Letter to reviewer #1

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

Observability analysis:
'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'
Not including the theoretical analysis in a work is not the issue at hand. A large portion of this paper is dedicated to an observability analysis. So this analysis should be precise and rigorous

We agree with the reviewer. To this effect, we consider that the reviewer's comments regarding the observability analysis have been very interesting and really helpful to improve the quality of the paper.

Scale:
The definition of scale is wrong. The scale of the map can be defined independent of the pose of the camera. But as mentioned it may sometimes be estimated by using the average of known depths of landmarks at particular point in time. It does not change given poor estimates of the pose of the camera.

The basic equation $s = d_real/d_est$ (49) used for defining the scale factor is well known (See for instance Hanel 2018, and Song 2016). and really very helpful.

Indeed, as we already said in the previous letter, for problems like monocular SLAM there are several distances involved, so the exact computation of the metric scale can be in practice quite difficult. In this case, we consider that the proposed methodology (the average ratio of estimated depth to actual depth) provides a reasonable approximation of the metric scale, which is useful for the comparative study purposes. For instance, it is well known that for the monocular SLAM systems the bigger source of uncertainty, and thus of error, is the unknown depth of the landmarks. Also, in our experience with monocular SLAM systems, we have observed empirically how the estimated landmarks depth correlates with the metric scale of the estimated map and trajectory.

As the reviewer has pointed out, with the proposed methodology the computation of the metric scale does not depend on the estimation of the camera pose, which is consistent with the reviewer's next comment regarding the difference between the pose drift and the scale drift.

On the other hand, if the reviewer's comment is pointing that the scale must be defined only in terms of the map, we don't have problems in redefining the formula for computing the scale in those terms.

Therefore, in the revised version of the paper, the methodology for computing the metric scale has been redefined in terms of the distances between landmarks. Also, a reference and an explanatory paragraph have been added.

-Alex Hanel et al. "Metric Scale Calculation for Visual Mapping Algorithms" The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-2, 2018.

-Shiyu Song et al. "High Accuracy Monocular SFM and Scale Correction for Autonomous Driving" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 38, NO. 4, APRIL 2016

I would guess that the scale of the monocular slam system from fig. 9 is ~= 1.5 for 0<=t<=30s. If the other methods were to be loosely coupled and use the same monocular slam system on the same data then the scale should be the same for all of the methods.

Not necessarily. For a loosely coupled system, the scale will also depend on the particular sensor used for providing the metric information. For instance, the characteristics (eg. error, and bias) of inertial sensors (i.e. accelerometers) are different from the characteristics of an altimeter. Moreover, in the past, we have carried out some experiments with loosely coupled architectures and we have observed these differences.

To improve the quality of the comparative study, in the revised version of the paper, we have included a loosely coupled method.

As the proposed method is tightly coupled and does not include the scale (it seems to be taken as measurement noise) the predicted map will be somewhere between the scaled map and actual values depending on a lot of different factors.

The scale is implicitly considered in all the metric variables, and therefore it is not taken as measurement noise. (A good reference to understand this fact is Civera et al. 2007)

We agree that our proposed method makes stronger assumptions (e.g. cooperative target) than other related methods, but if these assumptions are satisfied, then the results show that the metric scale will be very well recovered.

Civera, Javier & Davison, Andrew & Montiel, J. (2007). Dimensionless Monocular SLAM. 4478. 412-419. 10.1007/978-3-540-72849-8_52.

Drift:

In fig 9 only an error in the ratio of estimated depth to depth can be seen.

As we mentioned before, we still consider than the average ratio of estimated depth to depth provides a good approximation of the metric scale, but you are right, Fig 9 mainly shows an error in the metric scale. This is the expected behavior since the camera-robot is following a circular trajectory with loop closure (the initial low-uncertainty landmarks are revisited).

To have a better insight into the performance of the methods in terms of the drift of the scale, in the revised version of the paper, a new experimental setup has been added, where the camera-robot moves far from its initial position, thus causing error drift (both in pose and in scale).

In the figure in the response letter there is not enough information to know if there is a scale drift or a scale estimate error.

We agree that the scale drift and the scale error is not the same, but they are related. In this case, a reasonable model for the scale error could be:

$s_e = s_0 + s_d$

where s_e is the scale error, s_0 is an initial (fixed) scale error, and s_d is a time-varying error (scale drift).

In this case, we consider that the figure is very explicit, since it can be appreciated how at the beginning of the trajectory, the scale of the estimated map is consistent with the actual map, but as the camera moves far from its initial position, it is clear how the scale of the map is overestimated, being independently of the pose drift.

But of course, the figure was included only for illustration purposes in the response letter and does not provide rigorous information about the drift on the scale in that experiment.

The convergence (or divergence) of the estimates is not drift. Dead reckoning error has absolutely nothing to do with this problem (there is no IMU in the purely monocular SLAM nor the proposed

method). The pose drift is from the incremental estimation error in the camera pose/its covariance. The scale drift is the incremental estimation of the map between landmarks. The two values should not be confused.

From our experience in the field, the differences between the pose drift and the scale drift are clear. But it is important to note that since in SLAM the localization (camera pose estimation) and mapping are coupled, then, for instance, the error on the scale affects the pose estimation (thus the scale drift affects the pose drift). Of course, the scale drift is not the only source of the pose error/drift.

Re-time-varying. Yes, the scale is not constant, but it drifts very slowly and will almost not be seen in 30s, but of course this depends on a lot of different factors.

We agree that the scale depends on a lot of different factors, and just for this reason, not necessarily drifts slowly. Working with monocular SLAM systems, we have observed plenty of times that the scale can drift very quickly.

Weiss:
I agree that Weiss's method (loosely coupled, optimization based slam + filter based fusion) versus the proposed method (tightly coupled, filter based) are quite different. His thesis was just an example of something similar that includes an observability analysis and scale. The proposed method has better results, but also include stronger assumptions.

It is important to note that our claim is not that the proposed method outperforms all the other related methods in every possible aspect, but if the required (yet strong) assumptions are satisfied, the proposed method can provide very good performance.

(6-8),(22) Looks good.

Another proofread for English is still required. eg. the new figure has the x axis labeled in Spanish.
Thanks for the observation.