Virtual Insertion: Robust Bundle Adjustment over Long Video Sequences

Ziyan Wu*[1]
ziyan.wu@siemens.com
Zhiwei Zhu[2]
zhiwei.zhu@sri.com
Han-Pang Chiu[2]
han-pang.chiu@sri.com

[1] Siemens Corporate Technology
Princeton, NJ, USA
[2] SRI International
Princeton, NJ, USA

Bundle Adjustment is a key process to enhance the global accuracy of the 3D camera pose and structure estimation in the framework of structure from motion over long video sequences. However, most bundle adjustment algorithms require sufficient visual feature correspondences from each camera frame to its neighboring frames in video sequences, which are hard to collect in real environments, especially for indoor real-time navigation applications. A camera may not observe enough common scene points over a long period of time due to occlusions or non-texture background such as the white walls etc.. With the use of video images as the only input, bundle adjustment will easily fail due to the constant link outage of visual landmarks in the scene. We call it the effect of "visual breaks", and the issue of "visual breaks" has hindered the usage of bundle adjustment. It is particularly critical for sequential Structure from Motion (sSfM) applications where motion estimation is from "chaining" neighboring key frames.

On the other hand, to deal with this issue of "visual breaks", vision-based navigation systems, such as Simultaneous Localization and Mapping (SLAM), typically do not rely on the video cameras only for robustness. Different techniques have been proposed to reduce the drift caused by"visual breaks" and other sources (e.g. inaccurate calibration) by fusing non-vision sensors, such as Inertial Measurement Unit (IMU) [1], LiDAR [3] or GPS [2]. As a result, good motion measurements from non-vision sensors or motion assumptions can be obtained easily at these "visual breaks" locations. However the bundle adjustment is still not able to use the motion estimates from these techniques directly due to a missing approach to incorporate them inside the cost function during optimization.

In this paper, in order to overcome the above issue, we propose a "Virtual Insertion" scheme to construct elastic virtual links on these "visual breaks" positions to fill visual landmark link outage with the measurements provided by other sensors or motion assumptions, so that all camera positions can be linked in the long video by the real or virtual scene landmarks before bundle adjustment. This way enables the traditional bundle adjustment algorithms to achieve robust large-area structure from motion over long video sequences. Specifically, with the measurements from non-vision sensors at the "visual break" positions, we actually convert them into a set of virtual landmark links that will serve as 3D-2D projection constraints in the cost function of bundle adjustment optimization. As a result, measurements from other sensors can be integrated into existing bundle adjustment framework. Experiments on real-world long video sequences show that the virtual insertion scheme can significantly enhance both robustness and global accuracy of bundle adjustment over long video sequences in challenging real-world environments.

A "visual break" is critical especially for sequential structure from motion, where usually a camera position has feature correspondences only with neighboring positions. With the help of IMU and Kalman filter [3], a visual odometry system is able to output reasonable and continuous poses using measurements from IMU especially at the "visual breaks". However the measurements from IMU cannot be integrated into the framework of bundle adjustment directly, resulting large jumps and drifts. This is because "visual breaks" can severely affect the bundle adjustment, in the sense that the global-minima of the whole sequence becomes the combination of local-minimas in each of the two segments of the sequence because the transition between the two sets of locations is unconstrained. This is the reason why large jumps can be found in the output trajectory from bundle adjustment when "visual breaks" exist in the sequence.

Figure 1 shows the illustration of a typical motion estimation over a video sequence with a "visual break" annotated with a red link arrow. Due to drift in the initial poses estimation, the loop does not close although the person travels back to the origin. It can be seen from Figure 1 that the initial estimated trajectory from is continuous and smooth. After feature matching, frames at the end are matched with frames at the beginning, and for the other locations, frames are only matched with their neighboring frames. During bundle adjustment, with the constraints provided the loop closure, the drifts at the end can be reduced. However a large jump can be observed at the "visual break" location, as showed in Figure 1, since the constraints cannot be propagated to the other locations because of the "visual break". It is straightforward to consider this "visual break" as a "broken joint".

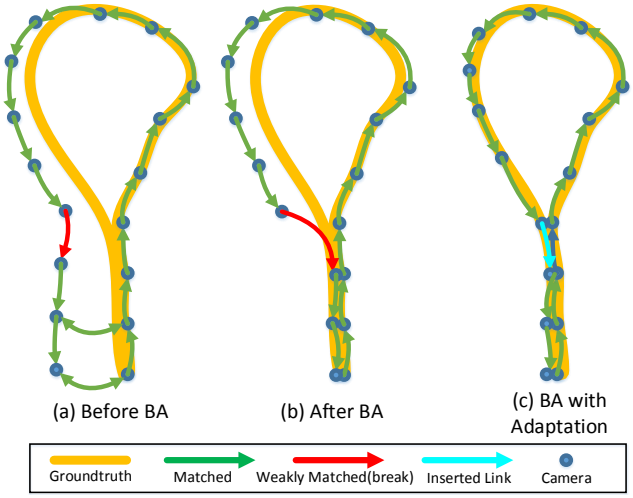

Figure 1: Linking the "visual break".

It is natural for us to think about adapting the "broken joints" with artificial links. From initial estimation of the camera poses fused with IMU, we can set up artificial links on the "visual breaks". Although drift will accumulate over long period in general, within a small period of time, the estimation fused with IMU can be considered as reliable and trustworthy. As shown in Figure 1, a virtual link estimated by IMU motion estimation can be inserted to the break location so that the constraints from loop closure can be propagated to the whole sequence. Hence, as it can be seen that the whole trajectory can reach global optima with drifts reduced on every location. In other words, this method is transferring the motion measurements from non-vision sensors into 3D-2D visual projection constraints, which are integrated into the cost function of bundle adjustment for a joint global optimization. This forms the base of proposed virtual insertion techniques.

[1] H. Chiu, S. Williams, F. Dellaert, S. Samarasekera, and R. Kumar. Robust vision-aided navigation using Sliding-Window Factor Graphs. *ICRA*, 2013.

[2] M. Lhuillier. Incremental fusion of Structure-from-Motion and GPS using constrained bundle adjustments. *IEEE PAMI*, 34(12):2489–95, December 2012.

[3] Z. Zhu, H. Chiu, T. Oskiper, S. Ali, R. Hadsell, S. Samarasekera, and R. Kumar. High-precision localization using visual landmarks fused with range data. *CVPR*, 2011.

*This work was done while the author was a student associate at SRI International, Princeton, NJ, USA.