PHYSCON 2017, Florence, Italy, 17–19 July, 2017

AUTOMATIC SCALING IN 3D MAP BUILDING FOR SLAM

Emanuel Trabes[1,2] Mario Alberto Jordan[1,2,3]

[1]Argentinean Institute of Oceanography (IADO-CCT-CONICET).
Florida km.8, 8000, Bahía Blanca, ARGENTINA.

[2]Dto. Ingeniería Eléctrica y de Computadoras- Univ. Nacional del Sur (DIEC-UNS)
Av. Alem 1253, 8000, Bahía Blanca, ARGENTINA.

***Abstract*—In this work we develop a novel approach for indirectly estimation of the metric scale for dense mapping of 3D environments. The scaling factor emerges as a local estimation for providing in real time a metric scale to the depth map. The approach is suitable for real-time monocular SLAM applications. It employs a laser arrangement fixed to the camera whose beams commonly impinge on middle-distant 3D points that are tracked in the frame sequence. The method employs a so-called wildcard frame and a keyframe to estimate the scaling factor along with the optimization of an energy functional to provide good depth estimations at middle-distance. Purpose-built experiments are led to illustrate the approach performance in mapping and tracking and show its feasibility.**

Keywords: Global scaling factor, Monocular SLAM, Direct dense method, Scale estimation, Structured laser beams, Map resolution

I. Introduction

Direct methods for visual SLAM are employed extensively in Robotics and Computer Vision to take advantage of the whole image information in contrast to indirect featured-based methods which focus on parametrized sets of 2D points found in the image [1], [2].

Direct methods include both direct dense, semidense or sparse depth models for structure mapping of a scene [2]. In particular, direct dense and semidense methods emerge as a sound alternative to indirect SLAM, in areas as diverse as aerial, ground, mobile robot navigation, space, swarm and underwater applications.

One important family of direct methods deals with monocular vision, consider, for instance, the two pioneer works in [3] and [4]. The main restrictive and still unresolved issue that delays its massive employment in robotics remains the scale drift. However, motivated by the vast applications of monocular cameras, great efforts are continuously routed in the scientific community to confer dense and semidense methods more accuracy and robustness with ground-truth data.

Moreover, chip technologies involving monocular vision imply an engaging tail wind to many applications in SLAM that have to contend with other recent technologies like kinect sensors or more consolidated alternatives like stereo or multistereo vision [5]. However, the main yardstick to contrast them is the flexibility to adapt the depth range of a given, wherein monocular vision has a decisive and categorical advantage.

Metric information is the key issue in order for a robot to navigate autonomously by unknown and haphazard scenarios. Specially in monocular SLAM, the absolute scale of the world cannot be directly observed. Even when the relationship with the real world is unknown, the map will result precise up to scale. However, the real case is that the arising scale (even if initially it is the scale of real world) drifts over time bringing about an often slow but continuous warping of the map and a deflected camera pathway.

State-of-the-art methods attempt to incorporate constrains derived from semantic and pattern recognition like a tree, a car, a building or employing a sort of feature dictionary or using a planar target of known size to external scale reference in order to convey real metric scale into the map. Actually, owing to the lack of sufficient constrains among resulting local maps, the underlying scale yields inconsistent. Novel proposals with scale-aware formulation like in [4], [6], [7] bring a

[3] Corresponding Author: Mario Alberto JORDAN. E-mail: mjordan@criba.edu.ar. Address: CCT-CONICET-IADO, Florida km.7, B8000FWB, Bahía Blanca, ARGENTINA

palliative in favor of the autonomy, but definitively not a solution for long-term SLAM with scarce reliable loop closures or even more pressing in close-free paths.

Alternatively, monocular camera can be fused with an IMU as scale provider [8]. Furthermore, *ad-hoc* sensors are included like fused GPS-inertial navigation system in semidense SLAM [9], [8], breakthrough technologies like RGB-D sensors that employs time-of-flight cameras (ToF) or structured light by projecting known patterns with diffracting optical elements. With them, mapping results generally much more reliable, albeit these technologies entail much higher cost, complexity and severe limitations for outdoors environments or short depth ranges.

In this work, we develop a novel approach for indirectly estimation of the metric scale for dense mapping of 3D environments which is suitable for real-time SLAM applications. By using ground-truth datasets it is shown the effectiveness of the approach in SLAM and its potential to extend it for any indirect or direct vision method.

A. Objective description

While some direct methods suppose certain limits of the environment depth for later mapping and tracking, other methods provide an ad-hoc mechanism to tackle the scale-drift problem. For instance, LSD-SLAM, suggests a explicitly scale-drift aware formulation that allows the approach to operate on challenging sequences including large variations in scene scale. By contrast, DTAM assumes a fixed depth range which is defined relying on a previous knowledge of land surveying along with the camera altitude in its specified mission pathway.

From the viewpoint of the robotics metric, it is more relevant to focus on short and middle distances of objects to the camera in a manner that enables an autonomous vehicle to react in time against obstacles and to schedule a reference, at best energy-saving, pathway within the scenario.

The objective in this paper is to develop a mechanism for SLAM that aims the improvement of the achievable precision of the SLAM by directly estimating a reliable scaling factor.

II. Scale estimation

A. Cost volume

One of the own characteristics of DTAM is the depth cost volume, see [3]. Each selected keyframe is associated with a sequence of ancillary frames which are to some extent superposed with each other at the time the camera moves on the scene. In other words, each ancillary frame provides a slightly different point of view of the scene with respect to the predecessors, whereby it maintains a small baseline with each other.

The cost volume relates discrete depth hypotheses for each pixel in the keyframe wherein the photometric error between it and someone in every epipolar line of the ancillary frames must be null (see Fig. 1). The photo-consistency of pixels that satisfy the same hypothesis represent essentially the same physical point. The hypotheses are ultimately valid in noise-free situations along with brightness constancy of the scene.

By contrast, in practical use, assumptions are to some extent breached and one can at best expect to test the hypotheses in mean employing to this effect the so-called photometric error functions (PEFs) defined along the epipolar line of every ancillary frame and for every keyframe pixel. This averaging is in our opinion just the property of DTAM that makes it robust in noisy footage. Thereby, one can search for a minimum of the averaged PEF corresponding to every pixel in the keyframe.

In general, an energy functional composed of a data term containing all the per-pixel PEFs and a regularization term that penalizes excessive depth scattering is employed for the optimization. The optimal depth function is found iteratively and is also suitably for parallelism.

B. Modified cost volume

At this point let us introduce a kind of wildcard frame (acronym WF) in the sequence of ancillary frames (acronym AF) situated at best close by the keyframe (acronym KF) as represented in Fig. 1.

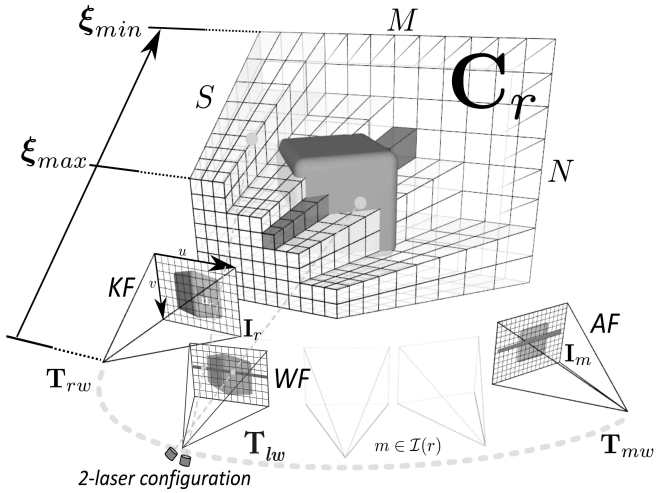

Figure 1 - Integration of the wildcard frame in the cost volume with 2 laser points (*cf.* DTAM in [3])

This wildcard frame (WF) entails the scene almost from the same viewpoint as the keyframe. Consider a laser structure fixed to the camera like the illustrated in Fig. 1. The beams are appropriately oriented to the scene so that they impinge on physical points leaving spots registered on the wildcard frame.

Pixels in the wildcard frame do not relate with any depth hypothesis in the cost volume. They are rather a perturbation of the frame that impairs the intensity of the signalized pixels. However they will contribute with valuable depth measurements.

Further down, we will elaborate on this structured laser in greater details and the way it can beneficially be applied.

C. Scene-depth properties

In the construction of the cost volume we assume statistical consistency in the sense the more dense the set of ancillary frames the better the quality of the mean depth estimation. Moreover, it holds the fact that the closer are the physical points to the camera the wider is the parallax and the more accurate its depth estimation.

However, as physical points of the scene are observed within different timeframes according to the pixel position in the keyframe, the certainty of the photo-consistency hypothesis may result different depending not only on position but also on the permanence in the footage. Certainly, there are physical points in the horizon that remain visible longer than points that are closer to the camera.

In summary, parallax and residence time may have opposing influences over the hypothesis test of every pixel of the keyframe.

Independently of the motion type, features of physical points captured from the camera pathway roughly obey the following table. As seen, parallax and residence time of a physical point can define three depth regions.

Table 1: Classification of depth regions

Point proximity	*Parallax*	*Residence time*	*Uncertainty (variance)*
Distant	poor	large	high
Middle	good	acceptable	small
Close by	excellent	small	medium

Far distant points include almost static points at the horizon with practically null parallax. On the contrary, close-by points produce rapid movements of pixels with tendency to blurriness and short permanence; therefore the uncertainty is not negligible. Thereby it is expected that normal distant points provide the most reliable estimations of depth.

D. Scaling factor

Further reasoning lead us to draw out that if some suitable estimated depth samples of middle-distant points could be compared with some true reference measure of the same point, scaling errors might be corrected to resize the map, at least locally.

Indeed, the availability of depth measures of normal-distant points help calculating the scaling factor as

$$s_i = \rho_{r_i} \; / \; \widehat{\rho}_{x_i,y_i} \tag{1}$$

where i is the index for the laser beam, ρ_{r_i} is the measure, $\widehat{\rho}_{x_i,y_i}$ is the depth estimation and s_i is the localized scaling factor estimation for the point (x_i, y_i). In noise-free situations both for the measures and the estimations, all values s_i for the keyframe would have to coincide.

As this is not the case and added the fact that the scaling factor is not regular on the keyframe self, a simple averaging of all of them is quite desirable. However, certain conditions for the candidate points (x_i, y_i) have to be attached in order to ensure the quality of every s_i. This will be addressed subsequently.

E. Point-depth measurement

The correct scaling of the estimated map relies on external measures of depths along with their depth estimates.

The distance measurement in this work is indirectly obtained as the separation of a pixel spot to the frame center point.

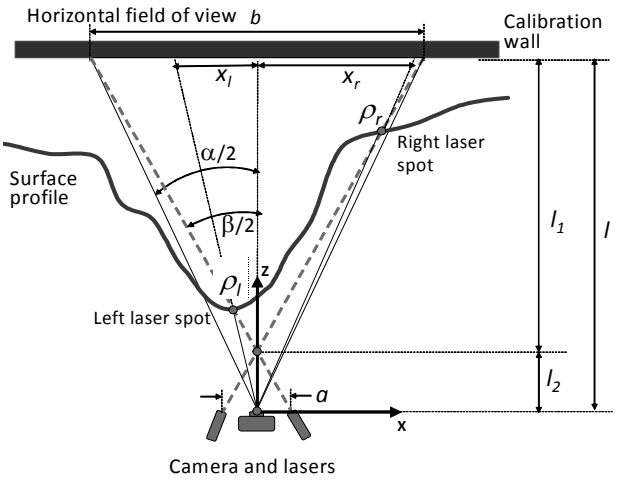

Figure 2 - Top view of the camera-laser configuration with two laser for point-based depth measurement

By way of example consider the Fig. 2 which depicts a two-laser configuration. The beams impact in a 3D

surface of the scene in two physical points with commonly different depths ρ_l and ρ_r, producing spots at proportional distances x_l and x_r in pixels.

The laser beams are configured on the horizontal plane containing the focal center, in which it is expected to find good reference candidates.

On sees the greater the opening angle β between beams with respect to the viewing angle α (the so-called H-FOV degree), the better is the sensitivity of the measurement, albeit the smaller ensues the depth span. However, the practical construction imposes restrictions whereby the lengths l and l_2 are the main design parameters.

F. Candidates for depth references

The reliability of the averaged scaling factor rests, above all, on how good the depth estimations $\widehat{\rho}_{x_i,y_i}$ are for the selected points. It is expected that such favorable points are represented in the middle of the frame around the x axis.

Thereby, for the sake of enhancing the accuracy of the estimated scale, it is advisable to append some conditions for reference candidates before accomplishing (1). These are

1) the spot position on the frame corresponds to a physical point of middle distance.

2) the averaged PEF regarding a particular spot position in the keyframe must have a sharp minimum on the respective epipolar line.

3) the spot should be desirably far from the keyframe center but also far from the side.

At first glance, these conditions seem to be reasonably achievable.

G. Laser calibration

The calibration demands a precise coupling between the intrinsic parameters of the camera and the beam configuration geometry. The intrinsic parameters considered for the calibration are focal length f, image center point offset (x_0, y_0), scaling factors k_x and k_y, skew parameter $\gamma = 0$, video format $w \times h$ (width×height, for instance 360×240) and the H-FOV angle α. Nonlinear camera parameters are not considered here because the footage is previously corrected from these aberrations.

In this work we have constructed a 2-laser configuration like in Fig. 2 because is simple to implement and calibrate.

The calibration is carried out by simply directing the beams to a frontal wall. They have an initial open angle β and a separation a of the beam source points.

For the utmost points in the frame, *i.e.*, $x_r = x_l = w/2$, with w being the frame width in pixels, the interval $[l_2, l/\cos\alpha/2]$ defines the measure range.

Using the set of intrinsic and geometric parameters one obtains

$$Z_l = \frac{a/2}{\tan\beta/2 - x_i/f} \quad \text{and} \quad \rho_l = \sqrt{Z_l^2 + X_l^2} \tag{2}$$

$$X_l = \frac{x_i\ l}{f} \quad \text{and} \quad x_i = x_l/k_x, \tag{3}$$

wherein (X_l, Z_l) are the coordinates of the physical point which is impinged on with a laser beam, x_i is the coordinate x of the spot in the focal plane. Thereupon, solving for ρ_l one gets

$$\rho_l = \frac{a/2}{\tan\beta/2 - x_l/fk_x}\sqrt{1 + \frac{x_l^2}{f^2k_x^2}}. \tag{4}$$

Similarly as for ρ_l one can deduce an equation for the right-side measure ρ_r.

The accuracy is much higher for short and middle distances which is the situation of mayor obliquity between light ray and laser beam.

The detection of spots can seamlessly be refined until reaching a degree of accuracy around subpixel.

H. Laser spot tracking

In principle, between two keyframes there is only one wildcard, preferentially close by the keyframe. However, the shot of the lasers has to be perfectly synchronized with the keyframe and this implies additional electronic and fine adjustment.

We prefer to circumvent this extra hardware by keeping the lasers always powered up and designing a mechanism that is able to draw out the impressed spots in the sequence of KF and related AFs. In so doing, first one should have to search for the spot in the KF and track it on the subsequent AFs.

Secondly, taking into account that the spots have disrupted the brightness information of impressed pixels, some reconstruction technique has to be applied to retrieve the actual intensity values.

With respect to the recovery of the true intensity values of impressed pixels , we can proceed as following:

1) One select the wildcard as the upcoming frame after the keyframe. Only this can afford measures x_l and x_r

2) One tracks spots as well in KF and AFs beginning around the places x_l and x_r. Once the spots are identify, one replaces the intensity of them by the averaged intensities of their neighbors

3) One carries out the optimization on the cost volume as done originally.

The final approach is summarized further down.

III. Scaling-factor approach

The approach for adaptive setting of the scaling factor is outlined in Fig. 3. The procedure includes steps of the approach described insightfully in the previous sections (see blocks at bottom of the figure) concatenated with the cost volume and energy functional optimization of DTAM (blocks at top of the figure).

Ideally, the estimated scaling factors should be keep constant for every keyframe. By contrast in ground truth, the estimated values change permanently on a par with the map quality along with the presence of outliers in the laser measurement.

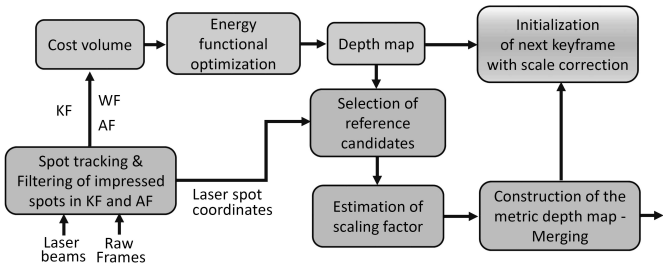

Figure 3 - Estimation of a scale-corrected depth map

Thereby one can construct a metric global map up to drift distortion as

$$\overline{s}_0\widehat{\mathcal{M}}_0 \frown \overline{s}_1\widehat{\mathcal{M}}_1 \frown \overline{s}_2\widehat{\mathcal{M}}_2 \frown ... \frown \overline{s}_t\widehat{\mathcal{M}}_t, \qquad (5)$$

where $\frown$ stays for point cloud merging, $\overline{s}_i$ is the mean value of the estimated scales for the keyframe K_i, $\widehat{\mathcal{M}}_0$ is the initial map patch and $\widehat{\mathcal{M}}_i$ is a subsequent nonscaled 3D-view map patch corresponding to K_i.

IV. Case studies

In order to test the feasibility and performance of our approach we have documented some important tests. The camera pathway consists of a hike through a wooded area. In this particular environment the surface consists partially of grass which produces noisy depth estimations and measures due to a slotted-surface effect.

Two landmarks are put *ad-hoc* into the scene as yardstick for scale verification after the experiments have concluded. They are separated exactly $0.48m$ each other and will appear at the beginning and end of the footage.

We employ a 2-red-laser arrangement in terms of Fig. 2, each one with a power of $50mW$ and an opening angle $\beta = 57.6°$ between beams, laser separation $a = 36cm$. The settings for a camera Gopro are: wide H-FOV degree $\alpha = 170°$, focal length $f = 15mm$, scaling factors $k_x = 0.579pix/mm$ and $k_y = 1.0411pix/mm$, resolution$= 848 \times 480$, principal point offset $(x_0, y_0) = (419.19, 211.65)$. The settings aimed to have well depth estimations in the interval $[0.5m, 4m]$.

Fig. 4 depicts the impact of the estimated maps on the determination of metric distances between physical points, as for instance the distance between the reference landmarks. The estimations are in general noisy, however in the case of the scaled map the errors are comparatively much more small and stable over time. On the contrary, the errors achieved with the unscaled map seem to increase unboundedly.

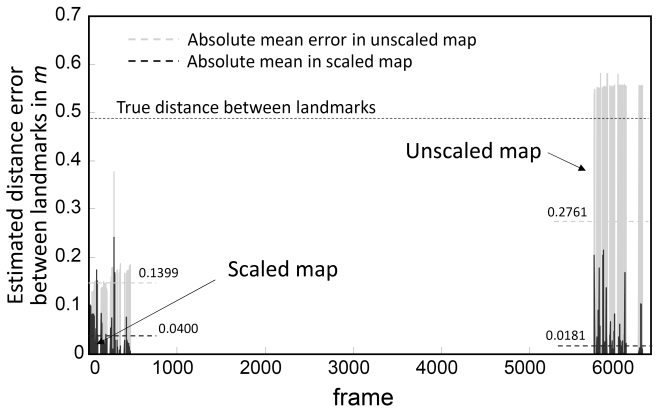

Figure 4 - Error in the estimation of the reference distance at the start and revisiting stretches

Fig. 5 shows the evolution of the estimated scaling factor over time. It is seen that the estimation is very changing and intermittent and occurs when the set of conditions for reference candidates are satisfied.

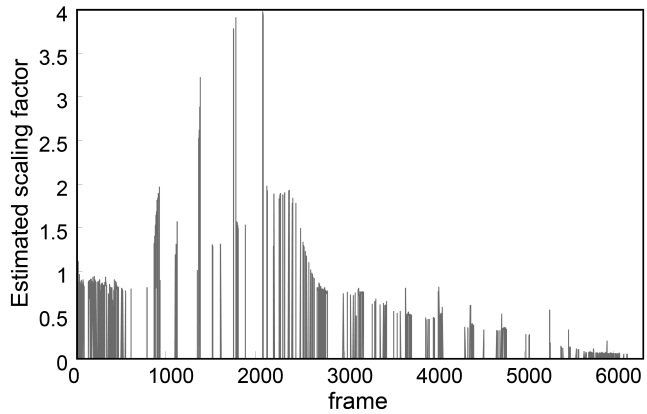

Figure 5 - Evolution of the scaling factor over time

Fig. 6 illustrates the achieved quality in the map estimation through the regularized and the data-term-based depth maps. The slotted-surface effect is partially appreciated in the white and black specks corresponding to close by points on the grass.

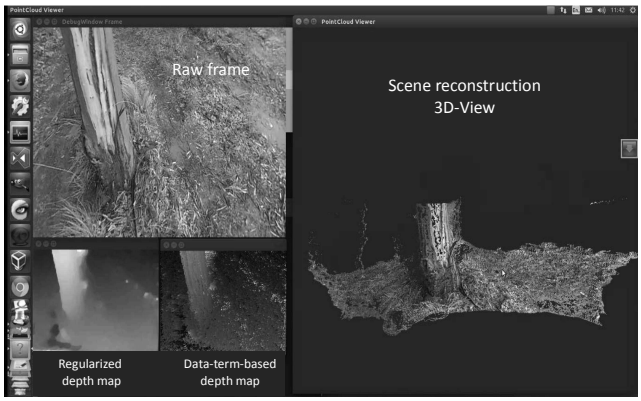

Figure 6 - Scene of a wooded environment

Fig.7 portrays the effect of scale drift in 3D-view of the same scene reconstructed by the scaled and unscaled map at the final stretch of the pathway. The landmarks are tracked in the frames based on the depth information provided from both cases. Clearly, the proportions of the object containing the landmarks are despair according to the case, wherein the unscaled map enlarges the objects significantly in relation to the real wold.

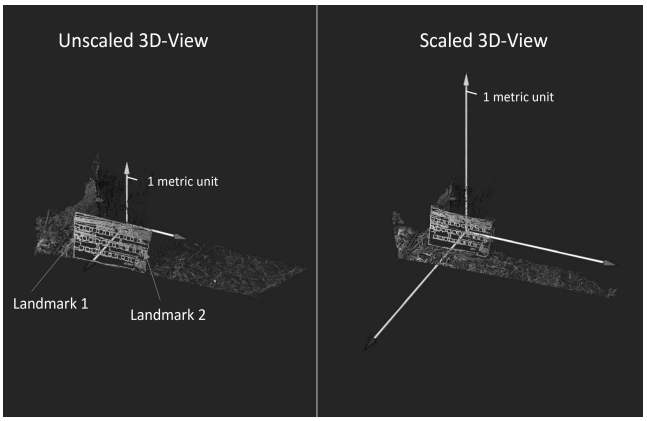

Figure 7 - Scale drift in the unscaled mapping after a long time respecting the real metric unit in the axes

V. Conclusions

A scaling factor is a necessary information in SLAM robotics for navigation in metric spaces. This approach presents a method for estimating the world scale in real time associated with a depth estimator in real time, in this work we uses DTAM for depth mapping.

An arrangement of laser beams is fixed to the camera and directed to horizontal zone of the frame where it is supposed the best estimations are located, precisely at medium distances. The impact of laser beams on the scene is detected by tracking the spot on the horizontal line which give reference measure of depth.

The corresponding depth estimations are processed to find suitable candidates to establish a local scaling factor. A set of design parameters of camera and laser arrangement enables the scaling-factor estimator to suit to a desired depth field.

Purpose-built experiments show that a good performance for real scaling a scenery outdoors.

References

[1] S. Lucey. *Direct Visual SLAM*. In 16-623 - Designing Computer Vision App. Carnegie Mellon. The Robotics Institute, 2016.

[2] X. Xu and H. Fan. Feature based simultaneous localization and semi-dense mapping with monocular camera. In Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, 2016, pp. 17-22. IEEE, 2016.

[3] R.A. Newcombe, S.J. Lovegrove, S.J. and A.J. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *2011 IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2320-2327. IEEE, 2011.

[4] J. Engel, J. Schöps, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *European Confonference on Computer Vision (ECCV)*, pages 834-849. Springer, 2014.

[5] F. Endres, J. Hess and N. Engelhard. An Evaluation of the RGB-D SLAM System. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2012.

[6] J. Grater, T. Schwarze and M. Lauer. Robust Scale Estimation for Monocular Visual Odometry using Structure From Motion and Vanishing Points. In *Intelligent Vehicles Symposium (IV)*. IEEE, 2015.

[7] H. Strasdat, J.M.M. Montiel, and A.J. Davison. *Scale Drift-Aware Large-Scale Monocular SLAM*. In Robotics: Science and Systems VI, 2010.

[8] S. Weiss. Dealing with Scale. Tutorial Computer Vision Group, NASA-JPL/CalTech, 2014.

[9] D. Bender, F. Rouatbi, M. Schikora, D. Cremersy and W. Koch. Scaling the World of Monocular SLAM with INS-Measurements for UAS Navigation. In *19th International Conference on Information Fusion (FUSION)*. IEEE, 2016.