# Scalable Structure from Motion for Densely Sampled Videos

B. Resch[1,2] H. P. A. Lensch[2,3] O. Wang[1] M. Pollefeys[3] A. Sorkine-Hornung[1]
[1]Disney Research Zurich. [2]Tübingen University. [3]ETH Zurich

Videos consisting of thousands of high resolution frames are challenging for existing structure from motion (SfM) and simultaneous-localization and mapping (SLAM) techniques. We present a new approach for simultaneously computing extrinsic camera poses and 3D scene structure that is capable of handling such large volumes of image data. The key insight behind this paper is to effectively exploit coherence in densely sampled video input. Our technical contributions include robust tracking and selection of confident video frames, a novel window bundle adjustment, frame-to-structure verification for globally consistent reconstructions with multi-loop closing, and utilizing efficient global linear camera pose estimation in order to link both consecutive and distant bundle adjustment windows. To our knowledge we describe the first system that is capable of handling high resolution, high frame-rate video data with close to realtime performance. In addition, our approach can robustly integrate data from different video sequences, allowing multiple video streams to be simultaneously calibrated in an efficient and globally optimal way. We demonstrate high quality alignment on large scale challenging datasets, e.g., 2-20 megapixel resolution at frame rates of 25-120 Hz with thousands of frames.

The input to our method is one or more image sequences. We focus on extrinsic calibration and assume the intrinsics to be fixed and known (in practice they can be computed from a few frames of the image sequences by using Bundler [1]). On a high level our strategy is as follows. First, we perform a modified KLT [2] 2D tracking of feature points utilizing data coherence to reduce drift. Next, we apply a window BA strategy on a set of *confident* frames only. We initialize the window based on the accidental motion approach of Yu and Gallup [5]. We apply bundle adjustment only to subsets of window camera poses and keep the others consistent using a linear camera pose solver proposed by Jiang et al. [3]. To incorporate loop closing, we further establish global anchor links between frame pairs that are carefully selected based on the work of Wang et al. [4]. Those frame pairs link different parts of the video or even different video streams. In addition to these global constraints, relative camera pose constraints from the window BA are integrated with the efficient linear camera pose estimation [3]. We then perform global BA, and finally add all the less confident images by interpolation and BA of their poses. During this step, we keep the scene structure fixed as determined by the confident images. All bundle adjustment is carried out on subsampled point data. The final result is a globally consistent calibration of all input frames from all input sequences.

Figure 1 and 2 show visualizations of some of the reconstructions we computed. Figure 3 compares our technique to several other SfM and SLAM approaches. Our method is more of a SfM approach than SLAM, as it features global BA steps typical to SfM. However, Figure 3 shows that it runs orders of magnitude faster than SfM systems, at comparable speed to current SLAM systems, while producing results which are an order of magnitude more accurate than SLAM systems.

A more detailed description of the method and evaluation is given in the paper. One of the key insights in this work is that the coherence of high spatiotemporal resolution material enables the use of modified tracking, subsampling, and global optimization schemes, which in combination allow for considerably faster and more robust computation than other approaches.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

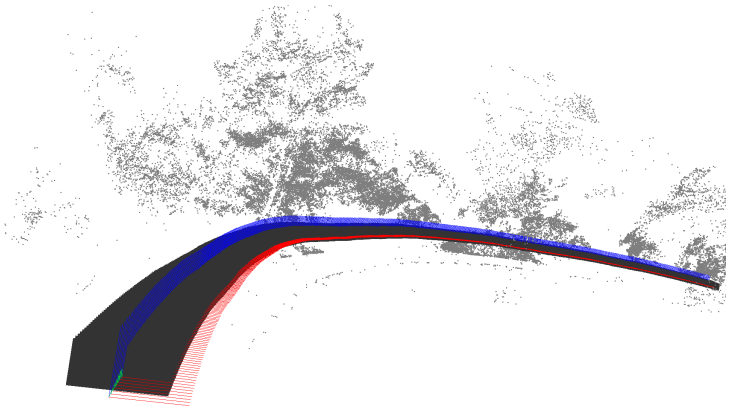

Figure 1: Reconstruction from a very high resolution (5K) sequence with 901 frames. Reconstruction time was 2630 seconds.

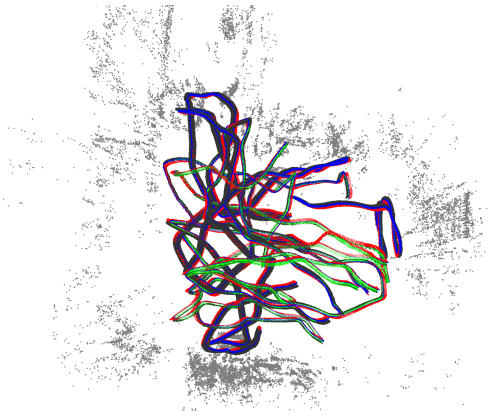

Figure 2: Top down view on the reconstruction from cooperatively captured video sequence set containing 14254 frames from three different cameras. All reconstructed camera trajectories are linked into a common global context by the anchor constraints. Reconstruction time was 1750 seconds.

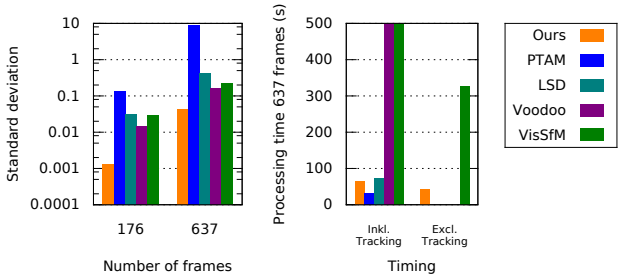


Figure 3: Evaluation based on synthetic ground truth from the Open Movie Project "Sintel". We give the standard deviation of the reconstructions fitted to the ground truth with an affine transformation plus timings. Our approach runs orders of magnitude faster than other SfM systems while producing results which are an order of magnitude more accurate than SLAM systems. PTAM failed after 176 frames due to too slow map update.

[1] Bundler Structure from Motion Toolkit. https://github.com/snavely/bundler_sfm. [Online; accessed 09-Nov-2014].

[2] G. Bradski. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] Nianjuan Jiang, Zhaopeng Cui, and Ping Tan. A global linear method for camera pose registration. In *ICCV*, pages 481–488, 2013.

[4] Oliver Wang, Christopher Schroers, Henning Zimmer, Markus H. Gross, and Alexander Sorkine-Hornung. Videosnapping: interactive synchronization of multiple videos. *ACM Trans. Graph.*, 33(4):77, 2014. doi: 10.1145/2601097.2601208. URL http://doi.acm.org/10.1145/2601097.2601208.

[5] Fisher Yu and David Gallup. 3d reconstruction from accidental motion. In *CVPR*, 2014.