Leveraging Deep Learning Based Object Detection for Localising Autonomous Personal Mobility Devices in Sparse Maps

Maleen Jayasuriya[1], Gamini Dissanayake[1], Ravindra Ranasinghe[1], and Nathanael Gandhi[1]

Abstract— **This paper presents a low cost, resource efficient localisation approach for autonomous driving in GPS denied environments. One of the most challenging aspects of traditional landmark based localisation in the context of autonomous driving, is the necessity to accurately and frequently detect landmarks. We leverage the state of the art deep learning framework, YOLO (You Only Look Once), to carry out this important perceptual task using data obtained from monocular cameras. Extracted bearing only information from the YOLO framework, and vehicle odometry, is fused using an Extended Kalman Filter (EKF) to generate an estimate of the location of the autonomous vehicle, together with it's associated uncertainty. This approach enables us to achieve real-time sub metre localisation accuracy, using only a sparse map of an outdoor urban environment. The broader motivation of this research is to improve the safety and reliability of Personal Mobility Devices (PMDs) through autonomous technology. Thus, all the ideas presented here are demonstrated using an instrumented mobility scooter platform.**

I. INTRODUCTION

The demand and market for Personal Mobility Devices (PMDs) is predicted to sky rocket within the next decade [1], [2]. Burgeoning trends in urbanisation, improvements in battery and motor technology, demand for more environmentally friendly transport, along with a rising ageing population, have been cited as reasons for this. Although exact definitions vary according to the legislative and regulatory frameworks of different nations [3], [4] "Personal Mobility Devices" broadly refer to a class of compact electric vehicles that facilitate individual, human transportation [5]. Examples include (but not limited to) powered wheelchairs, mobility scooters, segways, and hover boards.

These devices generally travel at very low speeds (under 15 Kmph), come in small form factors, and travel in spaces typically reserved for pedestrians, such as pavements and footpaths. This encroachment into pedestrian spaces has raised legitimate safety concerns and calls for stricter regulation in many nations [6], [4], [7]. However it must also be noted that a majority of PMDs help raise the living standards of many individuals with mobility restrictions. Furthermore, PMDs cannot be ignored as a potentially integral part of a more environmentally friendly, intelligent, future urban transportation system, specially in the context of first/last mile transportation [8], [9]. Considering these facts, the recent boom in self-driving vehicle technology has immense promise to balance the potential benefits of PMD usage, with their safety concerns.

[1]Maleen Jayasuriya, Gamini Dissanayake, Ravindra Ranasinghe and Nathanael Gandhi are with the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

An autonomous vehicle in general is a collection of many complex overlapping subsystems, of which the localisation module is fundamental. The localisation sub-system is responsible for locating a vehicle within a global coordinate frame. This is an elemental task in autonomous driving, since an accurate and reliable location estimate informs many other subsystems ranging from global route planning, to obstacle avoidance.

Over the last two decades, Global Positioning Systems (GPS) have traditionally aided human drivers in basic navigation tasks in urban driving scenarios. However, they do not currently posses the accuracy ($\sim$10m), precision, or reliability required for autonomous driving [10]. Thus, the current generation of autonomous vehicles focus on the use of additional sensors to aid in the localisation task. A collection of 2D/3D LIDARs and vision sensors, along with detailed high definition 3D maps [11] are commonly used to achieve the reliability and robustness required for a high speed vehicle operating in cities [10], [12]. Building and maintaining such high definition maps places high demands on resources such as data collection, storage, computational power and data transmission [13]. Thus, approaches that are capable of localising in sparse or low resolution maps are gaining popularity.

In [14], authors propose using the low resolution crowd sourced, Open Street Map (OSM), together with GPS and wheel odometry, to determine way points for global navigation. Local navigation in-between way points is achieved using a LIDAR based local perception system. However, this approach dubbed the MapLite system [15] is heavily reliant on GPS and thus primarily targeted to operate in rural environments. The synthetic LIDAR approach outlined in [16] and [17] uses a tilted 2D LIDAR input, with a 3D rolling window to form a 2D local map of the environment. Localisation of an autonomous golf-cart [18] and a mobility scooter [19] using this technique has been demonstrated. However, the approaches outlined above rely on the use of multiple expensive LIDARs, which is difficult to justify in the context of low-cost devices such as PMDs.

In general, cameras and vision based systems can offer more compact and low cost localisation solutions [10]. Popular Visual Odometry (VO) methods range from appearance based techniques such as optical flow to feature based methods [20]. Feature based systems employ feature detectors and descriptors such as FAST, SIFT, SURF, ORB and BRIEF [20], [21]. However most of these approaches tend to fail under extreme appearance, illumination, occlusion and weather changes, common in autonomous driving scenarios

[22], [23], [24], [25]. Thus, vision based approaches are currently dominated by techniques based in Deep Convolutional Neural Networks (CNNs), due to their ability to learn generic feature extractors that are robust to appearance and viewpoint changes [21].

One approach to CNN based localisation involves training neural networks to carry out complete end-to-end pose regression of a camera, based on an image [END TO END REFERENCE ICRA] [26], [27], [28]. Another approach involves semantic segmentation where a CNN is used to label each pixel with semantic information. For instance [29] proposes using semantic information to reject dynamic objects such as pedestrians, cars and bikes while only using static objects such as trees and posts, to aid popular vision based navigation frameworks. Alternatively, [30] uses a semantic segmentation process to extract geometric information from landmarks such as poles, street signs and traffic lights to match them against a 3D map of such features, using an optimisation process to obtain the corresponding camera pose. However, the CNN based techniques of end to end pose regression and semantic segmentation discussed above, have very high computational and training requirements that are difficult to be met in real time, specially in the context of an autonomous PMD.

Unlike pose regression or complete semantic labelling/segmentation of each pixel in an image, the general task of CNN based object detection is demonstrably less computationally demanding, and also easier to train. This involves detecting and placing bounding boxes over objects of interest. State of the art object detection methods include R-CNN [31] and YOLO [32]. Recent improvements to YOLO, specifically YOLOv2 [33] and YOLOv3 [34] have consistently outperformed the competition both in terms of accuracy and speed.

In this paper, we propose a resource efficient, real time, vision based localisation system that operates on a given sparse map. The sparse map only consists of common, persistent and easily discernible landmarks such as streetlamps, trees, parking meters, traffic lights and road signs that are typically found in the operating environments of PMDs. The overall system relies on low cost vision sensors and recent developments in deep learning based object detection, to form a robust perceptual front end. Information from this perception system is then used to carry out landmark based bearing only localisation using an Extended Kalman Filter (EKF). To validate the merits of the proposed localisation approach in unstructured real world environments, an off-the-shelf mobility scooter was retrofitted with a low cost computation and sensor package.

The remainder of this paper is structured as follows. Section II outlines the core framework and methodology behind our localisation system. Section III provides a brief overview of the hardware platform used to validate the proposed concepts. Section IV presents details and results of the conducted experiments. Finally, section V concludes the paper with with a brief discussion on the experimental results and some thoughts on future work.

II. LOCALISATION FRAMEWORK

The proposed localisation framework consists of a deep learning based perceptual front end, and an Extended Kalman Filter (EKF) based back-end for 2D pose estimation in a given map (See figure 1). The map consists only of the 2D locations of the landmarks, relative to a global coordinate frame.

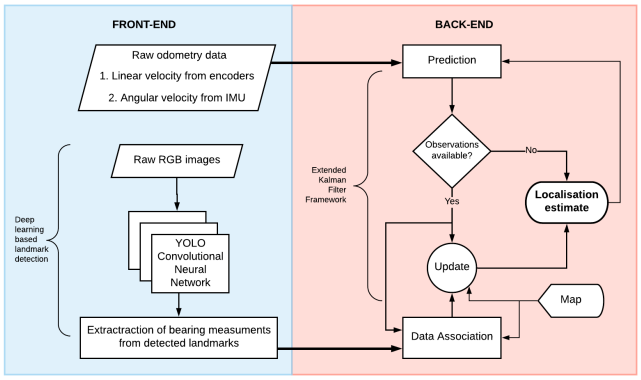

Fig. 1: Localisation framework

A. *Perceptual front-end*

As reviewed in section II, the YOLO object detection framework provides state of the art real time object detection with realtively high accuracy. Thus, YOLO version 2, which operates at a frame rate of 15-17 FPS on our hardware platform (described in section III), forms the basis of the perceptual front-end of the proposed system. Although YOLO v2 is less accurate than YOLO v3, the level of trade off between speed and accuracy suited our application and operating parameters better.

The underlying neural network architecture of YOLO v2 is known as Darknet-19, and consists of 19 convolutional layers and 5 maxpooling layers. This network is pre-trained on the standard ImageNet 1000 class classification dataset, and acts as the base feature extractor of YOLO v2. This base network can be further trained to detect any custom object class, through the process of transfer learning. This is achieved by removing the last convolutional layer and adding 3 more convolutional layers with 1024 filers each, followed by a final convolutional layer with the number of outputs matching the number of custom object detection classes required.

For the purpose of training YOLO v2 to detect landmarks required for localisation, we curated our own image dataset, collected by driving the hardware platform in the streets of Sydney. Commonly found features seen in pedestrian environments, such as street lamps, road signs, traffic lights, parking meters and trees, were identified as potential candidates for localisation landmarks. YOLO v2 was then trained using this dataset along with additional images obtained through the Imagenet database and web scraping. Once trained, and an RGB image is provided to YOLO v2; it detects the required landmarks in real time. Detections are visualised as bounding boxes around the landmarks of interest.

(a) Wentworth Park (b) Mary-Ann St

Fig. 2: Landmark detections from experiments

Figure 2 depicts examples of detected landmarks, observed during the experiments outlined in section IV. To obtain bearing information to these detected landmarks, the coordinates of the centroid of the bounding boxes; (u_c, v_c) are used to calculate the horizontal bearing θ_t^i of a detected landmark i at time t. Thus, a set of b bearing observations $\Theta_t = [\theta_t^1, \theta_t^2 .... \theta_t^i .... \theta_t^b]$ is obtained at time t.

The perceptual front end also includes odometry information, $U_t = [v_t, \omega_t]$ consisting of linear velocity v_t and Z-Axis angular velocity ω_t.

The hardware platform outlined in section III provides the RGB images required for YOLO via two Realsense D435 cameras. Odometry information is obtained from two rotary wheel encoders and the yaw gyro of an IMU unit.

B. EKF back-end

The back-end of the localisation system consists of an Extended Kalman Filter. Here, the information obtained from the perceptual front end, consisting of odometry data; $U_t = [v_t, \omega_t]$, is fused with the landmark bearing observations $\Theta_t = [\theta_t^1, \theta_t^2 .... \theta_t^i .... \theta_t^b]$.

The prediction step (Lines 2 to 6 of Algorithm 1) of the EKF uses the Odometry Motion Model g described by equation (1):

$$\hat{X}_t = g(X_{t-1}, U_t)$$

$$\begin{bmatrix} \hat{x_t} \\ \hat{y_t} \\ \hat{\phi_t} \end{bmatrix} = \begin{bmatrix} x_{t-1} + \Delta T * v_t * \cos\phi_{t-1} \\ y_{t-1} + \Delta T * v_t * \sin\phi_{t-1} \\ \phi_{t-1} + \Delta T * \omega_t \end{bmatrix} \quad (1)$$

Where, $X = [x, y, \phi]$ describes the 2D pose of the robot, $\hat{X}_t$ the predicted pose at time t, and ΔT the time between t and $t-1$.

The predicted pose covariance $\hat{P}_t$ is calculated based on odometry noise Q, and the relevant Jacobians (See line 6 of Algorithm 1).

When a landmark is detected by the perceptual front-end, the predicted pose is corrected by the update step of the EKF (Lines 7 to 16 of Algorithm 1). First, each observation is associated with the relevant landmark (see section II-C for more details). Let the map of l landmarks be $M = [m_1, m_2 ... m_j .. m_l]$, where $m_{j,x}$ and $m_{j,y}$ represents the x and y coordinate of the j^{th} landmark respectively. Then, the observation model h is given by equation (2):

Algorithm 1 EKF

1: Inputs: $X_{t-1}, P_{t-1}, \Theta_t, U_t, M$
2: Calculate predicted pose using motion model

$$\hat{X}_t = g(X_{t-1}, U_t)$$

3: Set control noise covariance Q
4: Calculate Jacobian ∇G, of $g(X_{t-1}, U_t)$ w.r.t to X_{t-1}
5: Calculate Jacobian ∇U, of $g(X_{t-1}, U_t)$ w.r.t to U_{t-1}
6: Calculate predicted pose covariance:

$$\hat{P}_t = \nabla G * P_{t-1} * \nabla G^T + \nabla U * Q * \nabla U^T$$

7: **if** Bearing measurements Θ_t are available **then**
8: Carry out data association (see section II-C)
9: Calculate corresponding predicted observations

$$\hat{\Theta}_t = h(\hat{X}_t, M)$$

10: Calculate Innovation

$$v = \hat{\Theta}_t - \Theta_t$$

11: Calculate Jacobian ∇H, of $\hat{\Theta}_t = h(\hat{X}_t, M)$ w.r.t $\hat{X}_t$
12: Set bearing measurement noise R
13: Calculate innovation co-variance:

$$S = R + \nabla H * \hat{P}_t * \nabla H^T$$

14: Calcuate Kalman gain K:

$$K = \hat{P}_t * \nabla H^T * S^{-1}$$

15: Correct pose estimate:

$$X_t = \hat{X}_t + K * v$$

16: Correct pose covariance:

$$P_t = \hat{P}_t - K * S * K^T$$

17: **else**
18: $X_t = \hat{X}_t$
19: $P_t = \hat{P}_t$
20: **end if**
21: **Return** X_t, P_t

$$\hat{\Theta}_t = h(\hat{X}_t, M) \quad (2)$$

where the i^{th} component of $\hat{\Theta}_t$ is:

$$\hat{\theta}_t^i = \text{atan2}(m_{j,y} - y_r, m_{j,x} - x_r) - \hat{\phi}_t$$

(x_r, y_r) is the location of the camera with respect to the global frame. This is calculated by equation (3), where a and b are the x and y offset of the camera relative to the platform's local frame.

$$\begin{aligned} x_r &= \hat{x_t} + a * \cos(\hat{\phi_t}) - b * \sin(\hat{\phi_t}) \\ y_r &= \hat{y_t} + a * \sin(\hat{\phi_t}) + b * \cos(\hat{\phi_t}) \end{aligned} \quad (3)$$

Finally the predicted pose estimate is corrected based on the computed Kalman gain K and innovation v (See lines 10 to 16 of Algorithm 1). Thus, the EKF returns the final pose estimate X_t and associated covariance P_t, successfully localising the system relative to the given map.

C. Data association

Once a landmark is detected, the observed bearing is associated with the correct landmark using an innovation gate,

based on the Mahalanobis distance d of each observation, calculated by equation (4):

$$d^2 = v^T * S^{-1} * v \tag{4}$$

d^2 is calculated using the innovation v between each observation θ_t^i and the predicted observations to each landmark. S is the corresponding innovation co-variance. d^2 is distributed as a chi-squared random variable with 1 degree of freedom. Observations can be associated with landmarks when d^2 is below a bound that is defined using a desired level of confidence. If multiple associations are made to one observation, the association with the lowest d^2 value is considered as the final association. All observations that do not pass the innovation gate and do not meet the above criteria, are ignored.

III. HARDWARE SYSTEM OVERVIEW

The Pride Pathrider 10, one of the most popular and reliable mobility scooters on the market, was selected as the base vehicle for our experimental hardware platform. As depicted in figure 3, the mobility scooter was retrofitted with a low cost computation and sensor package. Final physical specifications of the scooter post modification is outlined in Table I.

Fig. 3: Hardware overview of retrofitted mobility scooter

TABLE I: Pathrider 10 Specifications

Dimensions (L x W x H)	1.9 x 0.56 x 1.65 m
Weight	105 Kg
Maximum speed	8.85 km/h
Turning Clearance Circle	1.575 m (Turning radius)

A. Vision sensors

Two Intel®D435 cameras mounted at 45°angles to the heading of the mobility scooter are used as the primary vision sensors for localisation. A single D435 camera posses a horizontal field of view of 69.4°. This mounting configuration allows for a larger field of view to be dedicated towards the left and right sides of the scooter, which are generally richer in landmark features as opposed to facing forward. This also ensures that bearing measurements to detected landmarks are larger than when facing forward, reducing the measurement's percentage error. Furthermore, this configuration eases the landmark data association problem, as it reduces the chance of similar landmarks (Eg: streetlamps) overlapping due to parallax.

Depth measurements from the Realsense camera were found to be somewhat unreliable and requires further processing and calibration. Hence, as discussed in section II, localisation is carried out based only on bearing information obtained by the RGB sensor of the Realsense. Thus the algorithm presented here could be implemented using any monocular RGB camera. The Realsense camera was selected with the view of using it for other navigation tasks such as obstacle avoidance further down the line.

B. Computing

The primary computational unit of the scooter is an NVIDIA®Jetson AGX Xavier embedded system, with a 512-core Volta GPU and 16GB of RAM. The scooter is also equipped with an Intel®UP2 board that interfaces with the on board sensors. All systems interface, operate and communicate using ROS (Robot Operating System) Melodic.

C. Odometry

The scooter is equipped with two rotary encoders to measure wheel rotation. An MPU-9250 IMU is also attached to the camera mounting plate to provide heading information.

D. RTK-GPS

The Piksi Multi Real Time Kinematic (RTK) GPS unit is a multi-band, multi-constellation RTK GNSS receiver that provides centimetre-level accuracy. However this level of accuracy is only available when a large portion of the sky is visible. It is therefore unsuitable as a sensor for localisation. The scooter is currently equipped with this module to evaluate the performance of our localisation algorithm, when the vehicle is travelling through regions where RTK information is available. The RTK GPS was also used to physically survey landmark locations for map building.

IV. EXPERIMENTAL RESULTS

A. Algorithm validation

Initial experiments were carried out along pedestrian footpaths in an approximately ∼2000 square metre area of Wentworth Park, Sydney, Australia (Figure 4). This location was chosen due to the lack of obstructions in the skyline, enabling us to obtain a high quality, centimetre accurate RTK GPS fix. These continuous fixed RTK GPS readings were used as ground truth to evaluate our algorithms, and also obtain the locations of important visual landmarks such as lamp posts, trees and street signs, to generate a sparse geometric map of the park environment.

Figure 5 shows the error of the reported location estimates, together with their corresponding 2σ covariance bounds. It is

Fig. 4: Localisation result at Wentworth Park

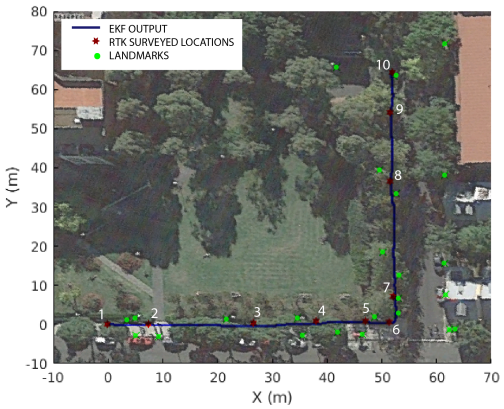

Fig. 6: Localisation result at Mary-Ann St tIno Bulwara Rd

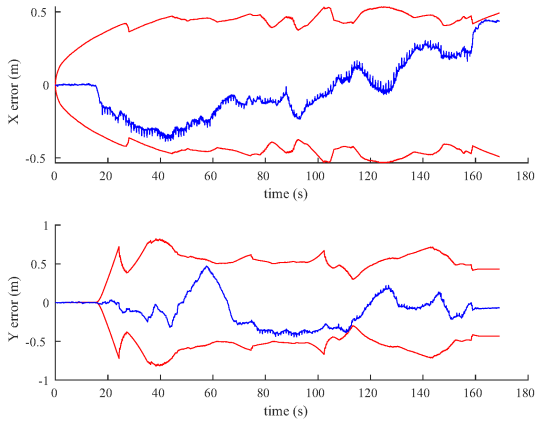

Fig. 5: Estimation error at Wentworth Park (blue), with 2σ covariance bounds (red)

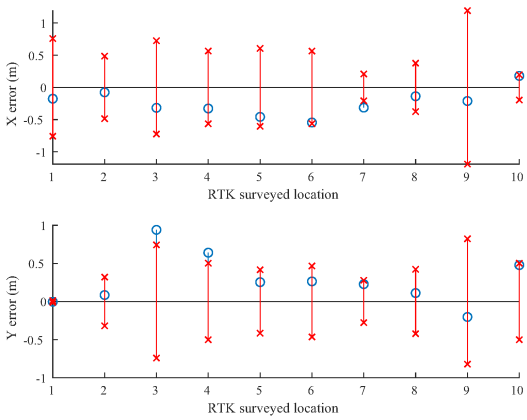

Fig. 7: Estimation error (blue) at Mary-Ann St to Bulwara Rd, with 2σ covariance bounds (red)

clear that the errors appear to be within the 95% confidence bounds defined by the 2σ gate, indicating that the EKF is well tuned. The maximum error values reported in either X or Y location estimates, during multiple runs ranged from $0.40 - 0.60m$.

B. Demonstration

To demonstrate the capabilities of the proposed framework in a a more real world urban environment, experiments were subsequently carried out along a typical suburban environment near the university campus, along Mary Ann St to Bulwara Rd, Sydney (Figure 6). The platform was driven along the pavement amidst pedestrians and uneven terrain, covering a distance of roughly 130m. Continuous fixed RTK ground truth information while in motion however, was not available due to surrounding trees and buildings occluding the skyline. Thus to provide an evaluation of the localisation error involved, ten separate locations were surveyed beforehand using the RTK receiver. These locations were selected based on a combination of available fixed RTK readings and low variance float RTK readings measured over a period of time. The error was then calculated by comparing the localisation output of the EKF against these surveyed locations when the platform was driven over them.

Figure 7 shows the localisation error at the ten surveyed locations along with their corresponding 2σ covariance bounds. The maximum error values reported in either X or Y location estimates, during multiple runs ranged from $0.50 - 0.94m$.

V. DISCUSSION AND CONCLUSION

It was observed during the experiments that the quality of the bearing information, is highly susceptible to the size and shape of the landmark being detected. This is the reasoning behind selecting vertical shaped objects as candidate landmarks. However, bearing noise from landmarks such as trees are still highly variable due to their shape and size. This poses challenges in terms of data association and overall functioning of the EKF.

In order to guarantee that only high quality bearing information and data associations feed into the EKF algorithm, a strict innovation gate of 0.46 (corresponding to 50% confidence in the 1-DOF Chi-Squared distribution) was set heuristically based on practical results. Although this was found to discard approximately 60-70% of observations, the information gathered from accepted observations was clearly adequate to produce a good quality location estimate.

Thus one key area of future work will focus on creating a more holistic and unique noise profile for each landmark

based on their size and shape, during the mapping process. One mapping avenue currently being explored is to remove the reliance on RTK-GPS and consider SLAM based frameworks. In terms of perception, we plan on also exploiting information from features such as pathways, pavement edges and curbs. Exploiting the depth information available from the Realsense cameras as well as the semantic labels provided by Yolo, are also avenues for further investigation.

The current iteration of our localisation framework has delivered promising results during initial experiments. Sub metre accuracies were reported using only a minimal map representation and a low cost computation and sensor package. We plan to build on these results with the goal of providing a low resource intensive localisation framework, suitable for the safe and efficient operation of autonomous PMDs.

REFERENCES

[1] PR-Newswire, "Personal mobility devices market worldwide 2014-2023," 2015. [Online]. Available: https://www.statista.com/statistics/485524/global-personal-mobility-devices-market-size/

[2] PR-Newswire, "Global Medical Mobility Scooters Market 2018-2022," 2018. [Online]. Available: https://www.prnewswire.com/news-releases/global-medical-mobility-scooters-market-2018-2022-analysis–forecasts-by-4-5-and-3-wheeler-medical-mobility-scooters-300727772.html

[3] Government-Australia, "Non-Road Vehicle Option," 2019. [Online]. Available: https://infrastructure.gov.au/vehicles/

[4] S. Chin, "Singapore reins in personal mobility devices," 2018. [Online]. Available: https://theaseanpost.com/article/singapore-reins-personal-mobility-devices

[5] "personal mobility device," 2019. [Online]. Available: https://medical-dictionary.thefreedictionary.com/personal+mobility+device

[6] R. Mealey, "Brisbane proposes 6kph speed limit for mobility scooters to protect pedestrians," 2018. [Online]. Available: http://www.abc.net.au/news/2018-06-12/brisbane-city-council-proposes-mobility-scooter-speed-limit/9860676

[7] S. O'Kane, "2017 will be an important year for personal electric vehicles of all sizes," 2017. [Online]. Available: https://www.theverge.com/2016/12/31/14134924/electric-skateboards-boosted-bikes-vehicles-hoverboards

[8] T. Birtchnell, G. Waitt, and T. Harada, "Dont ignore the mobility scooter. It may just be the future of transport," *The Conversation*, 2017. [Online]. Available: https://theconversation.com/dont-ignore-the-mobility-scooter-it-may-just-be-the-future-of-transport-85170

[9] R. Dowling, J. D. Irwin, I. J. Faulks, and R. Howitt, "Use of personal mobility devices for first-and-last mile travel: The Macquarie- Ryde trial," in *2015 Australasian Road Safety Conference*, 2015, p. 13.

[10] S. Kuutti, Fallah, S., Katsaros, K., Dianati, M., Mccullough, F., and Mouzakitis, A., "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.

[11] B. Templeton, "Many different approaches to Robocar Mapping | Robohub," 2017. [Online]. Available: http://robohub.org/many-different-approaches-to-robocar-mapping/

[12] F. Poggenhans, J.-H. Pauls, J. Janosovits, S. Orf, M. Naumann, F. Kuhnt, and M. Mayr, *Lanelet2: A high-definition map framework for the future of automated driving*, Nov. 2018.

[13] 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

[14] T. Ort, L. Paull, and D. Rus, "Autonomous vehicle navigation in rural environments without detailed prior maps," in *2018 IEEE International Conference on Robotics and Automation*, 2018, p. 8.

[15] A. Simons and R. Gordon, "Self-driving cars for country roads," 2018. [Online]. Available: http://news.mit.edu/2018/self-driving-cars-for-country-roads-mit-csail-0507

[16] Z. J. Chong, B. Qin, T. Bandyopadhyay, M. H. Ang, E. Frazzoli, and D. Rus, "Mapping with synthetic 2d LIDAR in 3d urban environment." IEEE, Nov. 2013, pp. 4715–4720. [Online]. Available: http://ieeexplore.ieee.org/document/6697035/

[17] Z. J. Chong, B. Qin, T. Bandyopadhyay, M. H. Ang, E. Frazzoli, and D. Rus, "Synthetic 2d LIDAR for precise vehicle localization in 3d urban environment," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 1554–1559.

[18] S. Pendleton, T. Uthaicharoenpong, Z. J. Chong, G. M. J. Fu, B. Qin, W. Liu, X. Shen, Z. Weng, C. Kamin, M. A. Ang, L. T. Kuwae, K. A. Marczuk, H. Andersen, M. Feng, G. Butron, Z. Z. Chong, J. Ang, E. Frazzoli, and D. Rus, "Autonomous golf cars for public trial of Mobility on Demand Service," *Rus*, Sep. 2015.

[19] H. Andersen, Eng, Y. H., Leong, W. K., Zhang, C., Kong, H. X., Pendleton, S., Ang, M. H., and Rus, D., "Autonomous personal mobility scooter for multi-class mobility-on-demand service," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, Nov. 2016, pp. 1753–1760.

[20] Reza Zekavat and R. Michael Buehrer, "Localization for Autonomous Driving," in *Handbook of Position Location: Theory, Practice, and Advances*. IEEE, 2019, pp. 1051–1087. [Online]. Available: http://ieeexplore.ieee.org/document/8633804

[21] S. Lowry, N. Snderhauf, 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.

[22] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, Jun. 2008. [Online]. Available: https://doi.org/10.1177/0278364908090961

[23] H. Chiu, M. Sizintsev, X. S. Zhou, P. Miller, S. Samarasekera, and R. Kumar, "Sub-meter vehicle navigation using efficient pre-mapped visual landmarks," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Nov. 2016, pp. 505–512.

[24] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 901–906.

[25] A. Ranganathan, S. Matsumoto, and D. Ilstrup, "Towards illumination invariance for visual localization," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 3791–3798.

[26] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2938–2946.

[27] A. Kendall and R. Cipolla, "Geometric Loss Functions for Camera Pose Regression with Deep Learning," in *2017 IEEE Conference on Computer Vision and Pattern*, Jul. 2017.

[28] A. Valada, N. Radwan, and W. Burgard, "Deep Auxiliary Learning for Visual Localization and Odometry," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 6939–6946.

[29] 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.

[30] 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.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2014, pp. 580–587. [Online]. Available: http://ieeexplore.ieee.org/document/6909475/

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv:1506.02640 [cs]*, Jun. 2015, arXiv: 1506.02640. [Online]. Available: http://arxiv.org/abs/1506.02640

[33] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *CoRR*, vol. abs/1612.08242, 2016.

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