2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Active Perception for Outdoor Localisation with an Omnidirectional Camera

Maleen Jayasuriya[1], Ravindra Ranasinghe[1], Gamini Dissanayake[1]

Abstract— **This paper presents a novel localisation framework based on an omnidirectional camera, targeted at outdoor urban environments. Bearing only information to persistent and easily observable high-level semantic landmarks (such as lamp-posts, street-signs and trees) are perceived using a Convolutional Neural Network (CNN). The framework utilises an information theoretic strategy to decide the best viewpoint to serve as an input to the CNN instead of the full $360°$ coverage offered by an omnidirectional camera, in order to leverage the advantage of having a higher field of view without compromising on performance. Environmental landmark observations are supplemented with observations to ground surface boundaries corresponding to high-level features such as manhole covers, pavement edges and lane markings extracted from a second CNN. Localisation is carried out in an Extended Kalman Filter (EKF) framework using a sparse 2D map of the environmental landmarks and Vector Distance Transform (VDT) based representation of the ground surface boundaries. This is in contrast to traditional vision only localisation systems that have to carry out Visual Odometry (VO) or Simultaneous Localisation and Mapping (SLAM), since low level features (such as SIFT, SURF, ORB) do not persist over long time frames due to radical appearance changes (illumination, occlusions etc) and dynamic objects. As the proposed framework relies on high-level persistent semantic features of the environment, it offers an opportunity to carry out localisation on a prebuilt map, which is significantly more resource efficient and robust. Experiments using a Personal Mobility Device (PMD) driven in a representative urban environment are presented to demonstrate and evaluate the effectiveness of the proposed localiser against relevant state of the art techniques.**

I. INTRODUCTION

Localisation remains one of the most fundamental and challenging tasks for an autonomous vehicle. Among the sensor modalities that aid in localisation, vision sensors stand out as low-cost compact sources that offer rich information about the operating environment of a mobile robot [1]. This paper presents an outdoor vision based localisation framework, targeted at low speed (under 15 Kmph), resource constrained autonomous vehicles operating in urban and suburban pedestrian environments such as footpaths and pavements. Typical examples range from Personal Mobility Devices (PMDs) such as mobility scooters and powered wheelchairs, to delivery robots. There is a significant interest in incorporating self driving capability to PMDs for improving their safety and efficacy. The framework proposed in this paper has been implemented and evaluated on an instrumented mobility scooter platform described in [2].

Vision based systems operating in outdoor environments need to deal with the challenges caused by appearance changes (illumination, occlusions) and the presence of dynamic objects. Traditionally this has been addressed by resorting to Simultaneously Localisation and Mapping (SLAM) or pure Visual Odometry (VO) which do not require the environment representation to be invariant. However, this is at the expense of a significant increase in computational cost and decrease in robustness. Comparatively, given a suitable sensor and a representation of the environment that does not change with time, localisation on a prebuilt map is a relatively straightforward task. Thus, the proposed framework aims to address this by focusing on high-level semantic observations of environmental landmarks (street lamps, tree trunks, parking meters etc) and ground surface boundary observations (pavement edges, manhole covers etc) obtained from state of the art Convolutional Neural Networks (CNN). Localisation is carried out on a prebuilt sparse 2D map of the landmark features and Vector Distance Transform (VDT) representation of ground surface boundaries. The work presented in this paper builds on our previous work [2], [3] by leveraging an omnidirectional camera and an active vision paradigm to improve robustness and performance.

The use of an omnidirectional camera overcomes an inherent limitation most vision based localisation approaches face due to the limited field of view (FoV) of conventional cameras, in comparison to the blanket coverage that sensors such as LIDARs offer. This is specially crippling in highly dynamic environments such as outdoor urban settings where significant portions of the available FoV maybe occluded [4]. However, utilising omnidirectional cameras for localisation also introduces it's own set of challenges. In particular, a high resolution sensor is required to counteract the impact of the large field of view on the environmental features resulting in an increase in the computational cost. An active perception paradigm [5] to determine when and where to make observations is proposed to deal with this issue. An information gain based metric is employed to select the portion of the image that is to be processed for feature extraction instead of the entire 360 image, in order to provide the best trade off between processing limitations and localisation accuracy. The contributions of this paper are:

- A localisation framework that takes advantage of omnidirectional cameras through CNN based perception of high-level semantic information of the environment.
- An information gain based active vision paradigm for perspective viewpoint selection to improve the efficacy

[1] Maleen Jayasuriya, Ravindra Ranasinghe and Gamini Dissanayake are with the Centre for Autonomous Systems (CAS), University of Technology Sydney, Australia.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

and functionality of the proposed framework.

- A detailed evaluation, comparison (against ORB-SLAM2 [6], VINS [7] and RTAB [8]) and demonstration of the proposed system based on real world experiments carried out using an instrumented PMD.

The remainder of this paper is structured as follows: Section II discusses the state of the art related to the work presented in this paper. Section III details the proposed localisation framework. Section IV provides a brief overview of the configuration of the hardware platform that was used for evaluation. Section V presents an evaluation of the proposed system on real world data, including a comparison with state of the art open source visual localisation schemes. Finally, Section VI offers a brief discussion on the results and proposed future work.

II. RELATED WORK

The past decade has witnessed enormous progress in the field of vision based localisation. Feature based approaches that first extract low level handcrafted features such as SIFT, SURF or ORB features in order to carry out pose estimation, such as ORB-SLAM [6], [9], VINS [7], Kimera [10] and RTAB [8] are noteworthy state of the art contributions. Direct and semi-direct appearance based approaches that take into account pixel intensities of the entire image for motion estimation such as LSD-SLAM [11] and DSO [12] have also proven to be remarkably effective. However, all these techniques rely on generating a location estimate while at the same time building a local or global representation of the environment, significantly adding to the computational effort required. Furthermore, these rely on "loop closure" which is essential to avoid the inevitable drift in the location estimate, which is challenging in outdoor urban settings due to dynamic objects and appearance changes [13] [14].

Some of these challenges can be dealt with by extending the above techniques to accommodate omnidirectional cameras with the expectation that the large FoV will offer better feature tracking and robustness to occlusion [4]. Notable examples include both feature based techniques [15]–[18] and direct techniques [19]–[22]. Complications associated with using omnidirectional cameras include large image distortions and low pixel resolutions which make calibration and selecting the appropriate camera models and image projections of paramount importance [23].

Another approach to tackle some of these challenges is to leverage high-level semantic information to aid in the localisation task. For instance, dynamic objects such as cars are recognised and rejected in work presented in [22], [24], [25]. However, the localisation task can be made significantly simpler and robust if a map of the environment that can be reused as and when required, can be prebuilt. For instance most autonomous cars rely on prebuilt point cloud based high definition 3D maps [26] in order to achieve the robustness and accuracy required for a fast moving vehicle [1], [27]. However, constructing and maintaining such maps demand expensive sensors and high resources in terms of data collection, storage, and processing [28]. Alternatively, Xaio et. al. [29] extracts landmarks such as poles, street signs and traffic lights obtained through a semantic segmentation process to match these against a 3D map of such features. Authors in [30] use a combination of LIDARs and cameras to observe similar high-level features as well as road-markings and facades to localise on a 3D map.

Our previous work [3] proposed an Extended Kalman Filter (EKF) based localisation scheme that fuses CNN based object detection of common environmental landmarks and ground surface boundaries for localisation on a sparse 2D map. However, the sparse nature of the bearing only environmental landmark observations is a major limitation to the robustness and accuracy of the framework presented in [3]. The work presented in this paper, proposes to use an omnidirectional camera to overcome this challenge. Although CNN based semantic segmentation and object detection schemes that directly apply to distorted fish-eye images have been attempted [31]–[33], these do not match the real-time performance and accuracy required to carry out visual localisation. Comparatively, the YOLO CNN object detection framework [34], [35] which was used in our previous work, offers state of the art performance and accuracy. Although it is possible to re-project the image from a fish-eye camera into multiple perspective images that are suitable for YOLO, using the set of all re-projected images incurs a significant computational cost and therefore is not suitable in a resource constrained platform. Hence an active vision paradigm is proposed to maintain an efficient processing pipeline for YOLO, in order to perform real-time localisation on our mobility scooter platform.

Active perception that directs the sensors on board a mobile platform in order to improve localisation performance has been proposed for use with sonar [36], LIDAR [37] and vision. Active vision is generally achieved through some form of visual servoing of cameras mounted on mechanical pan-tilt units [38]–[41] or selecting optimal navigational trajectories that maximise information gathered [42]. Metrics based on the trace of the covariance matrix [43], the D-optimality criterion [44], entropy [45] and fisher information [46] have been used to make active perception decisions. In this work, active panning of a virtual perspective camera to select the best viewpoint within a 360° spherical image is proposed. The trace of the covariance matrix is used as a metric to determine this optimal viewpoint.

III. LOCALISATION FRAMEWORK

A. Overview

The overall framework (see figure 1) consists of a CNN aided perceptual front end that consists of environmental landmark observations obtained through YOLO [34] and ground surface boundary observations obtained via HED [47]. Observations for the environmental landmark model which is the focus of this work, are obtained using an omnidirectional camera. Bearing only observations to these detected landmarks are then fused with the VDT based ground surface observations using an Extended Kalman Filter

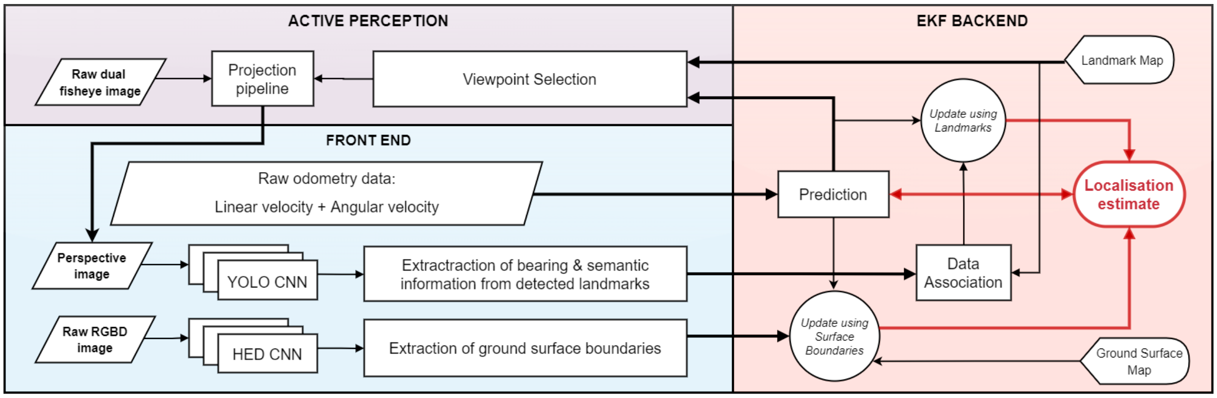
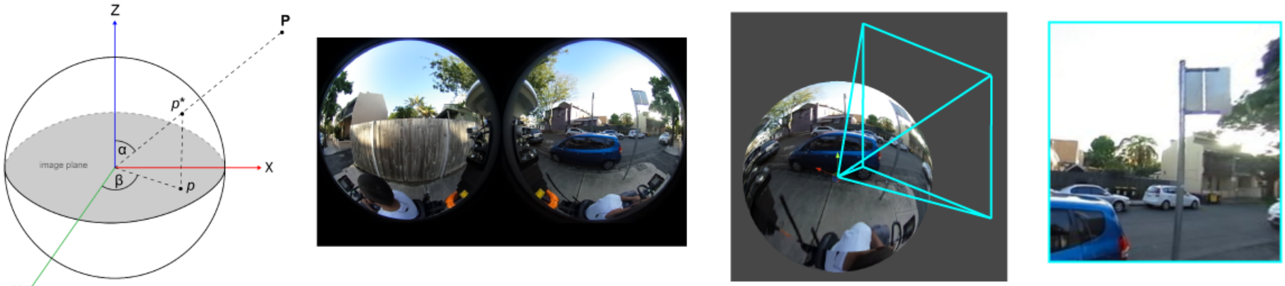


Fig. 1: Overview of Framework



(a) Fisheye camera model

(b) Raw dual-fisheye image from omnidirectional camera

(c) Unit sphere projection with virtual pinhole camera

(d) Perspective image

Fig. 2: Projection Pipeline

(EKF) back-end in order to localise on a sparse 2D map of the environment.

B. Omnidirectional perception

Omnidirectional cameras include catadioptric sensors, multiple camera rigs with 360° coverage, as well as dual-fisheye lenses. In principle, the proposed framework is applicable to all of the above omnidirectional sensor types, subject to the availability of an appropriate camera models and image projections. The following section is described based on a dual-fisheye lens setup due to the wide availability of low-cost consumer grade dual-fisheye 360° cameras and its use in the operational hardware platform (see section IV).

Dual-fisheye lenses consist of two fisheye lenses that each cover a FoV of 180° horizontally and vertically. The fisheye camera model is based on spherical projection. In the case of a single fisheye lens, a 3D point $P = [X_p, Y_p, Z_p]$ in space (refer figure 2a), is projected onto a sphere of radius f_s at point p^*. This describes the mapping between real-world spatial points, to points on the surface of a sphere given by:

$$p^* = [f_s \cdot sin\alpha \cdot sin\beta, f_s \cdot sin\alpha \cdot cos\beta, f_s \cdot cos\alpha] \quad (1)$$

Thus, $p^* = \lambda P$ where $\lambda = f_s / \sqrt{X_p^2 + Y_p^2 + Z_p^2}$. p^* can be projected onto the 2D image plane at point p given by equation:

$$p = [f_s \cdot sin\alpha \cdot sin\beta, f_s \cdot sin\alpha \cdot cos\beta, 0] \quad (2)$$

In the case of two fisheye lenses that provide 360° coverage, a dual-circular fisheye image (one for each lens) is obtained as in figure 2b. This is the format of the raw images obtained from a dual fisheye camera.

Reprojecting the dual-fisheye image onto a spherical projection with $f_s = 1$, gives an image that behaves as a scaled down version of the 3D spatial information of the real world, while preserving the bearings (see 2c). Sections of this unit sphere can then be projected from the centre of the sphere onto an image plane using a pinhole camera model to obtain undistorted perspective images (see figure 2d). These images are equivalent to images obtained from a regular perspective camera inside a scaled down version of the world.

C. Feature extraction with a CNN

CNN based object detection framework YOLO provides a convenient way to extract persistent and easily observable landmarks in pedestrian environments such as road-signs, traffic lights, parking meters and trees. However, YOLO is designed to be used with perspective images. Retraining YOLO with a curated custom dataset of circular fisheye images of landmarks collected in an around the Sydney City area was unsuccessful as the detection accuracy was poor. The detection rate was also poor due to the large resolution of the dual fisheye images making it untenable for real time operation. Resizing the images to speed up the process resulted in further reduction in detection accuracy due to the landmarks being too small relative to the whole image. Similar results with CNN based object detection on fisheye images are reported in [48].

Use of the undistorted perspective images as seen in figure 2d resulted in detection performance comparative to that obtained using a conventional camera. However, for real-time performance on the current hardware platform only a projected perspective window representing a 416x416 px image with a horizontal and vertical FoV of 60° could be processed in a single iteration. This problem is addressed by our active vision strategy described in section III-E.

D. EKF backend

The EKF framework used for location estimation is described in detail in [3]. The relevant equations are presented below for completeness and later used for describing the active vision strategy described in section III-E. The predicted pose estimate $\hat{\boldsymbol{X}}_{\boldsymbol{t|t-1}}$ is calculated using the generic motion model of the system (given by equation 3), based on odometry information at time t, U_t:

$$\hat{\boldsymbol{X}}_{\boldsymbol{t|t-1}} = g(\hat{\boldsymbol{X}}_{\boldsymbol{t-1}}, U_t) \tag{3}$$

The associated prediction covariance $P_{t|t-1}$ is given by

$$P_{t|t-1} = \nabla G_x \cdot P_{t-1} \cdot \nabla G_x^\top + \nabla G_u \cdot Q \cdot \nabla G_u^\top \tag{4}$$

where the ∇G_x and ∇G_u are the jacobians of the motion model with respect to X and U. Q is the noise in odometry measurements.

The YOLO framework produces bearings $\Theta_t = [\theta_t^1, \theta_t^2, ..., \theta_t^n]$ and the associated semantic label vector $L_t = [l_t^1, l_t^2, ..., l_t^n]$ after processing a given perspective image.

Each observation (θ_t^i, l_t^i) is associated with a landmark located on the map M_L, using an innovation gate based on the Mahalanobis distance. The generic landmark based observation model h is given by:

$$\hat{\Theta}_t = h(\hat{\boldsymbol{X}}_{\boldsymbol{t|t-1}}, M_L) \tag{5}$$

The innovation can then be calculated as:

$$\nu_l = \hat{\Theta}_t - \Theta_t \tag{6}$$

and corresponding innovation covariance S_l by:

$$S_l = R + \nabla H_x \cdot P_{t|t-1} \cdot \nabla H_x^\top \tag{7}$$

where R is the bearing measurement noise. ∇H_x is the Jacobian of the observation model with respect to X.

The Kalman gain K_l is then calculated as:

$$K_l = P_{t|t-1} \cdot \nabla H_x^\top \cdot S_l^{-1} \tag{8}$$

Then the pose estimate and covariance are updated using:

$$\begin{aligned} \hat{\boldsymbol{X}}_{\boldsymbol{t}} &= \hat{\boldsymbol{X}}_{\boldsymbol{t|t-1}} + K_l \cdot \nu_l \\ P_t &= P_{t|t-1} - K_l \cdot S_l \cdot K_l^\top \end{aligned} \tag{9}$$

A similar EKF update is carried out based on ground surface observations obtained using the HED neural network. This is implemented using a 2D Vector Distance Transform representation of a binary image map of ground surface boundaries. A detailed explanation of this can be found in our previous work [3].

E. Active vision

Considering the sparse nature of environmental landmark observations in a typical urban scene, consistent tracking of important landmarks over a longer time period is more important than observing the whole surroundings at a given moment. Thus an active vision approach was developed to select the best view point for projecting a perspective image for YOLO to process, in order to achieve optimal localisation results. The active view point selection process was limited to a selection between four possible fixed viewpoints. These were selected to avoid the camera mounts and the highly distorted seam where the two 180° images are stitched together (see viewpoints: A,B,C,D in figure 3). These images therefore provide a coverage of 240°. Raw dual-fisheye images from the camera was processed using the Unity 3D physics engine [49] to obtain the perspective images using a virtual pinhole camera placed in the unit sphere projection. At each iteration of the EKF at time t, a predicted pose estimate $\hat{\boldsymbol{X}}_{\boldsymbol{t|t-1}}$ and it's associated covariance $P_{t|t-1}$ are calculated based on the motion model described in equation 3. This pose estimate is fed into the observation model h (See equation 5) along with the 2D map of the landmark locations, to generate a set of predicted bearing observations for each possible camera view point. Based on these predicted observations, the EKF covariance update (See equation 9) is carried out for each of the possible camera viewpoints which generates 4 possible predicted covariances for each viewpoint: $\hat{P}_t^{vp}$ where $vp = A, B, C, D$.

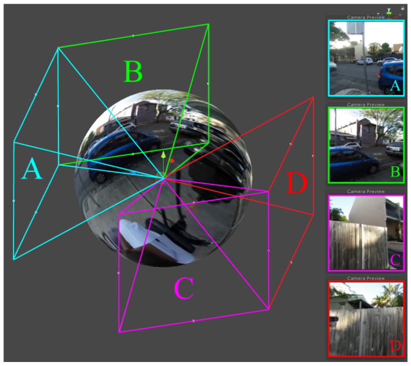


Fig. 3: Virtual perspective camera viewpoints

In order to quantify the information related to each possible viewpoint, the trace of these covariance matrices $tr(\hat{P}_t^{vp})$, which provides a strong measure of the pose uncertainty is calculated. The metric, ΔP^{vp} that quantifies the impact on the overall pose uncertainty from each viewpoint (A,B,C,D) is then calculated by equation 10:

$$\Delta P^{vp} = tr(P_{t|t-1}) - tr(\hat{P}_t^{vp}) \tag{10}$$

This provides a straightforward scoring system for each viewpoint. The maximum reduction in the trace of the covariance (highest ΔP^{vp}) corresponds to the viewpoint that provides the most information required for better localisation performance. Once the best viewpoint is selected based on this scoring system, a perspective projection corresponding to that viewpoint from within the unit sphere is captured and sent to the YOLO object detection framework. Bearing information from successful detections are then sent to the EKF back-end described in section III-D.

To deal with scenarios where a landmarks maybe occluded due to environment dynamics, the system uses a heuristic to switch to the viewpoint with second highest score when no detections are reported from the selected best view point.

IV. HARDWARE SYSTEM OVERVIEW

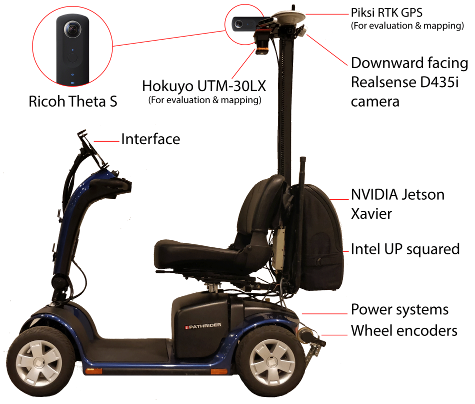


Fig. 4: Hardware overview of retrofitted mobility scooter

The hardware platform (Figure 4) comprises of the Pride Pathrider 10 mobility scooter instrumented with a low cost computation and sensor package (See [2], [3] for a detailed description). A Ricoh Theta S, omnidirectional dual fisheye camera is mounted on the platform and acts as the primary visual sensor for the environment landmark observations. For evaluation and map building purposes, the scooter is also equipped with a Piksi Multi Real Time Kinematic (RTK) GPS unit and Hokuyo UTM-30LX 2D laser.

A. Calibration

Intrinsic and extrinsic calibration of the cameras were carried out with the aid of an opti-track system and motion capture markers placed on a calibration target. The virtual pin-hole camera was treated as if it were a physical camera placed at each view point A,B,C,D as seen in figure 3. Four discrete viewpoints were selected instead of having a continuous virtual panning camera to ease the intrinsic and extrinsic calibration process.

V. EXPERIMENTAL RESULTS & DISCUSSION

A. Results

This section describes the experimental evaluation of the proposed framewsork based on real world data obtained by driving the mobility scooter platform (described in section IV) in a $\sim$8000m^2 area in Glebe, Sydney, Australia. The terrain was representative of a typical suburban pedestrian setting. A 2D map of the environmental landmarks and a VDT representation of ground surface boundaries were constructed based on known vehicle poses. These known poses were obtained by conducting careful mapping missions using the Real-Time Appearance-Based Mapping (RTAB) RGBD and LIDAR Graph-Based SLAM framework [8]. Mapping involved running RTAB on data collected in tightly controlled small loops, within short time frames (typically under one hour) during times of the day when the environment remained relatively static. These mapping missions were conducted roughly 6 months prior to the data collected for localisation.

Experimental evaluation of the proposed system is presented along with a comparison of the results obtained from two state of the art open source visual SLAM systems; ORBSLAM2 [6] and VINS-Fusion [7], [50]. Both the above techniques are run in stereo mode (in order to recover scale) using the mounted Realsense D435i. Additionally, results obtained from RTAB [8] based on the the RGBD data (from Realsense D435i) and laser data (from Hokuyo UTM-30LX) post graph based optimisation are also presented.

TABLE I: Root Mean Square Errors

Technique	RMSE: [m]
Proposed (Active vision)	0.11
Proposed (Unconstrained)	**0.11**
VINS	0.65
ORBSLAM2	1.35
RTAB	4.56

In order to ascertain the impact of the active vision strategy, results obtained by assuming no processing limitations is also presented. This was generated by feeding all 4 available viewpoints to YOLO in an unconstrained manner with no processing and time delay limitations considered.

Ground truth for this evaluation was obtained using RTK GPS. RTK GPS results are generally provided in 3 levels of precision. RTK fixed, RTK float, SPP (Single Point Positioning-acts as regular GPS) in descending order of precision. Only RTK-fixed provides centimetre level accuracy and was used during the error calculation. However, continuous RTK fixed GPS results are difficult to come by within a large urban environment. Thus, quantitative error calculations are carried out in sporadic locations along the trajectory of the mobility scooter whenever RTK fixes are available. Table I reports the root mean square error (RMSE) of all evaluated techniques. A qualitative analysis of the trajectories are also presented in figure 5.

B. Discussion

Results show that the accuracy of the proposed active vision strategy is comparable to that of the unconstrained processing of the full 360° FoV provided by the omnidirectional camera (see table I). This shows that in this particular environment, the active vision strategy is successful in taking

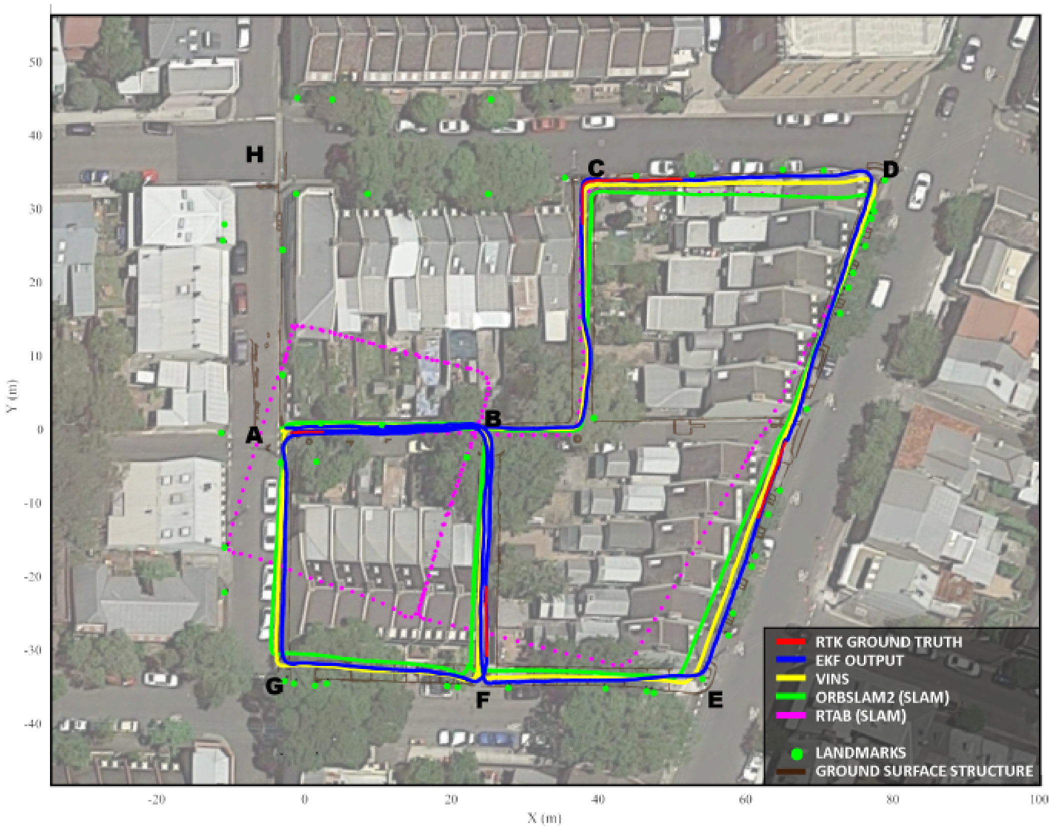


Fig. 5: Localisation trajectories

full advantage of the available field of view while balancing performance accuracy with processing limitations. This is perhaps because the landmarks in this particular environment are sparse enough that only choosing a particular viewpoint instead of processing the whole image has a negligible effect on accuracy. This might not be the case in a different environment with a dense number of landmarks spread evenly within a 360° view, in which case we posit that processing the entire 360° FoV although impossible in real time on the current platform, may yield better results.

The performance of ORB-SLAM2, VINS and RTAB is indicative of the inherent drift associated with SLAM systems, specially in highly dynamic outdoor environments. The scooter was driven along the path in the order A-B-C-D-E-F-B-A-G-F-B-A (see figure 5. It is clear that without any loop closure, the techniques under comparison start to show drift by D-E. However, when allowed to complete loop closure (See loop F-B-A-G-F) all techniques perform well, but the accumulated drift is not corrected since this forms a disconnected drifted loop.

The proposed framework performs better as localisation is carried out on a prebuilt map of high-level semantic landmarks and ground surface boundaries. Although the sparse nature of the landmarks is addressed by the use of an omnidirectional camera, a potential failure scenario of the proposed framework is locations with no observable environmental landmarks or ground surface information. Another possible but less probable failure scenario may occur if radical changes are made to the landmarks in the environment.

VI. FUTURE WORK AND CONCLUSION

In terms of future work we hope to explore the fusion of absolute global location information obtained through GPS into the EKF framework, specially to aid in initialisation. Furthermore, the fusing of VO based odometry will also be explored to aid in locations with little to no landmark features. A long term map maintenance strategy will also be considered in the future in order to account for changes that could potentially occur in the environment.

In conclusion, the proposed framework aims to offer a vision based localisation framework that can function successfully on a prebuilt map of high-level semantic features. A major challenge faced due to the sparse nature of landmarks is addressed using an omnidirectional camera and processing limitations are tackled using an active vision strategy to offer a low-cost, resource efficient outdoor vison based localisation framework.

ACKNOWLEDGMENT

This work is supported by the Centre for Autonomous System at University of Technology Sydney. Special thanks to Mr. Janindu Arukgoda, Mr. Nathanael Gandhi and Mr. Peter Morris for their assistance with the presented work.

References

[1] S. Kuutti, S. Fallah, K. Katsaros, M. Dianati, F. Mccullough, and A. Mouzakitis, “A Survey of the State-Of-The-Art Localization Techniques and Their Potentials for Autonomous Vehicle Applications,” *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 829–846, Apr. 2018.

[2] M. Jayasuriya, G. Dissanayake, R. Ranasinghe, and N. Gandhi, “Leveraging Deep Learning Based Object Detection for Localising Autonomous Personal Mobility Devices in Sparse Maps,” in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Oct. 2019, pp. 4081–4086.

[3] M. Jayasuriya, J. Arukgoda, R. Ranasinge, and G. Dissanayake, “Localising PMDs through CNN Based Perception of Urban Streets,” in *(Accepted) 2020 International Conference on Robotics and Automation (ICRA)*, 2020. [Online]. Available: https://www.researchgate.net/publication/339615610_Localising_PMDs_through_CNN_Based_Perception_of_Urban_Streets

[4] L. Payá, A. Gil, and O. Reinoso, “A State-of-the-Art Review on Mapping and Localization of Mobile Robots Using Omnidirectional Vision Sensors,” *Journal of Sensors*, vol. 2017, pp. 1–20, 2017.

[5] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, “Revisiting active perception,” *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, Feb. 2018.

[6] 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, Oct. 2017.

[7] T. Qin, P. Li, and S. Shen, “VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator,” *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[8] M. Labbe and F. Michaud, “Rtab-Map as an Open-Source Lidar and Visual Simultaneous Localization and Mapping Library for Large-Scale and Long-Term Online Operation,” *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, Mar. 2019.

[9] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “ORB-SLAM: A Versatile and Accurate Monocular SLAM System,” *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[10] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, “Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping,” *arXiv:1910.02490 [cs]*, Dec. 2019, arXiv: 1910.02490. [Online]. Available: http://arxiv.org/abs/1910.02490

[11] J. Engel, T. Schöps, and D. Cremers, “LSD-SLAM: Large-Scale Direct Monocular SLAM,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, vol. 8690, pp. 834–849.

[12] J. Engel, V. Koltun, and D. Cremers, “Direct Sparse Odometry,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[13] R. Zekavat and R. Buehrer, “Localization for Autonomous Driving,” in *Handbook of Position Location: Theory, Practice, and Advances*. IEEE, 2019, pp. 1051–1087.

[14] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, “Visual Place Recognition: A Survey,” *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[15] Z. Cao, S. Liu, and J. Roning, “Omni-directional Vision Localization Based on Particle Filter,” in *Fourth International Conference on Image and Graphics (ICIG 2007)*. Chengdu, Sichuan, China: IEEE, Aug. 2007, pp. 478–483.

[16] M. Ramezani, K. Khoshelham, and L. Kneip, “Omnidirectional visual-inertial odometry using multi-state constraint Kalman filter,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, BC: IEEE, Sep. 2017, pp. 1317–1323.

[17] M. G. Muller, F. Steidle, M. J. Schuster, P. Lutz, M. Maier, S. Stoneman, T. Tomic, and W. Sturzl, “Robust Visual-Inertial State Estimation with Multiple Odometries and Efficient Mapping on an MAV with Ultra-Wide FOV Stereo Vision,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Madrid: IEEE, Oct. 2018, pp. 3701–3708.

[18] H. Seok and J. Lim, “ROVO: Robust Omnidirectional Visual Odometry for Wide-baseline Wide-FOV Camera Systems,” in *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 6344–6350.

[19] D. Caruso, J. Engel, and D. Cremers, “Large-scale direct SLAM for omnidirectional cameras,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Hamburg, Germany: IEEE, Sep. 2015, pp. 141–148.

[20] P. Liu, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys, “Direct visual odometry for a fisheye-stereo camera,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, BC: IEEE, Sep. 2017, pp. 1746–1752.

[21] H. Matsuki, L. von Stumberg, V. Usenko, J. Stuckler, and D. Cremers, “Omnidirectional DSO: Direct Sparse Odometry With Fisheye Cameras,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3693–3700, Oct. 2018.

[22] Z. Cui, L. Heng, Y. C. Yeo, A. Geiger, M. Pollefeys, and T. Sattler, “Real-Time Dense Mapping for Self-Driving Vehicles using Fisheye Cameras,” in *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 6087–6093.

[23] Zichao Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, “Benefit of large field-of-view cameras for visual odometry,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. Stockholm, Sweden: IEEE, May 2016, pp. 801–808.

[24] V. Murali, H. Chiu, S. Samarasekera, and R. T. Kumar, “Utilizing Semantic Visual Landmarks for Precise Vehicle Navigation,” in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Oct. 2017, pp. 1–8.

[25] B. Bescos, J. M. Facil, J. Civera, and J. Neira, “DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4076–4083, Oct. 2018.

[26] B. Templeton, “Many Different Approaches to Robocar Mapping,” 2017. [Online]. Available: http://robohub.org/many-different-approaches-to-robocar-mapping/

[27] F. Poggenhans, J. Pauls, J. Janosovits, S. Orf, M. Naumann, F. Kuhnt, and M. Mayr, “lanelet2: A High-Definition Map Framework for the Future of Automated Driving,” in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov. 2018, pp. 1672–1679.

[28] H. G. Seif and X. Hu, “Autonomous driving in the iCity: HD maps as a key challenge of the automotive Industry,” *Engineering*, vol. 2, no. 2, pp. 159–162, Jun. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2095809916309432

[29] Z. Xiao, K. Jiang, S. Xie, T. Wen, C. Yu, and D. Yang, “Monocular Vehicle Self-Localization Method Based on Compact Semantic Map,” in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov. 2018, pp. 3083–3090.

[30] J. Kummerle, M. Sons, F. Poggenhans, T. Kuhner, M. Lauer, and C. Stiller, “Accurate and Efficient Self-Localization on Roads using Basic Geometric Primitives,” in *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 5965–5971.

[31] L. Deng, M. Yang, Y. Qian, C. Wang, and B. Wang, “CNN based semantic segmentation for urban traffic scenes using fisheye camera,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*. Los Angeles, CA, USA: IEEE, Jun. 2017, pp. 231–236.

[32] A. Saez, L. M. Bergasa, E. Romeral, E. Lopez, R. Barea, and R. Sanz, “CNN-based Fisheye Image Real-Time Semantic Segmentation,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*. Changshu: IEEE, Jun. 2018, pp. 1039–1044.

[33] M. Yahiaoui, H. Rashed, L. Mariotti, G. Sistu, I. Clancy, L. Yahiaoui, V. R. Kumar, and S. Yogamani, “FisheyeMODNet: Moving Object detection on Surround-view Cameras for Autonomous Driving,” *arXiv:1908.11789 [cs, eess]*, Aug. 2019, arXiv: 1908.11789. [Online]. Available: http://arxiv.org/abs/1908.11789

[34] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 6517–6525.

[35] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” *CoRR*, vol. abs/1804.02767, 2018.

[36] J. Manyika and H. Durrant-Whyte, “A tracking sonar sensor for vehicle guidance,” in *[1993] Proceedings IEEE International Conference on Robotics and Automation*. Atlanta, GA, USA: IEEE Comput. Soc. Press, 1993, pp. 424–429.

[37] S. He, H.-S. Shin, and A. Tsourdos, “Optimal Active Target Localisation Strategy with Range-only Measurements:,” in *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics*. Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 91–99.

[38] A. J. Davison and D. W. Murray, “Mobile robot localisation using active vision,” in *Computer Vision — ECCV’98*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, vol. 1407, pp. 809–825.

[39] A. Davison and N. Kita, “3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. Kauai, HI, USA: IEEE Comput. Soc, 2001, pp. I–384–I–391.

[40] M. Shibata and N. Kobayashi, “Image-based visual tracking for moving targets with active stereo vision robot,” in *2006 SICE-ICASE International Joint Conference*. Busan Exhibition & Convention Center-BEXCO, Busan, Korea: IEEE, 2006, pp. 5329–5334.

[41] T. Furukawa, C. Kang, B. Li, and G. Dissanayake, “Multi-stage Bayesian target estimation by UAV using fisheye lens camera and pan/tilt camera,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, BC: IEEE, Sep. 2017, pp. 4167–4172.

[42] M. Cognetti, P. Salaris, and P. Robuffo Giordano, “Optimal Active Sensing with Process and Measurement Noise,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. Brisbane, QLD: IEEE, May 2018, pp. 2118–2125.

[43] Ruijie He, S. Prentice, and N. Roy, “Planning in information space for a quadrotor helicopter in a GPS-denied environment,” in *2008 IEEE International Conference on Robotics and Automation*. Pasadena, CA, USA: IEEE, May 2008, pp. 1814–1820.

[44] H. Carrillo, Y. Latif, J. Neira, and J. A. Castellanos, “Fast minimum uncertainty search on a graph map representation,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vilamoura-Algarve, Portugal: IEEE, Oct. 2012, pp. 2504–2511.

[45] H. Carrillo, P. Dames, V. Kumar, and J. A. Castellanos, “Autonomous robotic exploration using occupancy grid maps and graph SLAM based on Shannon and R&#x00E9;nyi Entropy,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. Seattle, WA, USA: IEEE, May 2015, pp. 487–494.

[46] D. J. Agravante and F. Chaumette, “Active vision for pose estimation applied to singularity avoidance in visual servoing,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, BC: IEEE, Sep. 2017, pp. 2947–2952.

[47] S. Xie and Z. Tu, “Holistically-Nested Edge Detection,” in *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, Dec. 2015, pp. 1395–1403.

[48] P. Goodarzi, M. Stellmacher, M. Paetzold, A. Hussein, and E. Matthes, “Optimization of a CNN-based Object Detector for Fisheye Cameras,” in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*. Cairo, Egypt: IEEE, Sep. 2019, pp. 1–7.

[49] “Unity 3D documentation.” [Online]. Available: https://docs.unity3d.com/Manual/index.html

[50] T. Qin, J. Pan, S. Cao, and S. Shen, “A General Optimization-based Framework for Local Odometry Estimation with Multiple Sensors,” *arXiv:1901.03638 [cs]*, Jan. 2019, arXiv: 1901.03638. [Online]. Available: http://arxiv.org/abs/1901.03638