Letter to reviewers

Authors would like to thank again reviewer 1 and editors for their efforts and help offered to improve our work.

The changes in the paper have been highlighted in blue color.

Reviewer 1

I appreciate that the authors took my comments seriously and revised their methodology and presentation. The new simulations demonstrate the performance of their system relative to other common approaches. This comparison gives the paper more 'weight' in terms of contribution. The revisions in the literature review also clarify the author's contribution with respect to the literature. A few points remain:

(Numbering from the previous letter)

(1),(9) The paper is easier much easier to read, however, the abstract still does not explicitly state what problem it are trying to solve. 'The states of robot and target' needs to be explained. As previously mentioned the 'state' could be anything. From the observability section it appears that the goal of the paper is to estimate the velocities of the robot and target in a navigation frame. From the EKF-based SLAM section is it appears that the goal of the paper is to estimate the position and velocity of the robot and target in SLAM frame. Section 1.1 states the the objective is the estimation of the (unobservable) position of the robot and target. It also states that the proposed method seeks to provide scale recovery (see 4) and drift correction. Drift is never mentioned anywhere else in the paper.

The objective of the proposed method is to estimate the states of the camera-robot (position and velocity), target (position and velocity), and landmarks (position). All the states are defined as respect to the navigation reference system W (as it is shown in Figure 1).

For all the cases (section 1.1 (objectives and contributions), observability section, and EKF-based SLAM section) the goal remains the same than the mentioned one, indeed.

On the other hand, we know from our observability study, and also, from all the other observability analyses that can be found in the literature, that any world-centric (navigation-referenced) SLAM system is partially observable in the absence of global measurements.

So why does it appear that we are trying to estimate unobservable states?

To address that question, let us reformulate it, as it has been traditionally been stated, especially by those people with a strong control theory background:

Why do SLAM systems appear to work if they are only partially observable systems?

The fact is that, to our best knowledge, this question remains formally open from the point of view of the control theory!

Specifically, the roots of the monocular SLAM are the SFM (structure from motion) techniques developed from the computer vision research community. In this case, the general problem is to recover, in an offline manner, the structure of the scene, as well as the position and orientation of the camera (for each frame), given a set of image frames (or video). In this case, the real-time operation is not a requirement, then global optimization techniques (e.g. bundle adjustment) are used to find the optimal solution. Actually, nowadays, with the availability of powerful hardware, there is a very popular family of monocular-SLAM algorithms (e.g. PTAM-SLAM) that relies on optimization techniques.

For robotics, the (early) SFM techniques were inconvenient due to the real-time constraints for feed-back the state of the camera-robot to the control system. This fact favored the development of filter-based techniques, as the EKF-based monocular SLAM.

It is interesting to note that most of the available SLAM techniques (not only visual-based) that can be found in the literature are world-centric methods, perhaps because a lot of robotic applications require to locate the robot respect to a navigation frame.

Great advances and even impressive results have been achieved by the SLAM research community, but the problem is far to be yet considered a solved problem.

For instance, the analysis of robustness, convergence and stability properties of SLAM systems are of great interest. Therefore, those properties have propitiated the interest in analyzing the SLAM systems in a more rigorous manner using the control theory tools.

However, at this point, it is important to note that even the majority of the more recent works in SLAM, still does not provide any kind of theoretical analysis or foundation for their methods, but instead only present simulations and experiments as validation or their performance. E.g.

- De Croce, M., Pire, T. & Bergero, F. DS-PTAM: Distributed Stereo Parallel Tracking and Mapping SLAM System. J Intell Robot Syst 95, 365–377 (2019).
- López, E.; García, S.; Barea, R.; Bergasa, L.M.; Molinos, E.J.; Arroyo, R.; Romera, E.; Pardo, S. A Multi-Sensorial Simultaneous Localization and Mapping (SLAM) System for Low-Cost Micro Aerial Vehicles in GPS-Denied Environments. Sensors 2017, 17, 802.
- M. Filipenko and I. Afanasyev, "Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment," 2018 International Conference on Intelligent Systems (IS), Funchal - Madeira, Portugal, 2018, pp. 400-407.

Getting back to the topic of the observability, the only evidenced approach to make the SLAM an observable system (without global measurements) is defining the estimates respect to a local (moving) SLAM reference system as the robot-centric approaches. For some specific applications the robot-centric approach can be convenient, but the more general robotic problem of exploring and self-locating the robot in an unknown environment requires estimates respect to a navigation frame.

On the other hand, as it is shown in our paper, and also as it has been shown in other works, e.g.

- Teresa Vidal-Calleja, Juan Andrade-Cetto, Alberto Sanfeliu. Estimator stability analysis in SLAM. IFAC Proceedings Volumes. Volume 37, Issue 8, 2004. Pages 346-351.
- J. Andrade-Cetto and A. Sanfeliu, "The effects of partial observability in SLAM," IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, New Orleans, LA, USA, 2004, pp. 397-402 Vol.1.
- Kitanov, Andreja and Petrovic, Ivan “Generalization of 2D SLAM Observability Condition.” ECMR (2011).

The observability analyses of (partially-observable) world-centric SLAM systems can still provide very useful qualitative information about important properties of the system (e.g. robustness, convergence, stability). In this case, the objective is not to make the system completely observable, but instead, increase its degree of observability (e.g. make more states/modes observables).

Specifically, in our work, the objective of the observability analysis is to show that by the sole inclusion of an altimeter, some states/modes become observables, and to that effect, it is interesting that not only states related to the vertical axis become observables. In this case, simulation results are congruent with the expectation that the performance of the method should be improved if the observability is improved.

In the revised version of the paper, some sentences have been rewritten to better clarify the problem to solve.

The comment regarding the scale and drift is treated in the next answer.

(4) The scale is still not defined in the paper. It appears in the new Fig. 9 and is mysteriously set to 1. It is not inside of the observability analysis. More information on scale may be easier to find in computer vision textbooks compared to SLAM literature. Another good reference is Weiss's PhD thesis'Vision based navigation for micro helicopters'. In it he addresses the same method as proposed in this paper.

The metric scale as a concept can be simply defined by

s = d_real/d_est

where *s* is the scale, d_real is the actual distance, and d_est is the estimated distance. If the estimated distance is equal to the real distance, then the scale will be 1. If *s* is higher than 1 then the estimated distance is smaller than it should be, and the opposite holds.

In practice, in monocular SLAM there are a lot of real (and estimated) distances: distances between the camera and the landmarks, distances between the landmarks, distances defined by the positions of the camera in time (camera trajectory), and so on. In fact, in the related SLAM methods (that explicitly treat the metric scale), that can be found in the literature, there is not such a standard convention for determining the metric scale. But in general, for determining the scale the average of multiple real and estimated distances is taken.

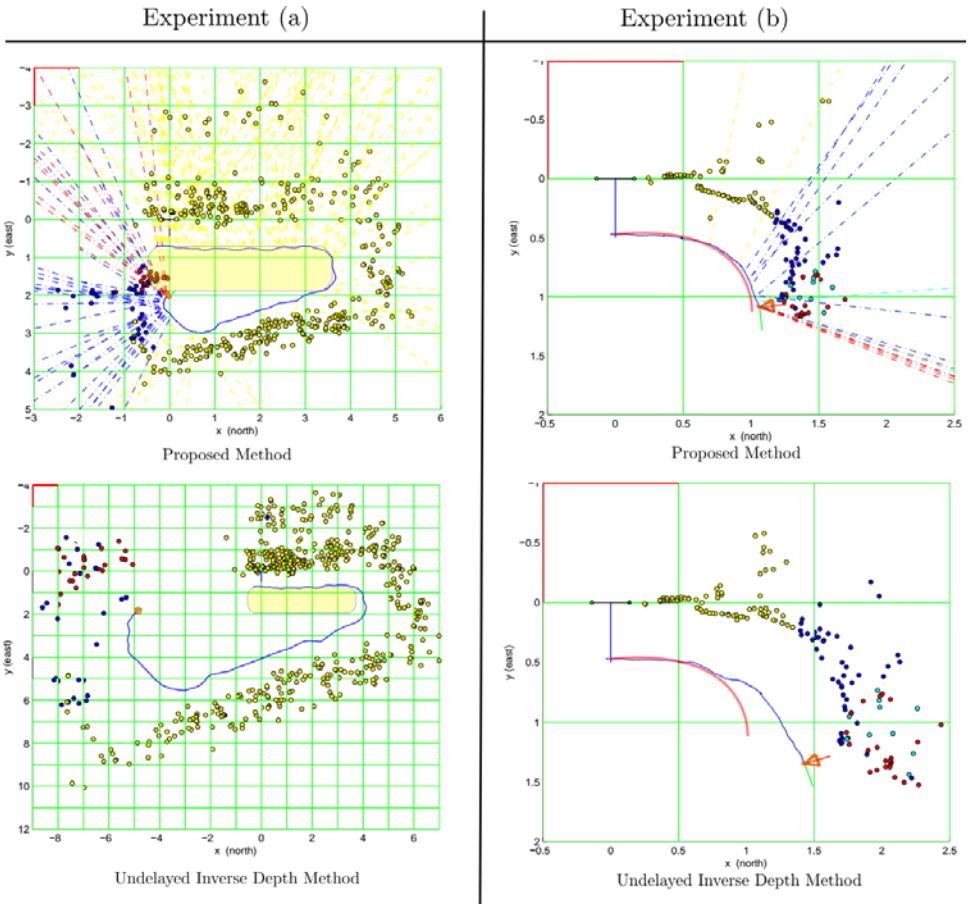
The revised version of the manuscript now includes the expression used for computing the metric scale as a measure of the quality of the estimations obtained from the methods used in the comparative study. We consider that the definition used in our work for the metric scale provides a good balance between practice (for computation) and accuracy.

Regarding the “scale recovery and drift correction” comment, as it has been mentioned, the metric scale recovery is a well-known problem in monocular SLAM, but also the drift is a well-known problem in the metric scale in estimates.

For instance, in monocular approaches relying on anchors at the beginning of the trajectory for setting the metric scale, it can be typically observed degradation of the metric scale as the camera travels far from the anchors. To this fact, the expected drift (dead reckoning error) in the estimated positions of the camera and the landmarks has to be added. In other words, in monocular SLAM methods that do not incorporate continuously metric information into the system, the estimated map and camera trajectory tends to shrink or expand as the camera travels far from its initial position. That means that the scale is time-varying instead of constant.

This fact can be observed in the following Figure. In all the cases there is an error drift in the estimations, but clearly, in the lower plots, a huge component of these errors is caused by a drift in the scale.

In the revised version of the paper, a note has been added to referencing the “scale recovery and drift correction” problem.



We appreciate the suggestion of Weiss's PhD thesis reference. It is important to note that there are several differences in the kind of problem addressed in that work respect the problem addressed in our work. Among others, the method presented in Weiss's PhD thesis does not involve the SLAM problem.

On the other hand, in the past, we already have tried something similar to the approach proposed by Weiss. In our case, we carried out experiments taking the (unscaled) estimations given by a purely monocular SLAM system (Weiss use optical flow instead), and then using an altimeter (Weiss uses inertial sensors) for computing explicitly the metric scale from the (metric) measurements given by the altimeter and the (non-metric) measurements given by the monocular SLAM. Then, the SLAM estimations can be “scaled” according to the computed metric scale.

However, we found the previous approach to be extremely sensitive to the noise of the measurements. The problem is especially notorious for small values of the measurements (where noise dominates) given that the metric scale is determined by: s = d_real/d_est. Of course, some adaptations can be done, as filtering the computed metric scale, but a trade-off is introduced between the time of convergence and the reduction of noise effects.

We have obtained better results by considering the metric scale implicitly in the system, by fusing directly into the system the (metric) measurements given by the aiding sensors (as the altimeters).

Also, from the theoretical point of view, we consider a little "tricky" to take the metric scale as a “virtual measurement”. For instance, what is the noise model for this kind of measurement? On the other hand, if the metric scale is implicitly considered, the required measurement and noise models are the actual ones of the sensors.

Finally, as a note, we also observe that Weiss also checks the rank of the observability matrix numerically.

(6-8) 'The analytical proof required to show that the rank of the observability matrix remains constant from some particular order derivative could be quite difficult'. Yes, this is why it would add value to the paper. The abstract states 'extensive analysis of the system nonlinear observability', however, the robot and the target's dynamics are linearized, scale is ignored, and a numerical rank that excluded the higher order terms was used.

It is important to clarify that higher-order derivatives were considered for computing the numerical rank, as it was explained in the previous letter. We did so to be sure that only the zero and first-order derivatives contribute to the rank.

At this point, we have not too much to add to the discussion of the observability analysis, apart from what it was already explained in the previous letter and the previous answers included in this letter.

On the other hand, as we understand, the reviewer's opinion still is that there is a discrepancy between the actual relevance of the analysis and the relevance that we claim in the paper.

So, to achieve a compromise with the reviewer's point of view, in the revised version of the paper, some sentences have been rewritten to "shade" our claims about the contribution of the observability analysis.

(22) The difference seen in the plot is within the error of GPS, but also within the error of the barometer. Without having a more exact ground truth it is difficult to know. If the behop has a trust worthy altitude measurement fused from different sensors perhaps it should be plotted too.

We agree with the reviewer regarding that a better ground truth would provide a more accurate evaluation of the method performance. Unfortunately for outdoor experiments in natural environments it is very difficult to obtain an accurate ground truth. For small-scale indoor lab-controlled experiments the ground truth can be provided by expensive equipments like Vicon or Optitrack.

On the other hand, it is very important to note that the main objective of the experiment with real data, besides to show that the proposed method can be practically implemented with commercial hardware, is to show that it can provide similar navigation capabilities to a UAV, but without the use of the GPS.

In the same manner that typical UAVs do, Bebop 2 uses its barometer-based altimeter (not the GPS) for controlling its altitude (for very close distance to floor a distance sensor is used). In other words, the altimeter measurements are the best ground truth that Bebop can offer. For long-run horizontal navigation Bebop uses GPS.

So, the important to observe from the experiment, using only the main camera and the altimeter of the Bebop, the proposed method can provide similar navigation capabilities than the original Bebop's navigation system (which additionally integrate GPS, distance sensor, optical flow sensor), in scenarios where a cooperative target is available.

In the revised version of the paper, a paragraph has been added to better clarify this important fact.

For the new figures it would help readability if a name was used instead of #1-5.

Done. Thanks for the suggestion.

Why does method #2 have a time varying scale when it was assumed that the scale is perfectly known from the anchors..

In the description of method 2 (monocular SLAM system with anchors) is said:

“But in this case, to set the metric scale of the estimates, it is assumed that the position of the **set** of landmarks seen in the first frame is perfectly known (anchors)”

But instead should say:

“But in this case, to set the metric scale of the estimates, it is assumed that the position of **a subset** of the landmarks seen in the first frame is perfectly known (anchors)”

Admittedly, if all the initial landmarks seen by the camera are anchors then the initial scale should be equal to one.

On the other hand, for the comparative study, we follow the standard approach that only a subset of the landmarks seen in the first frame are anchors. The initial landmarks that are not anchors (which are not perfectly known a priori) have some initial “weight” in the estimates. Because of the uncertainty of anchors is zero, the remain initial landmarks tend to converge to them, thus making the “average” scale also converging near to one.

Thanks to the reviewer for noting this inconsistency. In the revised version of the paper, the description of the method has been corrected.