 DOI: 10.3217/978-3-85125-603-1-24

A workflow for 3D model reconstruction from multi-view depth acquisitions of dynamic scenes*

Christian Kapeller[1], Braulio Sespede[1], Matej Nezveda[2], Matthias Labschütz[3], Simon Flöry[3], Florian Seitner[2] and Margrit Gelautz[1]

***Abstract*— We propose a workflow for generating high-quality 3D models of dynamic scenes for the film and entertainment industry. Our 3D scanning system comprises multiple synchronized 3D Measurement Units that incorporate stereo analysis. We give an overview of the involved algorithms for stereo matching, point cloud registration, semi-automatic post-processing and mesh generation, and demonstrate selected steps of their implementation. The computed 3D models will provide content for 360 degree video production and 3D augmented reality applications.**

I. Introduction

The extensive usage of computer graphics in the film and advertisement industries has drastically raised the demand for high-quality 3D model generation from real film content. Related applications include the creation of realistic special effects and upcoming trends towards immersive 3D augmented reality, virtual reality and 360 degree video content production. While several passive and active depth measurement sensors have become available over the last decade, currently available multi-view 3D scanning solutions have various shortcomings. For example, ReCap [3] from Autodesk processes multiple images of an object from various viewing positions for 3D object reconstruction, however, it appears not to be capable of coping with dynamic scenes. Active devices based on structured light or time-of-flight techniques often place severe restrictions on the scene environment (for example, indoor locations only) or have significant limitations on the size of the scanned area or the number of viewpoints. Furthermore, the scanning information is usually only available as point cloud data, while professional film post-production workflows require accurate 3D mesh models.

We propose to overcome some of these limitations by the development of a scalable and cost-efficient workflow for accurate 3D scanning of film sets and the creation of corresponding high-quality 3D models. The system comprises two stages: online content acquisition (i.e., production side) and offline quality enhancement and mesh generation (i.e., post-production side). On the production side several stereo-based 3D Measurement Units (3*DMU*) are used to acquire a scene from multiple viewpoints in a highly synchronized way. On the post-production side the captured views of all 3*DMU*s enable the conversion of the recorded 2D footage into high-quality mesh models. The suggested workflow is explained in more detail in Section III.

*This work has been funded by the Austrian Research Promotion Agency (FFG) and the Austrian Ministry BMVIT under the program ICT of the Future (project "Precise3D", grant no. 6905496)

[1] Institute of Visual Computing and Human-Centered Technology, Vienna University of Technology, Favoritenstrasse 9-11/193-06, 1040 Vienna, Austria; {braulio.sespede, christian.kapeller, margrit.gelautz}@tuwien.ac.at

[2] emotion3D GmbH, Gartengasse 21/3, 1050 Vienna, Austria; {nezveda, seitner}@emotion3d.tv

[3] Rechenraum e.U., Stutterheimstraße 16-18/2/3/20a, 1150 Vienna, Austria; {matthias.labschuetz, simon.floery}@rechenraum.com

II. Related Work

Single camera calibration relates to the process of estimating the intrinsic camera parameters such as focal length, principal point offset and skew factor. When calibrating multiple cameras, the extrinsic parameters describing the geometric relationship between these cameras are essential. Popular approaches for single camera calibration either rely on 2D objects such as planes [26] or 1D objects such as a wand with multiple collinear points [27]. For multiple camera calibration, a pairwise stereo calibration procedure can be applied [11]. In contrast, the authors of [21] present a fully automatic multiple camera calibration procedure.

Depth reconstruction from stereo image pairs has been studied extensively in the literature, with algorithms traditionally being grouped into local, global and semi-global approaches. Local methods proposed during the last years often rely on adaptive support weight techniques [12] and cost-volume filtering [13]. Very recently, deep convolutional neural networks have been successfully used for depth estimation. For example, Kendall et al. [15] propose an end-to-end deep learning system for computing disparity maps. The 4Dviews [10] system uses an iterative patch sweep method.

Rough alignment of two or more 3D measurements of a scene by so-called global registration is typically based on geometric or topological features. In recent work, the Super4PCS algorithm [17], which matches approximately planar four-point configurations, has emerged as a fast yet robust method. In local registration (fine-tuning the results of global registration) the iterative closest point (*ICP*) algorithm [4] and its variants [18] are well established techniques formulated as optimization problems in a least-squares sense.

Sumner et al. [20] propose a method to interactively and non-rigidly deform meshes. They reduce the complexity of the problem of deforming a mesh by deforming a graph structure, termed the embedded deformation (*ED*) graph. Non-rigid deformation of a mesh can be used to perform non-rigid alignment of two meshes, as shown in the Fusion4D [9] pipeline, which reconstructs a non-rigid scene in real-time. In

Fusion4D, aligned data is stored and fused in a 3D voxel grid. The popular marching cubes [16] and dual contouring [14] algorithms provide the means to transform volumetric to mesh data.

Some further literature will be addressed in the context of the method description in the next section.

III. Proposed Approach

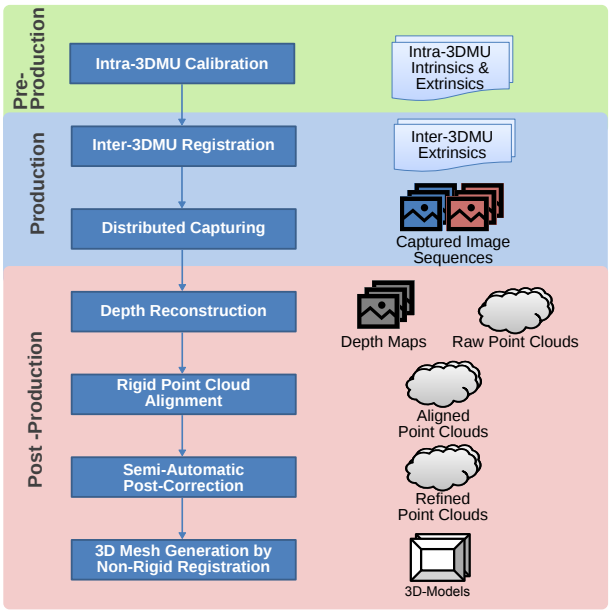

Fig. 1. Processing pipeline of the proposed workflow.

Our reconstruction framework, as shown in Figure 1, is designed to capture scenes with multiple 3*DMU*s. From the acquired data we reconstruct depth information individually per 3*DMU*. In a subsequent step, registration errors among the individual units are minimized using rigid point cloud alignment. After that, the results are refined with our semi-automatic post-correction tool. In the last step we generate mesh models by means of non-rigid alignment.

In the following, we give a detailed overview of the individual steps involved in our processing pipeline.

A. 3DMU Setup and Calibration

We illustrate our sensor setup comprising two 3*DMU*s as depicted in Figure 2. Each 3*DMU* is composed of two industrial-grade XIMEA cameras [24]. Each camera contains a 2/3 inch RGB sensor and is capable to record at a 2464×2056 (5 MPix) resolution in RAW format at a maximum of 60 frames per second. Thus, in total we use four cameras c_i to observe the scene. Here, $i \in \{0,1,2,3\}$ is the camera index, where 0 represents the left camera of $3DMU_1$, 1 the right camera of $3DMU_1$, 2 the left camera of $3DMU_2$ and 3 the right camera of $3DMU_2$. Captured image data is recorded by a controlling computer with USB3 interface. We achieve accurate synchronization of the cameras with a hardware interface operating in master/slave mode.

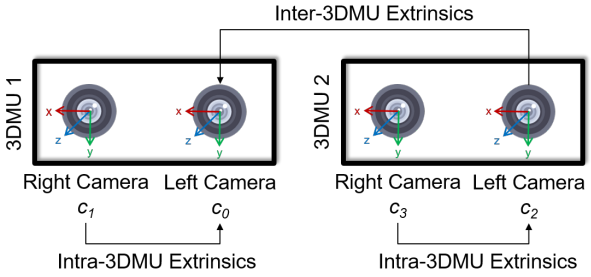

Fig. 2. Illustration of our sensor setup. Please note that cameras are facing the scene, thus left and right are flipped. A detailed description can be found in the text.

For calibration, we perform (i) intra-calibration to obtain the intrinsic and extrinsic parameters for each 3*DMU* individually and (ii) inter-calibration to obtain extrinsic parameters between multiple 3*DMU*s. For both we use the approach of Zhang [26] to compute calibration parameters based on a circle grid calibration pattern. For the former calibration parameters are obtained in a pre-production step as these parameters stay fixed for a 3*DMU*. In particular, we compute calibration matrices and distortion coefficients for each camera individually, and rotations and translations for c_1 to c_0 and c_3 to c_2. For the latter, calibration parameters are obtained during production as soon as the position of each 3*DMU* is fixed. In particular, the inter-3*DMU* extrinsics encode the rotation and translation from c_2 to c_0.

B. Depth Reconstruction

We reconstruct dense disparity maps from each individual 3*DMU* by means of stereo matching. Our method is based on the cost-volume filtering algorithm employed in [19]. For each pixel in the left image of a stereo pair with size $w \times h$, the costs of matching a corresponding pixel in the right image are computed using the Census dissimilarity metric [25] in a given disparity search range d. This gives rise to a cost-volume of dimensions $w \times h \times d$. Subsequently, the cost-volume is aggregated (i.e., filtered) using the fast edge preserving permeability filter [8]. The selected disparity values are those with minimum costs in the cost-volume. This step yields a raw disparity map. Finally, unreliable and occluded pixels are eliminated by means of a consistency check. Disparity values are compared with those of corresponding pixels in a second disparity map that was computed in the same fashion but with the right image as reference. Pixels that disagree by more than 1 disparity count are dropped. In order to improve the quality of the results and run-time, the disparity map computation is embedded into a hierarchical matching scheme. For each stereo image pair a Gaussian image pyramid with $k = 3$ layers is built. Stereo matching is performed first on the coarsest layer l_2. Based on this initial disparity map, an offset map is computed that guides disparity estimation on the next finer layer l_1. The process is then repeated for the finest layer l_0 of the Gaussian pyramid. In image sequences slightest changes of capturing conditions introduce temporal noise in the computed disparity maps. We

address this issue by temporal filtering in the cost-volume following the approach in [7]. Using the intra- and inter-3*DMU* calibration information we project the disparity maps into point clouds in a common coordinate system.

C. Semi-automatic Post-Correction

Depth estimation errors, e.g. due to untextured regions, make additional steps of correction necessary to improve the quality of the multi-view point cloud reconstruction for high-quality 3D film content generation. With this goal in mind, we develop a semi-automatic multi-view 2D-plus-depth visualization and correction tool. To effectively reduce noise and outliers from the resulting point cloud, we implement the multi-view consistency filter of [23] and outlier removal algorithm of [6]. The tool also allows the user to make spatio-temporally coherent local corrections on the disparity maps in a joint-view manner. When it comes to local corrections, we perform binary segmentation operations on 2D video [5] to extract areas we are interested in correcting through user-assisted scribbles in key-frames.

D. Point Cloud Registration and Mesh Generation

In order to fine-tune global 3*DMU* registration, we perform local rigid pairwise alignment of the raw point clouds for a specific frame (that differs by rigid-body transforms only). We employ the method of Pottmann et al. [18], representing an unknown rigid-body transform by its linear velocity vector field and minimizing approximations of a point cloud's squared distance function. In case extrinsic calibration information of the individual 3*DMU*s is not available, a global alignment step based on 4PCS [1] precedes the local alignment step.

We improve a reference frame by fusing each successive frame with the reference. The reference frame is stored as a signed distance field, a volume data-set. For alignment we first extract the reference mesh (the zero-crossing surface in the signed distance volume) via marching cubes. Non-rigid alignment is performed to align the next frame as a point cloud to the reference frame. We follow the approach of Sumner et al. [20] by generating an ED graph for the point cloud to be aligned. Skinning the vertices to the ED graph uses the weighting presented in Fusion4D [9], as this results in a smoother deformation. A least-squares problem that minimizes the distance of the point clouds while maintaining approximative local rigidity through regularization is formulated based on the linear velocity vector field. To integrate the aligned point cloud into the reference volume, we update the signed distance of each voxel grid point by projection onto an implicit point set surface [2]. In a final step, we extract a mesh from the merged signed-distance field via marching cubes.

IV. Experiments

We have conducted a capturing session with two 3*DMU* prototypes and acquired multiple data-sets with real world calibration and image sequence data. Results of the initial processing steps are shown in Figure 3. For these test scenes we created spherical objects of known size in order to assess the accuracy of the 3D reconstruction. Using the intra-3*DMU* calibration information the image sequences were rectified (Figure 3 (a,b) and (e,f)). The computed stereo-derived disparity maps (Figure 3 (c,g)) were afterwards projected into point clouds (Figure 3 (d,h)). We currently capture with a frame rate of 25 fps. By measuring the sizes of the captured spheres in the reconstructed point clouds, we have determined that the diameters of the two reconstructed spheres deviate from their true sizes by less than 6 percent for both 3*DMU*s. A screenshot of the implemented post-correction tool is shown in Figure 4. The displayed scribbles are projected to 3D space using calibration and disparity information and then back-projected to other views, reducing user interactions and aiding the user in the 3D segmentation process.

V. Summary and future work

We have proposed a multi-view depth reconstruction system in the context of a film and media production workflow. The approach aims at the generation of high-quality and temporally coherent 3D meshes from dynamic real world scenes. The implemented processing chain includes algorithms for stereo-based depth reconstruction and geometric 3D data processing, in conjunction with semi-automatic post-processing techniques for further quality enhancement.

The next step of the ongoing project will be to evaluate the results of the individual components and the system as a whole in terms of accuracy and efficiency. Besides quantitative measurements on scene targets of pre-defined size and shape, we plan to evaluate the system by rendering novel views into the viewpoint of an additional reference 3DMU and computing error maps [22]. A complementary qualitative user study will connect objective assessment and subjective judgement.

References

[1] D. Aiger, N. J. Mitra, and D. Cohen-Or, "4-points congruent sets for robust pairwise surface registration," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 85:1–85:10, 2008.

[2] M. Alexa and A. Adamson, "On normals and projection operators for surfaces defined by point sets," in *SPBG'04 Symposium on Point - Based Graphics 2004*, 2004, pp. 149–155.

[3] Autodesk, "Recap," accessed: 2018-04-30. [Online]. Available: https://www.autodesk.com/products/recap/overview

[4] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–607.

[5] N. Brosch, A. Hosni, C. Rhemann, and M. Gelautz, "Spatio-temporally coherent interactive video object segmentation via efficient filtering," in *Pattern Recognition*, vol. 7476, 2012, pp. 418–427.

[6] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5556–5565.

[7] C. Çığla and A. Aydın Alatan, "An improved stereo matching algorithm with ground plane and temporal smoothness constraints," in *European Conference on Computer Vision. Workshops and Demonstrations*, 2012, pp. 134–147.

[8] C. Çığla and A. Aydın Alatan, "Information permeability for stereo matching," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1072–1088, 2013.

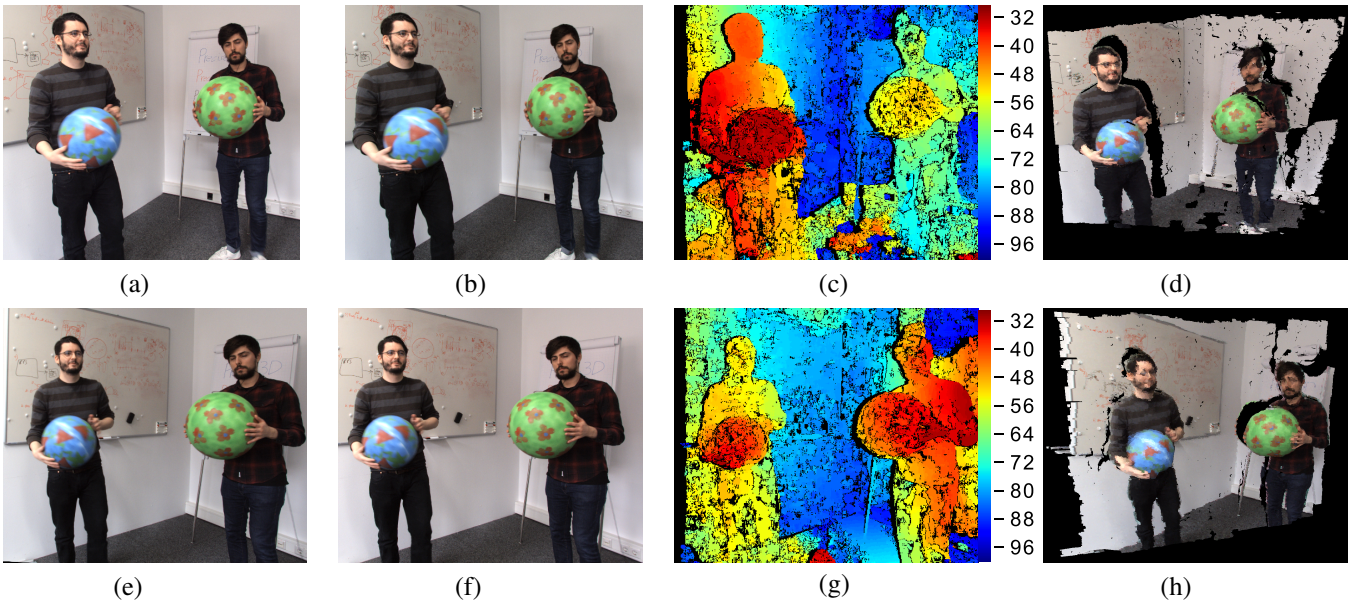

Fig. 3. Data acquired with our experimental setup. The top row shows results acquired by $3DMU_1$, the bottom row those of $3DMU_2$. (a)-(b) and (e)-(f): rectified RGB stereo image pairs; (c) and (g): color-encoded disparity maps, numbers indicate disparity values; (d) and (h) point clouds derived from depth maps.

Fig. 4. User interface of the post-correction tool with scribble placed by the user on a key-frame. The histogram displayed on the left shows the disparities covered by the scribble and supports the segmentation of the selected object and projection of the scribble to other views.

[9] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, *et al.*, "Fusion4D: Real-time performance capture of challenging scenes," *ACM Transactions on Graphics*, vol. 35, no. 4, p. 114, 2016.

[10] T. Ebner, I. Feldmann, S. Renault, O. Schreer, and P. Eisert, "Multi-view reconstruction of dynamic real-world objects and their integration in augmented and virtual reality applications," *Journal of the Society for Information Display*, vol. 25, no. 3, pp. 151–157, 2017.

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[12] A. Hosni, M. Bleyer, and M. Gelautz, "Near real-time stereo with adaptive support weight approaches," in *International Symposium 3D Data Processing, Visualization and Transmission*, 2010, pp. 1–8.

[13] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 504–511, 2013.

[14] T. Ju, F. Losasso, S. Schaefer, and J. Warren, "Dual contouring of hermite data," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 339–346, 2002.

[15] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-End Learning of Geometry and Context for Deep Stereo Regression," in *IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[16] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[17] N. Mellado, D. Aiger, and N. J. Mitra, "Super 4pcs fast global pointcloud registration via smart indexing," *Computer Graphics Forum*, vol. 33, no. 5, pp. 205–215, 2014.

[18] H. Pottmann, S. Leopoldseder, and M. Hofer, "Simultaneous registration of multiple views of a 3D object," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 34, no. 3/A, pp. 265–270, 2002.

[19] F. Seitner, M. Nezveda, M. Gelautz, G. Braun, C. Kapeller, W. Zellinger, and B. Moser, "Trifocal system for high-quality inter-camera mapping and virtual view synthesis," in *International Conference on 3D Imaging*, 2015, pp. 1–8.

[20] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 80, 2007.

[21] T. Svoboda, D. Martinec, and T. Pajdla, "A convenient multicamera self-calibration for virtual environments," *Presence: Teleoperators & Virtual Environments*, vol. 14, no. 4, pp. 407–422, 2005.

[22] M. Waechter, M. Beljan, S. Fuhrmann, N. Moehrle, J. Kopf, and M. Goesele, "Virtual rephotography," *ACM Transactions on Graphics*, vol. 36, no. 1, pp. 1–11, 2017.

[23] K. Wolff, K. Changil, H. Zimmer, C. Schroers, M. Botsch, O. Sorkine-Hornung, and A. Sorkine-Hornung, "Point cloud noise and outlier removal for image-based 3D reconstruction," in *International Conference on 3D Vision*, 2016, pp. 118–127.

[24] XIMEA GmbH, "Ximea MC050CG-SY product specification brochure," accessed: 2018-02-18. [Online]. Available: https://www.ximea.com/files/brochures/xiC-USB3_1-Sony-CMOS-Pregius-cameras-brochure-HQ.pdf

[25] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision*, 1994, pp. 151–158.

[26] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[27] Z. Zhang, "Camera calibration with one-dimensional objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 892–899, 2004.