A Post-processing Tool for Multi-view 2D-plus-depth Video

Braulio Sespede[1]
https://www.ims.tuwien.ac.at
Florian Seitner[2]
https://www.emotion3d.ai
Margrit Gelautz[1]
https://www.ims.tuwien.ac.at

[1] Institute of Visual Computing and Human-Centered Technology,
Vienna University of Technology,
Vienna,
Austria
[2] emotion3D GmbH,
Vienna,
Austria

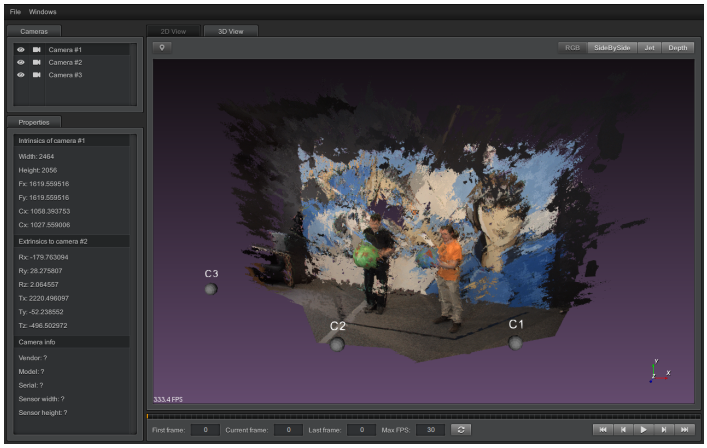

Figure 1: Screenshot of post-processing tool displaying a self-recorded scene in 3D view, displaying 3 cameras and their respective camera parameters (i.e., intrinsics, extrinsics).

Figure 2: Screenshot of post-processing tool displaying the semi-automatic post-processing algorithm. The resulting mask from user input is shown in green.

Over the last few years, the growth of immersive 3D experiences in visual media has increased the need for high-quality depth reconstruction techniques. Capturing multi-view dynamic scenes with high levels of detail remains challenging since errors are normally present in the 3D reconstruction of state-of-the-art algorithms (e.g., via stereo analysis). Artifacts due to erroneous stereo matching become particularly apparent when moving from single-view 2D disparity maps into 3D point clouds representations. With this in mind, a post-processing step that addresses noise removal while preserving details is necessary to prepare 2D-plus-depth content for real-world usage.

In this work, we implemented a semi-automatic post-processing tool to locally correct certain areas of a scene with user input. To achieve this goal, we first implemented multiple 2D and 3D multi-view video visualizations (Figure 1), as they are a key component in detecting errors during depth estimation. Then we implemented several state-of-the-art algorithms to improve the geometric accuracy. Among them, we included a novel semi-automatic multi-view algorithm to correct specific parts of a scene in a spatio-temporally consistent manner [1] (Figure 2). Local corrections are guided by a video segmentation algorithm that relies on an optimization framework based on efficient cost-volume filtering [3] to propagate user-made scribbles. Additionally, the tool incorporates other algorithms such as outliers removal, background subtraction, multi-view consistency enforcement, and the capacity to generate animations. The tool was implemented in C++ using the Point Cloud Library, OpenCV, and QT5. Currently it supports the Microsoft 3D Video format [6] and the Precise3D format, but can be easily be expanded to support any kind of 2D-plus-depth format (e.g., Kinect or RealSense).

The post-processing tool was validated qualitatively and quantitatively. Qualitatively, we incorporated our tool into a multi-view 3D reconstruction pipeline, where we validated the importance of dynamic visualizations to assess the quality of the recordings. Quantitatively, we measured the precision and recall of the underlying video segmentation algorithm against the Microsoft3D dataset [6]. Results have shown that the use of the additional depth channel increases the spatio-temporal consistency of the semi-automatic algorithm. To evaluate the noise reduction capabilities of the semi-automatic post-processing algorithm, we compared the resulting disparity maps against public ground truth datasets. We noticed a small decrease in error at the expense of user annotation effort. Compared to other popular 3D post-processing tools such as MeshLab and CloudCompare, we address the temporal aspect not only regarding dynamic visualizations but also regarding temporal consistency of the correction algorithms. This means that the corrections performed by the user can be propagated to following frames of the video effortlessly. Additionally, the aforementioned tools work exclusively on 3D data while we also preserve and make use of 2D video to simplify the correction process. In contrast to user assisted stereo algorithms such as [2], the main contribution of our algorithm is the introduction of the temporal element to the correction process, allowing coherent corrections over time with minimal user input. Several previous publications on semi-automatic post-processing have focused exclusively on single view scene [4, 5], while we specifically aim for multi-view content. This implies preserving not only temporal but also spatial coherence.

In future work, we plan to improve the semi-automatic post-processing algorithm in several manners. First, the temporal filtering of the cost-volume will be implemented on the GPU using CUDA. Secondly, we plan to incorporate an optional empty background model to the segmentation process, reducing user input and outliers during post-processing.

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] N. Brosch, A. Hosni, C. Rhemann, and M. Gelautz. Spatio-temporally coherent interactive video object segmentation via efficient filtering. In *Proceedings of the joint DAGM and OAGM Symposium*, pages 418–427, 2012.

[2] Y. Doro, N. Campbell, J. Starck, and J. Kautz. User directed multi-view-stereo. In *Proceedings of ACCV*, pages 299–313, 2014.

[3] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE TPAMI*, 35(2):504–511, 2013.

[4] C. Lin, C. Varekamp, K. Hinnen, and G. De Haan. Interactive disparity map post-processing. In *Proceedings of 3D IMPVTC*, pages 448–455, 2012.

[5] K. Ruhl, M. Eisemann, and M. Magnor. Cost volume-based interactive depth editing in stereo post-processing. In *Proceedings of ECVMP*, pages 1–6, 2015.

[6] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM TOG*, 23(212):600–608, 2004.