 2015). doi:10.1145/2766945. 2

[CL96] CURLESS B., LEVOY M.: A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1996), SIGGRAPH '96, Association for Computing Machinery, p. 303–312. doi:10.1145/237170.237269. 2, 3

[DSF*19] DIAZ R., SHEHU A., FELDMANN I., SCHREER O., EISERT P.: Region dependent mesh refinement for volumetric video workflows. In *2019 International Conference on 3D Immersion (IC3D)* (2019), pp. 1–8. doi:10.1109/IC3D48390.2019.8975991. 3

[EWR*14] EBEL S., WAIZENEGGER W., REINHARDT M., SCHREER O., FELDMANN I.: Visibility-driven patch group generation. In *2014 International Conference on 3D Imaging (IC3D)* (2014), pp. 1–8. doi:10.1109/IC3D.2014.7032597. 3

[FGG*20] FEILER E., GOVAERE F., GRIESS P., PURK S., SCHÄFER R., SCHREER O.: Archiving the memory of the holocaust. In *Culture and Computing* (Cham, 2020), Rauterberg M., (Ed.), Springer International Publishing, pp. 145–155. 1, 2, 4

[GLD*19] GUO K., LINCOLN P., DAVIDSON P., BUSCH J., YU X., WHALEN M., HARVEY G., ORTS-ESCOLANO S., PANDEY R., DOURGARIAN J., TANG D., TKACH A., KOWDLE A., COOPER E., DOU M., FANELLO S., FYFFE G., RHEMANN C., TAYLOR J., DEBEVEC P., IZADI S.: The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph. 38*, 6 (Nov. 2019). doi:10.1145/3355089.3356571. 2

[HHD99] HORPRASERT T., HARWOOD D., DAVIS L. S.: A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE ICCV'99 Frame-Rate Workshop* (1999). 3

[KH13] KAZHDAN M., HOPPE H.: Screened poisson surface reconstruction. *ACM Trans. Graph. 32*, 3 (July 2013). doi:10.1145/2487228.2487237. 3

[LFB17] LEROY V., FRANCO J., BOYER E.: Multi-view dynamic shape refinement using local temporal integration. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 3113–3122. doi:10.1109/ICCV.2017.336. 2

[LMU20] LMU: Learning with digital testimonies. https://www.en.lediz.uni-muenchen.de/projekt-lediz/index.html, 2020. Accessed: 19.10.2020. 2

[MHE19] MORGENSTERN W., HILSMANN A., EISERT P.: Progressive non-rigid registration of temporal mesh sequences. In *European Conference on Visual Media Production* (New York, NY, USA, 2019), CVMP '19, Association for Computing Machinery. doi:10.1145/3359998.3369411. 3

[NIH*11] NEWCOMBE R. A., IZADI S., HILLIGES O., MOLYNEAUX D., KIM D., DAVISON A. J., KOHI P., SHOTTON J., HODGES S., FITZGIBBON A.: Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality* (2011), pp. 127–136. doi:10.1109/ISMAR.2011.6092378. 3

[RCDAT17] ROBERTINI N., CASAS D., DE AGUIAR E., THEOBALT C.: Multi-view performance capture of surface details. *International Journal of Computer Vision 124*, 1 (Aug 2017), 96–113. doi:10.1007/s11263-016-0979-1. 2

[SFR*19] SCHREER O., FELDMANN I., RENAULT S., ZEPP M., WORCHEL M., EISERT P., KAUFF P.: Capture and 3d video processing of volumetric video. In *2019 IEEE International Conference on Image Processing (ICIP)* (2019), pp. 4310–4314. doi:10.1109/ICIP.2019.8803576. 2, 3, 4

[SH07] STARCK J., HILTON A.: Surface capture for performance-based animation. *IEEE Computer Graphics and Applications 27*, 3 (2007), 21–31. doi:10.1109/MCG.2007.68. 2

[Sti18] STITCHBRIDGE, INC.: Journey through the camps. https://www.stitchbridge.com/work/journeyvr, 2018. Accessed: 19.10.2020. 2

[USC20] USC ICT: New dimensions in testimony. https://ict.usc.edu/prototypes/new-dimensions-in-testimony/, 2020. Accessed: 19.10.2020. 2

[VCTB18] VITUCCIO R., CHO J., TSAI T.-Y. J., BOAK S.: Creating compelling virtual reality and interactive content for higher education: A case study with carnegie mellon university. In *ACM SIGGRAPH 2018 Educator's Forum* (New York, NY, USA, 2018), SIGGRAPH '18, Association for Computing Machinery. doi:10.1145/3215641.3215647. 2

[Vol20] VOLUGRAMS: Reconstruction technology. https://volograms.com/technology, 2020. Accessed: 18.10.2020. 2

[VPB*09] VLASIC D., PEERS P., BARAN I., DEBEVEC P., POPOVIĆ J., RUSINKIEWICZ S., MATUSIK W.: Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 Papers* (New York, NY, USA, 2009), SIGGRAPH Asia '09, Association for Computing Machinery. doi:10.1145/1661412.1618520. 2

[WFS11] WAIZENEGGER W., FELDMANN I., SCHREER O.: Real-time patch sweeping for high-quality depth estimation in 3D video conferencing applications. In *Real-Time Image and Video Processing 2011* (2011), Kehtarnavaz N., Carlsohn M. F., (Eds.), vol. 7871, International Society for Optics and Photonics, SPIE, pp. 133 – 142. doi:10.1117/12.872868. 3

[WFS*16] WAIZENEGGER W., FELDMANN I., SCHREER O., KAUFF P., EISERT P.: Real-time 3d body reconstruction for immersive tv. In *2016 IEEE International Conference on Image Processing (ICIP)* (2016), pp. 360–364. doi:10.1109/ICIP.2016.7532379. 3

[ZREK11] ZILLY F., RIECHERT C., EISERT P., KAUFF P.: Semantic kernels binarized - a feature descriptor for fast and robust matching. In *2011 Conference for Visual Media Production* (2011), pp. 39–48. doi:10.1109/CVMP.2011.11. 3