Report on Thesis of Marek Solony – A scalable multi-sensor reconstruction framework

Laurent Kneip

June 4, 2017

1 Thesis overview

The thesis presents a complete pipeline for structure from motion with fundamentally different types of exteroceptive sensors, meaning classical monocular cameras, laser range measurement sensors, and omni-directional stereo cameras able to sense depth from the application of a stereo optical flow algorithm. This is a timely topic, as the spectrum of possible sensor concepts for performing large scale 3D reconstruction is rapidly increasing, and their capacities are more often than not complementary. Standard cameras, as an example, have the advantage of providing dense photometric information without any limitations in terms of the depth of the scene. Newer sensors such as laser range measurement sensors and RGB-D cameras, on the other hand, provide direct depth measurements thus easing the geometric estimation problem. They may however be limited to a certain maximum depth, slower in capturing frames, and more constrained in terms of the possible environments. The thesis focusses on the two main aspects of the multi-sensor reconstruction problem, namely the so-called front-end problem in which the graphical model of the optimization problem is defined (i.e. neighbouring views, corresponding 3D landmark measurements, and initial values for all variables), as well as the back-end problem which consists of producing a final optimized result that takes all variables and measurement correspondences into account.

2 Main comments

1. The thesis is interesting as it nicely demonstrates how traditional epipolar geometry can be applied to the more exotic cases of spherical imagery. This can notably be achieved by—perhaps to the surprise of several readers—employing the classical epipolar constraints and the derived algorithms without any dedicate changes. The thesis nicely explains why this is the case, and subsequently proves it via successful registration of spherical imagery on multiple datasets.

2. There may be a small misconception regarding the concept of guided matching. As for instance taken from Ochoa and Belongie's work *Covariance Propagation for Guided Matching*, "Guided matching methods are often used to reduce the size of the search region from the entire image to a region expected to contain the corresponding feature.", so it affects the matching stage. The reduction to a region where a matched feature is expected is notably achieved by a geometric (perhaps statistic) prior about the relative transformation. The idea as described in the thesis in fact only represents robust pose estimation. It only encompasses the detection of outliers based on geometric consistency once the matches are given, but not the establishment/refinement of the matches themselves based on a geometric model. Same accounts for Section 4.3.3. This may be more of a detail, and does not change the validity of the presented approach for outlier-resilient relative pose computation.

3. Section 4.8 describes the complete SfM pipeline employed in the present thesis. While it is always impressive to see the implementation of a complete pipeline happening, the exposition could be completed by details about the initialisation of the bundle adjustment problem. To be more specific, what we obtain from robust pose estimation is estimates about the relative pose of the images. Bundle adjustment however optimises the absolute pose of the measurement frames. Since bundle adjustment needs sufficiently good initial estimates, it raises the question how to obtain these initial absolute pose estimates from the relative camera poses given by the initial robust computations. This is a graphical problem, and in the case of larger problems, it can for instance be solved using rotation/translation averaging algorithms. I presume that given the limited number of images in the present scenarios, this is not necessary. Initial values for absolute poses can perhaps be obtained by simply concatenating relative pose estimates, as is usual in incremental structure from motion. Some details about how exactly this has been done are perhaps missing though.

4. The work discusses how to register planar and spherical images. The final goal in particular of adding the planar views is to obtain a denser or at least more complete reconstruction of the environment. The present thesis successfully tackles the problem of registering the planar and spherical views with sparse correspondences. Densification of the spherical images from planar views could now be investigated. In particular, not using the Faro sensor at all and densifying the spherical images with many planar images is a very practically interesting scenario as it would significantly reduce the cost of the entire system.

5. Chapter 6 mentions how "*to obtain the best configuration of the sensor poses and structure points*". Although the 3D to 3D registration error for the stereo spherical images is of course a valid solution to this problem, I am just wondering whether relying exclusively on classical reprojection errors could have achieved the same optimality, or an even better result.

6. I do have one question about the incremental factorization. This may enter some theory that perhaps lies outside the focus the present thesis, but I am nonetheless wondering what happens if the damping factor of the Levenberg-Marquardt implementation changes? This should in theory affect the entire factorization. Since the damping factor is something that can possibly change in every iteration of the optimization, how is the incremental nature of the updates really exploited?

7. Comparison against ICP and Capturing Reality: It is impressive to see how the proposed pipeline is able to outperform both alternative results from academia as well as commercial solutions. With regards to ICP, my question would be whether the correspondences are recomputed as part of the computation or not (classical ICP alternates between a closed-form 3D-3D registration step and the establishment of hypothetical correspondences). This could naturally influence the running time as well as the quality of the registration. The result of the commercial alternative seems to degrade with increasing baseline between the views. So the first obvious question is whether there is some conceptual limitation in the Capturing Reality framework that limits the magnitude of the baselines? Second, are the frameworks really using the same features and correspondences? The type of features (SIFT or ASIFT) may influence the performance in wide-baseline situations. Third, what kind of error metrics are being used? Is the Capturing Reality framework really able to exploit the information in the exact same way. The CLIDAR scans for example may not be used in a similarly effective way, as Capturing Reality primarily relies on the images. Those questions are not to criticise the present work, as it seems to be clearly better suited to solve the present problem. I list those questions merely as hints for further weaknesses of alternative methods that could have been discussed in more detail.

8. Section 2.2, second last paragraph: The exposition in terms of graphical models is of course very interesting beneficial. To my view, BA problems simply are graphical problems, and it makes of course sense to introduce problems as such, and even use the elements of graphical models as parts of an interface for a backend optimization framework. The improved efficiency of recent BA frameworks is of course primarily a result of improved implementations, hardware acceleration, exploitation of sparsity, and concise variable reordering methods such as for instance the Bayes tree method. The author has been a direct collaborator on some breakthrough results in this direction. As a minor point though, the thesis could have provided a few more details about how the formulation in terms of graphical models itself permitted BA problems to become more efficient.

9. Chapter 3 mentions that "*The need for in situ visualisation of the 3D reconstructed environment and taking decision on which parts of the scene needs more sampling, motivated the development of a fast and accurate system for 3D reconstruction from multiple sensors*". This is interesting,

and perhaps would have deserved earlier attention more in the beginning of the thesis.

3 Recommendation

The thesis represents a very interesting contribution in the 3D computer vision community, as structure from motion with input from different exteroceptive sensors is a somewhat unaddressed problem with many potential applications. The content is certainly more on the engineering side, and a few more experiments to analyze the behaviour in different scenarios could have enriched the material. The outcome of the thesis however definitely meets the minimum requirement, as it successfully demonstrates an efficient and complete end-to-end pipeline for solving this very challenging problem, and the result remains something unique even within the research community. The author is however invited to carefully consider and include all above comments.

Sincerely,

Laurent Kneip