2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Madrid, Spain, October 1-5, 2018

Stereo Visual Odometry and Semantics based Localization of Aerial Robots in Indoor Environments

Hriday Bavle[1], Stephan Manthe[2], Paloma de la Puente[1],
Alejandro Rodriguez-Ramos[1], Carlos Sampedro[1], Pascual Campoy[1]

***Abstract*— In this paper we propose a particle filter localization approach, based on stereo visual odometry (VO) and semantic information from indoor environments, for mini-aerial robots. The prediction stage of the particle filter is performed using the 3D pose of the aerial robot estimated by the stereo VO algorithm. This predicted 3D pose is updated using inertial as well as semantic measurements. The algorithm processes semantic measurements in two phases; firstly, a pre-trained deep learning (DL) based object detector is used for real time object detections in the RGB spectrum. Secondly, from the corresponding 3D point clouds of the detected objects, we segment their dominant horizontal plane and estimate their relative position, also augmenting a prior map with new detections. The augmented map is then used in order to obtain a drift free pose estimate of the aerial robot. We validate our approach in several real flight experiments where we compare it against ground truth and a state of the art visual SLAM approach.**

I. Introduction

Nowadays, autonomous aerial robots have received increasing attention for indoor applications such as inspections, search and rescue. In order to perform autonomous indoor missions in cluttered environments, accurate localization of the aerial robots constitutes an important problem.

Several well known localization as well as simultaneous localization and mapping (SLAM) techniques presented for aerial robots require high precision 2D or 3D laser range finders. Due to the weight restrictions, usually such sensors require a larger size aerial robotic platform, which is a clear disadvantage in cluttered indoor environments. Most localization and SLAM techniques using lightweight RGB or RGB-D sensors which depend on low level characteristic features from the environment such as points or lines. These techniques suffer from inherent limitations regarding view point dependency, adequate lighting conditions and repetitive patterns, deteriorating the data association as well as loop closure capabilities.

Advances in faster and robust object detection and classification techniques have given rise to several semantic based localization and mapping methods using the higher level features from the environment, hence improving the problem of data associations and robust loop closures, overcoming to great extent the view point dependency limitations. Although higher level semantic information ensures robustness in terms of data associations, extracting the relative position of the semantic landmarks for accurate localization can be a challenging task with noisy sensors, and especially in unstructured indoor scenarios where the 3D structure of the objects may vary significantly.

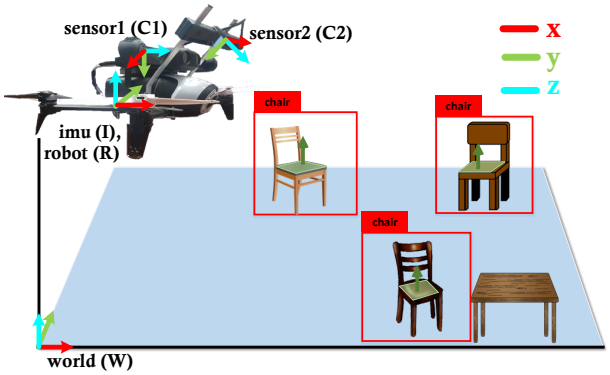

Fig. 1: The mini-aerial robotic platform used to validate the proposed approach, including its relevant reference systems.

Taking advantage of the lightweight property of RGB and RGB-D sensors and improving localization by means of semantic landmarks, in this paper we present a particle filter based localization algorithm for accurately estimating the pose of mini-aerial robotic platforms in unstructured and cluttered indoor environments. The prediction stage of the filter consists of our stereo VO algorithm which is accurate during short time intervals but suffers an accumulation of error during long time intervals. This accumulated error is corrected in the update stage, in two phases using: 1. The inertial measurement unit (IMU) data correcting the orientation obtained from the predicted pose. 2. The relative 3D position computed from the horizontal planar surface, associated with the corresponding semantic landmark in the map, updating the 3D predicted position of the aerial robot.

The rest of the paper is organized as follows. Section II presents the related work. Section. III explains the proposed approach, thus explaining the VO algorithm, semantic detection as well as segmentation and at the end particle filter based localization and mapping. Section. IV explains the performed experiments and the obtained results. In conclusion Section. V summarizes the paper and outlines future work directions.

[1] Computer Vision and Aerial Robotics (CVAR) Group, Centre for Automation and Robotics (UPM-CSIC), Calle José Gutiérrez Abascal 2, Universidad Politécnica de Madrid (Spain). hriday.bavle@upm.es
[2] Institute for Computational Visualistics, University of Koblenz-Landau, Universitätsstr. 1, 56070 Koblenz, Germany.

The authors would like to thank the Spanish Ministry of Science MICYT DPI2014-60139-R for project funding. This work is also partially funded by the Spanish Ministry of Economics and Competitivity: DPI2017-86915-C3-3-R.

II. Related Work

Properly recognizing references in the environment is fundamental to mobile robots localization. While existing methods based on low level geometric features usually present data association problems in cluttered or changing scenarios, employing high level landmarks and object representations enhances robustness and improves spatial scene understanding [1], [2], [3], [4]. Incorporating semantic information provides further disambiguation and is closer to human-like reasoning [5], [6], [7], [8], [9].

Several works in the literature have addressed the problem of building semantically meaningful maps. Rogers *et al.* [5] showed the benefits of interpreting door signs for mapping office environments. Pronobis *et al.* [6] proposed a probabilistic framework for inference of room categories considering common sense knowledge in domestic applications. Applying model-based object recognition, Günther *et al.* [10] create maps of several classes of furniture: office chair, conference chair, desk, conference table, and shelf. Ruiz-Sarmiento *et al.* [11] model and exploit contextual relations and estimate the uncertainty of possible groundings, evaluating their work on a domestic environment dataset. Murali *et al.* [12] add semantic classification to a visual SLAM system in order to reduce data association false positives in the presence of dynamic objects in urban environments. A survey comparing different semantic mapping methods and analyzing related trends can be found in [13]

Most previous approaches address the geometric and semantic representation of the objects separately. In contrast, Sünderhauf *et al.* [14] presented an integrated approach for building point cloud maps including semantic labels of previously unknown instances of object classes. This method is similar to ours in that it combines image based DL recognition of generic objects and 3D geometry, but we use a more compact representation based on horizontal planar surfaces and employ a prior map.

Localization given prior maps is usually the preferred method for navigation in environments where such maps are available or can be easily obtained in an initial setup phase. Toft *et al.* [8] apply an optimization method for localization using semantic categories of visual points and curves in urban areas. In order to improve global localization from multi-view data, Gawel *et al.* [15] propose semantics based graph generation and matching. Atanasov *et al.* [16] incorporate a sensor model based on semantic object recognition and data association, using only vision. Ma *et al.* [7] introduced semantic information based on DL detection in roads into particle filters, hence achieving lower computational cost and better reliability. The direction of the sun is modeled as a Gaussian distribution and other semantic cues -like the road type- are exploited by means of discrete distributions.

Semantic perception has also been employed for robot navigation based on situations classification [17]. Regarding aerial robotics, Ghasemi *et al.* [18] developed a Visual Teach and Repeat method in which semantic objects are identified as high level landmarks. Maravall *et al.* [19] presented a hybrid method for semantic localization in topological maps built from object images and for semantic autonomous navigation of a quadrotor.

III. Proposed Approach

A. Stereo Visual Odometry

Visual odometry can be performed either by a monocular or a stereo setup, although in monocular cases its difficult to estimate the true scale of the estimated trajectory and its performance degrades in presence of pure rotations. Stereo VO approaches overcome these problems, providing more reliable results. Hence the prediction stage of the particle filter consists of our previously developed stereo VO approach [20]. Its main feature is the maximization of information gain by utilizing key points with and without depth information.

The algorithm extracts FAST key points from the stereo image pairs and filters them by means of bucketing [21]. To obtain key point correspondences, a key point matching between the key points of the left and the right camera image is applied, while the key points of the left image are tracked with the KLT-Tracker in order to obtain correspondences over time. Based on the matched key points the algorithm triangulates 3D points and deduces 3D to 2D correspondences.

The motion estimation is divided into an initialization and a refinement stage:

- For motion initialization the algorithm extracts a relative pose from the essential matrix of two temporally consecutive images from the left stereo camera. However direct computation of the correct scale of the translation is not possible. Hence it is estimated from 3D points with correct scale triangulated from key point matches and corresponding 3D points triangulated with the wrong scaled translation vector of the relative pose.
- The motion refinement is done with a bundle adjustment approach. In many cases it is not possible to match key points between the left and the right stereo image while they get tracked correctly in the left stero image. Therefore a large number of key point correspondences without 3D information exist, which still contain information about motion. These can be exploited with the epipolar constraint, by computing the distance between an epipolar line and a corresponding key point. In addition for 3D points which were tracked correctly in the left image, the reprojection error is computed. Both errors get small if the estimated pose is correct. Therefore a sum of squared errors is minimized by means of the Levenberg-Marquardt algorithm.

B. Semantic Detection and Segmentation

For real time detection of the semantic data we use the You Only Look Once (YOLO) object detector and classifier [22], which can detect over 9000 object categories in real time providing a bounding box of the detected object along with its class and its probability.

In our case, we configure the object detector to detect and classify only the required semantic data types. The semantic detection is performed on a single RGB camera. The detected

bounding box is transformed into its corresponding depth image in order to extract the 3D point cloud data. The computed bounding box of the detected object can vary in size due to viewpoint changes, so the corresponding 3D point cloud data can include 3D points not relevant to the corresponding semantic data, e.g. floor points in the detections shown in Fig. 1. Hence taking a median of all the 3D measurements of the points inside the bounding box cannot provide an accurate relative 3D position of the semantic object. In order to minimize the errors in the extracted relative position, we propose to segment the horizontal plane from the detected semantic data. We divide this technique into two parts:

- *Normal Based Clustering:* In order to cluster the 3D planar surfaces we compute the point normals from the segmented 3D point cloud data. We use the integral normal estimation technique presented in [23]. The orientation of the computed normals is used to cluster all the 3D planes present in the detected semantic data. We use the k-means clustering algorithm, which provides the normal centroids of all the planar surfaces.
- *Horizontal Plane Segmentation:* In accordance to the defined world reference frame W as shown in Fig. 1, the normal orientation of horizontal planes is always parallel to the z-axis. Using this rule, we segment the normal centroid of all the horizontal planes obtained previously, along with their corresponding 3D points. These segmented 3D points are used for computing the vertical height of the horizontal planes. We apply a second k-means clustering in order to cluster all the height centroids. As in our case, we use a horizontal plane just above the ground plane, its corresponding height centroid can be easily segmented along with its corresponding 3D points. A third k-means clustering is used for finding the closest centroid of the obtained 3D points.

This method ensures that the relative 3D position of the detected semantic data is extracted only when its horizontal plane is present, discarding all the unnecessary 3D points within the detected bounding box. Thus providing an accurate semantic cue for the update stage of the particle filter.

C. Particle Filter based Localization

In order to accurately fuse the stereo VO data presented in Section. III-A, the extracted relative 3D position of the semantic data from Section. III-B as well as the inertial data, we use the non-parametric implementation of Bayes filter, known as the particle filter. We chose a particle filter over a standard Kalman filter implementation in order to incorporate different distributions from the detected semantic data in the future. We also selected this approach because of its global localization capabilities. The filter consists of M particles. Each state vector $\boldsymbol{x}^{[m]}$, where m is from 1 to M, models a state vector hypothesis comprising of $[\ x\ y\ z\ \theta\ \phi\ \psi\ ]^T$, where x, y and z are the positions of the aerial robot in the x, y and z-axis respectively, in the world reference frame W, and θ, ϕ and ψ are the respective pitch, roll and yaw represented by Euler angles with respect to frame W. We assume that the initial state of the aerial robot is well known and thus the initial state vector of all the particles has the same value.

The particle filter also contains a map of the semantic landmarks of the environment in the form of $\boldsymbol{L} = \boldsymbol{L}_1, .., \boldsymbol{L}_n$. $\boldsymbol{L}_i = (\boldsymbol{l}_i^z, l_i^c)$, where $\boldsymbol{l}_i^z$ is the 3D position vector of the i-th landmark in the world frame of reference and l_i^c is the class of the corresponding semantic landmark. The filter presents the following stages:

1) Prediction: The prediction stage uses the 3D pose measurements of the aerial robot obtained from the stereo VO algorithm in the frame W. The state of each particle is propagated from the previous measurements $t-1$ following a normal distribution as:

$$\boldsymbol{x}_t^{[m]} \sim p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}^{[m]}, \boldsymbol{u}_{s_t}); \tag{1}$$

where $\boldsymbol{x}_{t-1}^{[m]}$ is the pose of the m-th particle at time instant $t-1$. $\boldsymbol{u_s}$ is the pose increment obtained from the difference between the stereo VO measurements between time instances $t-1$ and t given as:

$$\boldsymbol{u_{s_t}} = \ominus \boldsymbol{x}_{s_{t-1}} \oplus \boldsymbol{x}_{s_t} \tag{2}$$

where $\boldsymbol{x}_{s_{t-1}}$ and $\boldsymbol{x}_{s_t}$ are the VO measurements received at time $t-1$ and t respectively. The prediction stage of the filter runs whenever a new VO measurement is received and is based on the composition of each particle's estimate with the obtained increment given by Eq. 2.

2) Update: The update stage of the filter consists of three parts:

- *Update using the IMU measurements:* The roll, pitch and yaw angles of the state predicted by the VO can accumulate error, especially in the absence of characteristic features in the environment. In order to overcome this error, we update the roll, pitch and yaw angles using the angles provided by the IMU. We calculate the importance factor for all the particles as follows:

$$w^{[m]} = \mu \exp\left(-(\phi_{diff} + \theta_{diff} + \psi_{diff})\right) \tag{3}$$

Where,

$$\begin{aligned} \phi_{diff} &= \frac{(\phi_m - \mu_\phi)^2}{2 \cdot \sigma_\phi^2} \\ \theta_{diff} &= \frac{(\theta_m - \mu_\theta)^2}{2 \cdot \sigma_\theta^2} \\ \psi_{diff} &= \frac{(\psi_m - \mu_\psi)^2}{2 \cdot \sigma_\psi^2} \end{aligned} \tag{4}$$

ϕ_m, θ_m and ψ_m are the predicted roll, pitch and yaw angles of each m-th particle respectively. μ_ϕ, μ_θ, μ_ψ are the roll, pitch and yaw angles measured by the IMU. σ_ϕ, σ_θ and σ_ψ being their respective standard deviations and μ a constant normalization factor based on their expected noise standard deviations. In the resampling stage, the roll, pitch and yaw angles of each particle are resampled according to this distribution of the calculated importance weights. This stage is called only when the corresponding IMU data is available.

- *Update using the semantic data:* The semantic data is

received in the form of $\boldsymbol{S}_i = (\boldsymbol{s}^z_{r_i}, s^c_i, s^n_i, s^p_i)$, where $\boldsymbol{s}^z_{r_i}$ is the relative position of the i-th semantic object extracted from Section. III-B, s^c_i being its class type, s^n_i being the number of segmented 3D points and s^p_i being the detection probability. For an accurate data association, we divide the problem into two stages:

- First, for the given semantic data it is checked whether its class label s^c_i equals to any of the class labels of the mapped landmarks $\boldsymbol{l}^c$ and whether its probability s^p_i is higher than a certain threshold. It is further checked whether the number of obtained 3D points s^n_i for the semantic data is greater than a certain threshold.
- Second, if the detected semantic data satisfies the first stage, the relative 3D measurement vector $\boldsymbol{s}^z_i$ composed of x_{r_i}, y_{r_i} and z_{r_i}, is converted to the world frame of reference (see Fig. 1) as follows:

$$\begin{aligned} x_{w_i} &= x_a \oplus x_{r_i} \\ y_{w_i} &= y_a \oplus y_{r_i} \\ z_{w_i} &= z_a \oplus z_{r_i}. \end{aligned} \tag{5}$$

  Where x_a, y_a and z_a are the average of the position measurements obtained from the prediction stage. In order to associate the semantic data with a mapped element, we compute the average difference between the semantic measurements in the world frame obtained from Eq. 5 with the already mapped measurements of the landmarks containing the same class type. If the difference is less than a given threshold, we select the minimum average difference in order to associate the current semantic data with the mapped data. This simplified data association strategy works well in our approach, since the semantic objects are quite far away from each other and the aerial robot position estimate does not have a lot of uncertainty.

Once the detected semantic data is associated with its corresponding mapped landmark k, the importance factor of all the particles is calculated as:

$$w^{[m]} = \frac{\exp\left(-(x_{diff} + y_{diff} + z_{diff})\right)}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \tag{6}$$

where,

$$\begin{aligned} x_{diff} &= \frac{(x^m_{w_i} - x^{map}_{w_k})^2}{2 \cdot \sigma^2} \\ y_{diff} &= \frac{(y^m_{w_i} - y^{map}_{w_k})^2}{2 \cdot \sigma^2} \\ z_{diff} &= \frac{(z^m_{w_i} - z^{map}_{w_k})^2}{2 \cdot \sigma^2} \end{aligned} \tag{7}$$

where $x^m_{w_i}$, $y^m_{w_i}$ and $z^m_{w_i}$ are the positions of the semantic data in the world frame calculated for each predicted m-th particle positions x^m, y^m and z^m obtained as:

$$\begin{aligned} x^m_{w_i} &= x^m \oplus x_{r_i} \\ y^m_{w_i} &= y^m \oplus y_{r_i} \\ z^m_{w_i} &= z^m \oplus z_{r_i} \end{aligned} \tag{8}$$

and $x^{map}_{w_k}$, $y^{map}_{w_k}$ and $z^{map}_{w_k}$ are the positions of the corresponding mapped semantic landmark k in the world frame. σ, is the standard deviation of the expected noise in the detected semantic data distance. Using the given distribution of the calculated importance weights, the positions of all the particles are resampled, converging to particle positions with higher importance factors. This stage is called only when the semantic segmented data is available.

- *Incorporation of new map elements:* New map elements are added in a Maximum Likelihood fashion. If the measurement of the semantic data belonging to a class type does not associate with the current mapped landmarks, due to the average difference being higher than the threshold, the corresponding semantic data is incorporated as a new map element along with its class type. The position of the new map element j is computed as:

$$\begin{aligned} x^{map}_{w_j} &= x_a \oplus x_{r_i} \\ y^{map}_{w_j} &= y_a \oplus y_{r_i} \\ z^{map}_{w_j} &= z_a \oplus z_{r_i}. \end{aligned} \tag{9}$$

IV. Experiments and Results

For a demonstration of the real flight experiments, the reader is advised to refer the following video: `https://vimeo.com/259349563`

A. System Setup

Fig. 1 shows the mini-aerial robot platform named Parrot Bebop-2, used for validating the performance of our proposed approach. We use the IMU on-board the Bebop-2 for the roll, pitch and yaw measurements. Additionally, we use two on-board sensors namely the Parrot-S.L.A.M.dunk and the Intel-RealSense R200 depth camera. The Parrot-S.L.A.M.dunk consists of a low performance ARM-architecture on-board computer along with two fisheye lens cameras. The RGB images from the two cameras, each with a resolution of 640×480 pixels, are utilized for the stereo VO algorithm (Section. III-A). The individual images are acquired at an approximate frequency of 15 Hz. Due to the hardware limitations of the Parrot-S.L.A.M.dunk in computing a decent quality depth image without delay, we use a comparatively accurate Intel-RealSense connected to the Parrot-S.L.A.M.dunk computer, for performing the semantic detection and segmentation explained in Section. III-B. The data from the RealSense is received at 3 Hz. All the images used for computation are compressed and sent along with other required sensor data through WiFi connection between the Bebop-2 to an Intel Core i7-8700K on-ground computer. Our algorithm runs at a real time frequency of 10 Hz. For performing autonomous flights, we embed it into our Aerostack software framework [24], which provides the position estimated by our algorithm to the position controller.

B. Results and Discussions

In order to validate our approach, we perform several autonomous real flight experiments in cluttered and unstruc-

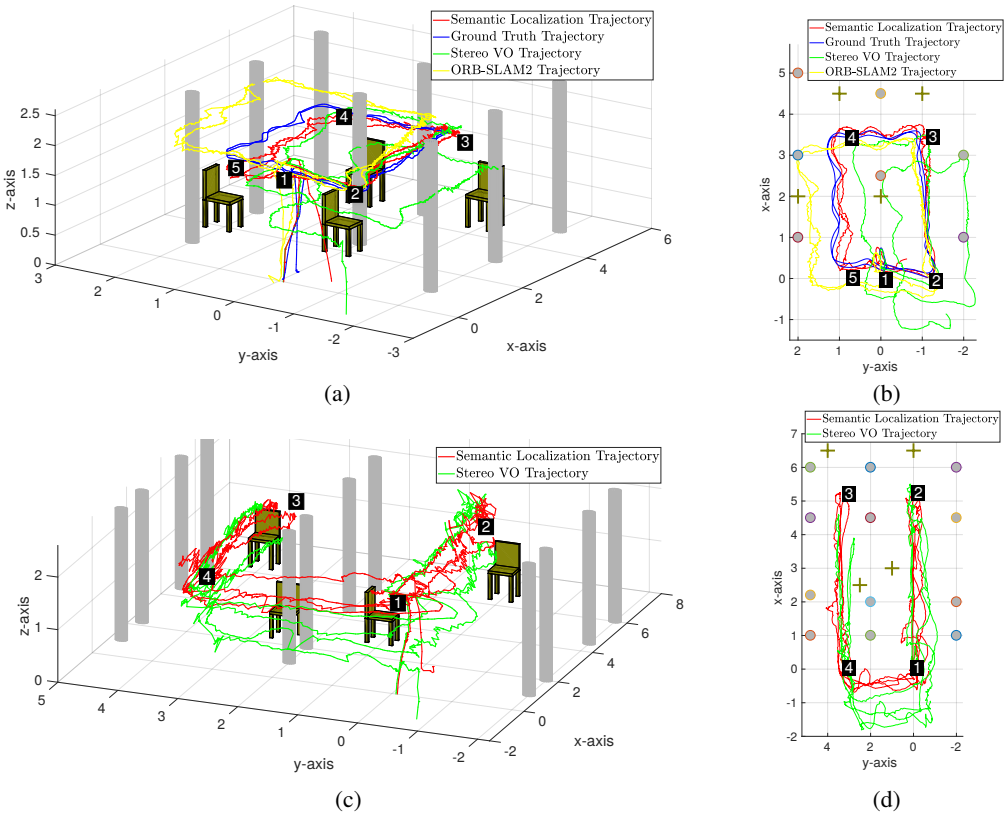

Fig. 2: Results obtained from autonomous real flight experiments of the aerial robot in a cluttered indoor environment. The numbered black boxes represent the order of the commanded 3D trajectory points. In (a) and (b) we present in the 3D and 2D plots for the ground truth evaluation experiment while we show in (c) and (d) the perspective 3D and 2D plot for the long trajectory evaluation experiment. The columns are obstacles for navigation, not included in the localization map.

tured indoor environments. Currently, we only use a single type of semantic data i.e chairs, although this approach can be easily extended to another type of semantic data containing horizontal planar surfaces. We evaluate our proposed algorithm in two different manners:

1) Ground Truth Evaluation: Firstly, we evaluate the performance of our algorithm with an optitrack ground truth system, in a 6 m $\times$ 4 m indoor environment containing several columns, which are solely used for the purpose of adding complexity to the environment in the form of obstacles and are not used in any way by our algorithm. For these experiments we use 4 different types of chairs as semantic landmarks. Fig. 2 depicts the results obtained for the estimation of the trajectory followed by the aerial robot, compared with the ground truth trajectory. The aerial robot follows predefined 3D trajectories in a selected area at an average velocity of 0.2 m/s. In order to demonstrate the robustness of the estimated pose, the trajectory loop is performed twice with a total length of 27.32 m. The sequence of commanded trajectory points can be appreciated in Figures 2a and 2b.

As shown in Fig. 2b, the stereo VO algorithm during the first trajectory loop has little error compared to the ground truth trajectory between points 1 and 3. Between points 3 and 4, as the aerial robot moves closer to a white wall representing a repetitive pattern, the estimated pose begins to accumulate errors. On the other hand, in our approach this accumulated error is corrected whenever a detection of a chair along with the 3D position of the segmented horizontal plane is received. As it can be observed in Fig. 2a, even though the environment contains several obstacles in the form of columns, the accurate position estimated by the algorithm enables the aerial robot to maintain the desired trajectory during the entire experiment.

We compare our algorithm, with the state of the art ORB-SLAM2 stereo visual SLAM algorithm [25]. Although ORB-SLAM2 uses only a stereo image pair provided by the Parrot-S.L.A.M.dunk, as compared to our method which additionally uses an Intel-RealSense, this comparison presents the data association errors of methods depending on low level characteristic features of the environment, as apposed to our method depending on high level semantic cues. ORB-SLAM2 maintains an accurate trajectory along the x and y-axis between trajectory points 1 and 4, whereas between points 3 and 4, consisting a repetitive pattern, the data association using low level characteristic features degrades and the errors in the estimated position increases (see Figures 2a and 2b). Table. I compares the absolute trajectory errors [26] obtained during the execution of this experiment.

2) Long Trajectories Evaluation: The purpose of this experiment is to validate the performance of our algorithm when the aerial robot has to traverse long trajectories in

TABLE I: Absolute Trajectory Error for the compared algorithms.

Error	Our approach	Stereo VO	Stereo ORB-SLAM2
ATE $[m]$	0.18	0.85	0.39

larger spaces. We perform the experiment in a larger cluttered indoor environment of 8 m $\times$ 7 m, also containing similar obstacles as in the previous experiment and 4 chairs as semantic landmarks. The aerial robot performs the 3D trajectory loop twice, spanning an approximate trajectory length of 57.8 m at an average velocity of 0.3 m/s. It can be seen from Figures 2c and 2d, that the aerial robot accurately reaches the commanded trajectory points and returns back close to the initial take-off point 1, whereas the pose estimated by the VO accumulates drift.

V. Conclusions and Future Work

In this paper, we have presented a particle filter based localization approach for mini-aerial robots, using stereo VO and semantic information from cluttered and unstructured indoor environments. We presented a novel technique to accurately estimate the relative position of the detected semantic objects in presence of noisy depth measurements, by segmenting the dominant horizontal planar surfaces, generalizing for several types of objects irrespective of their structure. We tested our algorithm in such indoor scenarios using a common class of semantic data i.e chairs, of different shapes and sizes, providing the real time pose information of the aerial robot at 10 Hz. We validated our approach by comparing the estimated pose with the ground truth data as well as comparing its performance with a state of the art visual SLAM algorithm in challenging conditions.

Even though the relative position estimation of the semantic data is accurate, it can suffer from inaccuracies when its horizontal surface is partially occluded. Hence, as future work, we plan to improve the relative position estimation of the segmented semantic data by obtaining the convex hull of the surfaces and we also plan to extract the relative orientation from the data. Furthermore, we plan to implement a full semantic SLAM based approach for aerial robots without using prior mapped landmarks.

References

[1] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 1352–1359.

[2] J. L. Sanchez-Lopez, J. Pestana, P. de la Puente, R. Suarez-Fernandez, and P. Campoy, "A system for the design and development of vision-based multi-robot quadrotor swarms," in *2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, May 2014, pp. 640–648.

[3] P. de la Puente and D. Rodriguez-Losada, "Feature based graph SLAM with high level representation using rectangles," *Robotics and Autonomous Systems*, vol. 63, pp. 80 – 88, 2015.

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec 2016.

[5] J. G. Rogers, A. J. B. Trevor, C. Nieto-Granda, and H. I. Christensen, "Simultaneous localization and mapping with learned object recognition and semantic data association," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2011, pp. 1264–1270.

[6] A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 3515–3522.

[7] W.-C. Ma, S. Wang, M. A. Brubaker, S. Fidler, and R. Urtasun, "Find your way by observing the sun and other semantic cues," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6292–6299, 2017.

[8] C. Toft, C. Olsson, and F. Kahl, "Long-term 3d localization and pose from semantic labellings," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 650–659.

[9] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 1722–1729.

[10] M. Guenther, T. Wiemann, S. Albrecht, and J. Hertzberg, "Model-based furniture recognition for building semantic object maps," *Artificial Intelligence*, vol. 247, pp. 336–351, 2017.

[11] J.-R. Ruiz-Sarmiento, C. Galindo, and J. Gonzalez-Jimenez, "Building multiversal semantic maps for mobile robot operation," *Knowledge-Based Systems*, vol. 119, pp. 257 – 272, 2017.

[12] V. Murali, H.-P. Chiu, S. Samarasekera, and R. Kumar, "Utilizing semantic visual landmarks for precise vehicle navigation," *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–8, 2017.

[13] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86 – 103, 2015.

[14] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. D. Reid, "Meaningful maps with object-oriented semantic mapping," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5079–5085, 2017.

[15] A. Gawel, C. D. Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multiview localization," *IEEE Robotics and Automation Letters*, vol. 3, pp. 1687–1694, 2018.

[16] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Localization from semantic observations via the matrix permanent," *I. J. Robotics Res.*, vol. 35, pp. 73–99, 2016.

[17] L. Wang, L. Zhao, G. Huo, R. Li, Z. Hou, P. Luo, Z. Sun, K. Wang, and C. Yang, "Visual semantic navigation based on deep learning for indoor mobile robots," *Complexity*, vol. 2018, 2018.

[18] F. S. Amirmasoud Ghasemi Toudeshki and R. Vaughan, "Uav visual teach and repeat using only semantic object features," *arXiv preprint arXiv:1801.07899*, 2018.

[19] D. Maravall, J. de Lope, and J. P. Fuentes, "Navigation and self-semantic location of drones in indoor environments by combining the visual bug algorithm and entropy-based vision," *Frontiers in Neurorobotics*, vol. 11, p. 46, 2017.

[20] S. Manthe, A. Carrio, F. Neuhaus, P. Campoy, and D. Paulus, "Combining 2d to 2d and 3d to 2d point correspondences for stereo visual odometry," in *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, 2018, pp. 455–463.

[21] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artificial intelligence*, vol. 78, no. 1-2, pp. 87–119, 1995.

[22] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016.

[23] S. Holzer, R. B. Rusu, M. Dixon, S. Gedikli, and N. Navab, "Adaptive neighborhood selection for real-time surface normal estimation from organized point cloud data using integral images," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 2684–2689.

[24] J. L. Sanchez-Lopez, M. Molina, H. Bavle, C. Sampedro, R. A. Suárez Fernández, and P. Campoy, "A multi-layered component-based approach for the development of aerial robotic systems: The aerostack framework," *Journal of Intelligent & Robotic Systems*, vol. 88, no. 2, pp. 683–709, Dec 2017.

[25] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[26] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 573–580.