Bachelor's Thesis Review
Two-Body Structure from Motion
submitted by *Petr Hrubý*

The topic of Petr Hrubý's bachelor thesis is Multi-Body Structure from Motion (MBSfM), focusing on the 3D reconstruction of scenes consisting of a single moving object in front of a static background captured by multiple image sequences.

MBSfM is an extension of classical SfM to dynamic scenes with multiple rigidly moving objects, which in general is still an unsolved problem. Situations where multiple objects are moving independently are quite common and therefore, MBSfM is an interesting problem with many applications. The studied simplified version of the MBSfM problem with a single moving object in front of a static background captured by multiple image sequences appears for example in systems for 3D scanning of small or medium sized objects. Moreover, this simplified version is an important step towards a more general MBSfM method with multiple rigidly moving objects.

The thesis presents several contributions and demonstrates that the author has a good and fundamental understanding of advanced concepts from linear algebra, graph theory, and camera geometry. In particular, the thesis proposes a complex method for the two-body SfM problem. This problem requires to solve several challenging sub-problems, such as merging different reconstructions and cameras, verification and clustering of motions, a correct segmentation of object and background points, and complex transformations between different coordinate systems. Peter Hrubý designed and applied different concepts like sequential PnP RANSAC, multi-view spectral analysis, algorithms from graph theory, and a specialized bundle adjustment to solve all these sub-problems and to obtain reasonable 3D models for this challenging task.

The studied two-body SfM problem is quite complex in terms of moving between several different coordinate systems and using different representations (projection matrices of different cameras from different takes (sequences of images) w.r.t. different coordinate systems, registered to different reconstructions (w.r.t. background or object), and newly proposed graph representations which sometimes represent motions, sometimes takes and sometimes 3D points). Even though Peter Hrubý spent a lot of effort to correctly describe all situations, concepts, and proposed solutions, and I appreciate especially the descriptions of the used symbols and notations at the beginnings of sections, some parts of the thesis are still hard to follow. This is sometimes caused also by several typos; not precisely or sufficiently described notations (sometimes more indices are necessary to fully and correctly describe a situation); confusions between used terminology - e.g., anchor/reference/origin take; or by using the same notation with different meanings in different sections, e.g., the overuse of the symbol b.

Listed below are a few examples of typos, incorrect statement, and other notes:

1. Fig 1.1 is showing just 3 takes, and in the description, there are 5 takes. Moreover, symbols t and s, in the background of takes 1 and 2, are most likely typos.
2. Page 6: Description of $\vec{O}$ is missing.
3. Page 6: X_δ should be X_β.
4. Page 7: z is not defined, the scalar σ is equal to $\frac{1}{x_3}$.
5. Page 8: Only 11 linearly independent equations are required to estimate P and not 12 as mentioned in the thesis (P can be estimated only up to scale, i.e., equations are homogeneous).
6. State-of-the-art methods presented in Section 3.2.1 Algebraic methods are not fully algebraic. I think most of these methods are based on some optimization techniques.
7. Page 18: "...for every point it holds true: $X^2_{O,\beta} = \ldots$" should be "for every point from the object it holds true:".
8. Page 18: "...its position is the same as it was in the second take": This is not clear and needs to be specified in more detail.
9. Page 20: Define what do you mean by $b < a$.
10. Page 20: Section 4.3 - "All calculated motions are in the coordinate system of the reconstruction of the anchor take, onto which the cameras have been registered. In order to cluster and verify the motions, they have to be transformed to the same coordinate system." - This part is a little bit confusing since it looks like all motions are already in the same coordinate system - the coordinate system of the reconstruction of the anchor take.
11. Section 4.3.1: Origin is used with two different meanings here - for the "origin coordinate system" and also for the "origin OF the coordinate system". The origin coordinate system should be renamed, e.g., to input or initial coordinate system.
12. Section 4.3.1: The point $X'_{\beta'}$ should be $X_{\beta'}$ since it is the same 3D point as X_β only in a different coordinate system.
13. Page 22: K is a calibration matrix of the camera and not a camera.
14. Equation (4.21) and (4.22): $o_i \rightarrow o'_i$
15. Section 4.3.2 and Algorithm 2 and 3 - There is some confusion between the functions Find Inliers and Find Inliers Basis and their inputs.
16. Section 4.3.5: References to Figures 4.7(b) and 4.7(d) should most likely be 4.7(a) and 4.7(c).
17. Page 32: "...zero motions appears to have too high extent" - It is not clear what this means.

18. Figure 4.6 - An improved description of the figure is needed as it is not clear what the difference between (a),(b),(c) and (d) is.

19. "A triplet of vertices can be a subset of a cycle if it does not contain multiple vertices corresponding to the same pair of takes..." $\rightarrow$ This should be "corresponding to the same takes...".

20. Page 44: "The cameras in the clusters in the selected cluster of clusters are used to distinguish the 3D points." - It is not clear what this means.

21. 45: "The weight of this edge is equal to the number of such 2D features." - It is not clear what kind of 2D features are meant here.

22. 4.7.2 - Symbol P in this section is used in a different context than in the rest of the thesis, where P means a camera matrix.

23. Equation (4.40) - $R \rightarrow r$

24. Page 48: "set $T^B_{r,s}$ of all tracks from the object" $\rightarrow$ This should be tracks from the background.

25. Page 61: ".... datasets 5.4(a),5.4(c), 5.4(f)" should most likely be 5.4(f),5.5(b), 5.5(h).

26. Page 65: "An interesting result is the reconstruction of a planar object "Catalog" which is depicted in Figure 5.2(b). The object "Lego" in Figure 5.2(b)" $\rightarrow$ Catalog should be 5.2(c) and Lego 5.2(d).

In the thesis, I'm especially missing a summary of the proposed method with a brief description of all its steps and sub-problems and a pseudo-code of the whole pipeline. Such a summary would significantly help in better understanding the goals and challenges of all 9 sub-problems (steps) of the proposed pipeline described in Sections 4.1-4.9. Even though the descriptions of the used symbols and notations at the beginning of Sections 4.1-4.9 are useful, it would also be helpful to better specify the goals of each step (each sub-problem), specify what is given, what is fixed (e.g. by starting with - "Let's fix the camera i, from the take j....."), and what is the output of each step. For clarity, it would also be useful to add more illustrations and to more precisely define not only the used symbols but also the used terminology, e.g. to define the motion $M_{ij} = (A_{ij}, b_{ij})$ as a rigid transformation of the object between a fixed take i and j, and then use the symbol M_{ij} at all places where it may be confusing which motion is considered. I would also appreciate a more detailed description of the contributions in Section 1.4.

In the experimental section, Peter Hrubý tested the new method on 12 different datasets that he collected. These datasets contain different objects, including planar, shiny, and objects with repetitive patterns. The objects were placed on a non-planar background (folded poster) and several takes (image sequences), each of which depicts a static configuration of the object towards the background were captured. Peter Hrubyý reconstructed the captured scenes using the proposed method, and he compared two different approaches to the final bundle adjustment step (the gradient descent and the alternating minimization). He also compared the proposed method to the standard single-body SfM pipeline implemented in COLMAP [3]. The single-body SfM pipeline is not able to handle moving

objects and was, therefore, run on the scenes without a background. In most cases, the reconstruction of an object with a background using the proposed two-body method produces a model which is worse than the model reconstructed in a single-body COLAMP pipeline without the background. The new two-body method with background produces better 3D reconstructions than the single-body SfM pipeline without the background on the objects which contain repetitive structures. These results are, however, not surprising and are expected. In general, it's hard to reconstruct objects with repetitive structures (that sometimes look very similar from different angles) without additional information from the background. On the other hand it is obvious that the standard single-body COLAMP pipeline would perform better on "standard" objects than the new two-body pipeline which may not correctly classify all object points and may filter some of them.

All tested datasets are quite "simple" and most likely state-of-the-art approaches for object or motion segmentation would correctly segment objects from the background on these datasets. Therefore it would be interesting to test state-of-the-art approaches for object or motion segmentation on these datasets and maybe compare the proposed method also with other MBSfM methods. It would also be interesting to see results on more challenging datasets with a more complex background (e.g., car moving in a real environment...).

General questions:

1. Regarding the sequential PnP RANSAC, it would be useful to clarify how this algorithm relates to the sequential RANSAC proposed in [15]. It's also important to explain why the author decided to use sequential RANSAC, since in Section 3.2.2 he mentioned some disadvantages of this method. Have you also considered Multi-RANSAC proposed in [16] or the method from [17]? How many points do you sample in PnP (what is n)? Have you also considered P4Pf or the calibrated P3P solver? Are you fixing the calibration of all cameras to the same calibration?

2. What are the main possible challenges and issues in extending the proposed method to a MBSfM method with more than one moving object?

In summary, the thesis fulfills all the stated goals, and the author demonstrates a good understanding of SfM pipelines, camera geometry, and linear algebra. The topic of the thesis is of importance to the field; the goals of the thesis were met, and a working two-body SfM pipeline was proposed. The text of the thesis could be improved (see the review). I recommend the thesis for defense and propose the grade of **B (very good)**.

31. 5. 2019

RNDr. Zuzana Kúkelová, PhD
Czech Technical University in Prague,
Faculty of Electrical Engineering