Fit for Purpose?
Predicting Perception Performance based on Past Experience

Corina Gurău, Chi Hay Tong and Ingmar Posner

Mobile Robotics Group, University of Oxford, United Kingdom
{corina, chi, ingmar}@robots.ox.ac.uk

Abstract. This paper explores the idea of predicting the likely performance of a robot's perception system based on past experience in the same workspace. In particular, we propose to build a place-specific model of perception performance from observations gathered over time. We evaluate our method in a classical decision making scenario in which the robot must choose *when* and *where* to drive autonomously in 60km of driving data from an urban environment. We demonstrate that leveraging visual appearance within a state-of-the-art navigation framework increases the accuracy of our performance predictions.

Keywords: robot perception, object detection, performance prediction

1 Introduction

Reliable robot perception is a difficult yet fundamental problem, as robots interact directly with the world and any misconduct can have adverse consequences. Our goal is to equip a robot with the introspective capability of predicting when the operational environment is challenging and its perception system is underperforming. Such a high level understanding of the operational environment constitutes a useful diagnostic tool for any decision making agent. Just as humans have the ability to anticipate a difficult road situation, such as an approaching busy intersection or a narrow and crowded street, the robot should be equipped with the ability to forsee its perceptual shortcomings. While significant effort is being devoted to building highly performant perception systems ([1], [2], [3]), the problem of predicting their failure in action is, to the best of our knowledge, overlooked. As robots operate in complex, continuously-evolving, dynamic workspaces, it is critical to analyse and predict how robustly their perception systems operate at any given moment in time.

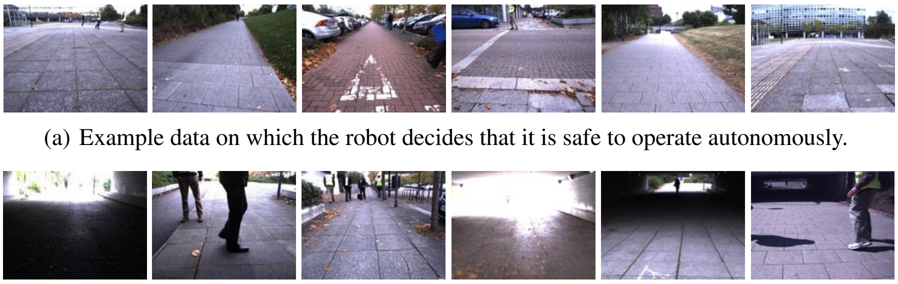
Our work is additionally motivated by our previous observations that perception performance for mobile robots is environment-dependent. In some places of operation performance is excellent, while in others failure occurs more often [4]. We attribute this to the vicissitudes of the environment – changes in appearance due to external factors such as weather or illumination conditions.

In this work we propose to model the robot's perception capabilities using a probabilistic framework. Our goal is to allow the robot to drive autonomously *only* when it is confident of its performance and require human assistance otherwise. Some examples of this scenario can be seen in Figure 1. Requiring a human to intervene in an autonomous operation falls under the *autonomy on offer* paradigm, in which the robot offers autonomy only when it is extremely confident in its capabilities and hands over control to a human otherwise. More specifically, the contributions of this work are:

- Introducing **performance records**: a probabilistic framework used to incorporate place-specific performance estimates gathered over time, which allow the robot at test time to estimate the likelihood of the perception system making a mistake.
- The description of two modalities for using performance records, one of which makes use of the visual appearance of a place.
- A classical decision making scenario which allows the robot to take an optimal action regarding offering autonomy.

2 Related Work

There are several works that touch upon the fluctuating performance levels of a robot during operation. However, we believe to be the first to estimate the likelihood of success of a vision system by modelling its outcome as a function of space and time. The system we propose is deeply relevant to the work of [5] who describe the sensitivity of object detectors to factors such as weather and location and train local experts by incorporating place specific hard negative examples in the training procedure. When data the robot is unlikely to encounter during operation is replaced with mistakes, they are able to significantly improve the detection results. Unreliable perception performance has also been observed by [6] and [7] who attribute it to sensor data integrity and analyse the effects of challenging operational conditions on the perceptual integrity of the robot. The works of [8] and [9] identify the use of biased training datasets as a cause for poor generalisation performance to new tesing conditions. Similar problems are reported for localisation performance. [10] and [11] propose embedding spatial models of expected localiser performance in localisation maps in order to aid trajectory planners.



(a) Example data on which the robot decides that it is safe to operate autonomously.

(b) Example data on which the robot can ask to switch the control back to a human.

Fig. 1: Example data encountered by a robot as it traverses an urban environment in the proximity of pedestrians, cyclists and other road users. On some sections of the road on which it belives its perception system is underperforming the robot can ask to switch control back to a human operator.

This higher-level characterisation of when and where an algorithm fails is similar in spirit to the concept of *introspection* introduced in [12]. In that work, the authors looked at the introspective capacity of different classification frameworks, which refers to a classifier's ability to assign an appropriate measure of confidence to any test data. Mistakes are not considered catastrophic when they are made with high uncertainty as this gives the system the ability to ask for help and correct itself. Our framework is independent of the classification algorithm. It bears some similarity with [13], which introduces ALERT, a system used to predict the accuracy of a computer vision system on various tasks. We share with ALERT an aspiration to prevent failure by flagging a warning when predicting that performance will be low. However, our work stands apart from that of [13] as our approach is tailored specifically to robot perception by exploiting location and past experiences of a place. These provide useful contextual information, which can improve the robot's decision making capabilities.

3 Approach

We rarely allow robots to drive autonomously somewhere totally new. In fact, most successful autonomous operation techniques exploit the fact that the robot often traverses the same workspace over and over again ([14]). If a robot has traversed a route in the past, then we would like to leverage its past experience in order to predict the robot's performance in subsequent visits of the same place. Based on these predictions, we would like the robot to offer autonomy only if its estimates of performance are high, and deny it otherwise. Figure 2 shows how we leverage past information: we drive the same route multiple times and gather performance estimates along it. Specifically, what we estimate in this paper is the image-based pedestrian detection outcome. In order to achieve this we need to address the following:

- estimating detection performance at a particular location
- formulating offering/denying autonomy as a decision making problem

3.1 Building Performance Records

We consider the environment (the place of operation) to be an underlying hidden influence on the detection outcome. For a traversal T of a route, we denote as T_i the i^{th} location along it. We define θ_i as the probability that at T_i the detection system will be successful, and we model it as a random variable with the probability $p(\theta_i)$ assumed to be a beta density of the form

$$p(\theta_i;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}, \qquad 0 \leq \theta_i \leq 1, \tag{1}$$

where $\alpha > 0$, $\beta > 0$ and $B(\alpha,\beta)$ is the Beta function. Our canonical prior at a new location that we see for the first time and where we have no knowledge of the success of the detector is given by $\alpha = 1, \beta = 1$. As the robot traverses the route, at each location T_i it observes a set of detections: true positive, false positive and false negative respectively.

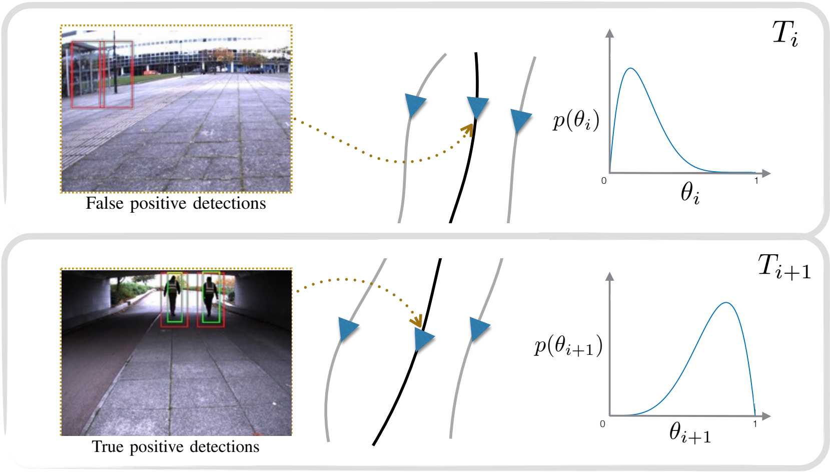


Fig. 2: A new traversal (black line) of a route which has been travelled previously (grey lines) can make use of past estimates of detection performance. For instance, at Location A where we have repeatedly observed false positive detections, the performance record yields a low probability of success for the detector, while at Location B, where the detector has only produced true positive detections, the probability of success is very high.

They represent the success or failure of the detection system and we record them as binary observations $x_i^j \in \{0,1\}$ such that:

$$x_i^j = \begin{cases} 1, & \text{if the } j^{th} \text{ observation at } T_i \text{ is a true positive} \\ 0, & \text{if the } j^{th} \text{ observation at } T_i \text{ is a false positive or a false negative} \end{cases} \tag{2}$$

We let the observations x be modelled by a Bernoulli random variable: $x \sim Ber(\theta)$ with probability mass function:

$$p(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad x \in \{0,1\}. \tag{3}$$

We additionally make the assumption that the set of obsevations $X_i = \{x_i^1, x_i^2, \ldots, x_i^{n_i}\}$ are conditionally independent given the probability of success θ_i, and express the likelihood of successful performance for a particular location T_i as:

$$p(X_i|\theta_i) \propto \prod_{j=1}^{n_i} p(x_i^j|\theta_i) \propto \theta_i^{k_i}(1-\theta_i)^{n_i-k_i}, \tag{4}$$

where k_i represents the number of observations indicating good performance ($x_i = 1$) out of a total of n_i observations at location T_i along the route. Using Bayes Theorem, we calculate the probability of the detector being successful at location T_i as:

$$p(\theta_i|X_i) = \frac{p(X_i|\theta_i)p(\theta_i)}{\int_{\theta_i} p(X_i|\theta_i)p(\theta_i)} \tag{5}$$

Since the Beta distribution is a conjugate prior to the Bernoulli distribution, the posterior $p(\theta_i|X_i)$ is also a Beta distribution. The hyperparameters of the posterior are updated as:

$$\widehat{\alpha}_i = \alpha + k_i, \quad \widehat{\beta}_i = \beta + n_i - k_i \tag{6}$$

This gives us a simple procedure for incorporating observations over time. We refer to all $p(\theta_i; \widehat{\alpha}, \widehat{\beta})$ at locations T_i as the performance record of the detection system on a chosen route after traversal T and use it to estimate the likely performance of the robot at test time.

3.2 Decision Making using a Performance Record

Using Bayesian decision theory we can translate the posterior probability of performance into optimal actions. In this paper we focus on the case in which the robot can take either of the following two actions: a^0, *denying autonomy* or a^1, *offering autonomy* at every location along a test route. The robot should choose action a^0 when it believes that its perception system is failing and a human operator should take over control and it should choose action a^1 when it believes that its perception system is functioning well and it can reliably operate autonomously.

We make the simplifying assumption that there are only two states that the perception system can be in: failing (and producing false detections), or performing well (and the robot presents no risk when operating autonomously). In order to discriminate between the two states, we introduce hyperparameter τ and denote by s^0 the event that the perception system is failing at location L_i. We compute its probability as

$$p(s^0|\theta,\tau) = p(\theta \leq \tau) = \int_0^\tau p(\theta; \widehat{\alpha}, \widehat{\beta})\,\mathrm{d}\theta, \tag{7}$$

where $p(\theta; \widehat{\alpha}, \widehat{\beta})$ has been estimated using the performance records proposed. We denote by s^1 the event that the perception system is performing well and compute the probability of it happening as $p(s^1|\tau) = 1 - p(s^0|\tau)$. In order to select an optimal action, we associate a loss to each of the event-action pairings, which reflects how serious it is to take action a^i when the actual state is s^j, for $i, j \in \{0, 1\}$:

$$L(a,s) = \begin{pmatrix} 0 & \mathrm{L}_{\text{offer}} \\ \mathrm{L}_{\text{deny}} & 0 \end{pmatrix}$$

We choose the action which minimises the expected loss computed as

$$\overline{L}_\tau(a) = \sum_i p(s^i) L(a, s^i). \tag{8}$$

In our scenario, denying autonomy and asking for help (even if un-neccessary) is more desirable than driving autonomously while the perception system is performing poorly, as the latter can have catastrophic consequences. In Figure 3 we show the effect of adjusting the losses associated with each type of error on the actions selected. Type I, or false positive errors, correspond to the situations in which the robot denies autonomy (a^0) but its perception system is in reality performing well (s^1) and incur a loss of L_{deny}. Type II, or false negative errors, occur when the robot fails to recognise that it is underperforming (s^0) and continues to operate autonomously (a^1). Figure 3 shows that by making type I errors more expensive (increasing $\mathrm{L}_{\mathrm{offer}}$), the robot employs the safer action of denying autonomy more often.

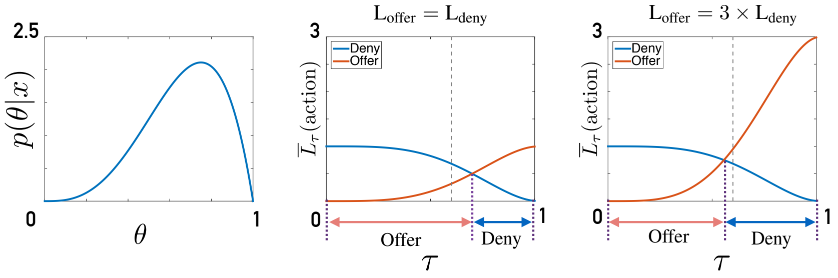


Fig. 3: Figure shows the expected loss of choosing an action for a posterior distribution $p(\theta|x)$ and two different loss matrices used. When $\mathrm{L}_{\mathrm{offer}} = \mathrm{L}_{\mathrm{deny}}$, for $\tau = 0.6$ (grey line), the action chosen by the robot is to offer autonomy because it has a lower expected loss $\overline{L}_{\tau=0.6}$. However, by setting $\mathrm{L}_{\mathrm{offer}} = 3 \times \mathrm{L}_{\mathrm{deny}}$, the optimal action becomes to deny autonomy. Increasing $\mathrm{L}_{\mathrm{offer}}$ creates a more cautious system that will offer autonomy less often.

3.3 Performance Records and the Experience Paradigm

In order to assign different observations to the same location we use geographical proximity given by GPS measurements. While this distance metric is useful for gathering all the observations close to a desired location, it does not take into account which of them are most relevant. Imagine the following test case: while driving at night, past observations gathered during night time should be more relevant than observations gathered during day time. Similarly, detection in bright sunny conditions might have a different outcome than detection during rain. In these situations having a distance metric that also incorporates visual similarity is crucial. Here is where Experience-Based Navigation (EBN) comes in. EBN ([15], [16]) is an ideal framework for our problem as it selects, through a camera-based localisation system, which of the past appearances of a location most resemble what the robot is experiencing at test time. In order to do this, EBN distinguishes between different visual appearances of a place and, as any vision based feature matching system, it works better at matching images when visual features are common. We hypothesise that visual features similar enough for localisation will produce a similar detection outcome.

We denote the method of estimating performance using all past observations, regardless of the visual appearance of the environment by **LOC**, since it only incorporates observations that are close in location. We denote a second method, which leverages EBN to select observations from locations that are close both in physical distance and visual appearance by **APP**. We expect the second method to give better estimates of performance as it accounts for more than structural changes of environment (different locations) but also for appearance changes caused by lighting, weather, or even time of the day, that can significantly influence a detection system.

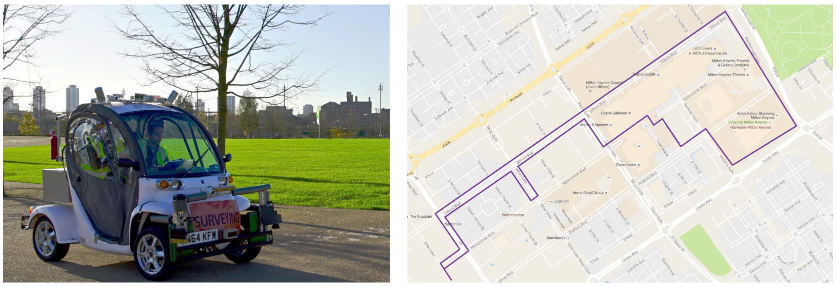


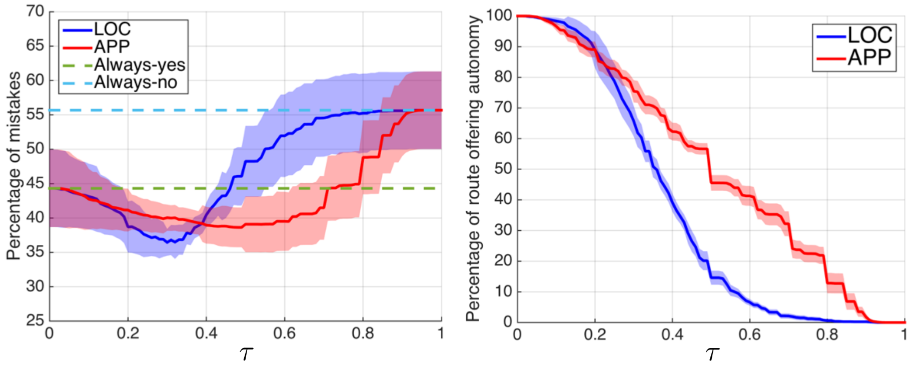
Fig. 4: Figure showing the platform and the route chosen for experiments. The vehicle is equipped with a Bumblebee3 stereo camera, Velodyne Lidars HDL32E and an INS system used for data collection. We produce both 2D and 3D pedestrian detections in image and laser data along the route in Milton Keynes shown on the right.

Estimating performance on a given image first requires localising it against an EBN map and returning the highest scoring localisation candidates. With APP, we build the performance record using observations from these candidates only. We refer the reader to [16] for a comprehensive description of the EBN framework employed.

4 Experiments and Results

We evaluate the two methods proposed for estimating performance, LOC and APP, on 60km of driving data gathered in an urban environment in Milton Keynes over the course of six months. The same route has been traversed eight times under different environmental conditions using the data collection platform shown in Figure 4 and comprise a total of 70k image frames. Some examples can be seen in Figure 1. Since manually annotating such large datasets requires a considerable effort, we make use of a surrogate metric of performance which evaluates the pedestrian detections against laser detections in order to obtain observations neccessary for building the performance record. The image detector used for the experiments presented in this paper is a support vector machine on Aggregate Channel Features (ACF) [17] trained on the INRIA Person dataset [18] following best practice. The laser detector used for providing a surrogate ground truth metric was trained on KITTI Velodyne data [19] and achieves high levels of performance as described in [20]. Note that although we require the laser sensor in order to build the performance record at training time, we do not require the sensor at test time. We estimate performance and take optimal actions either using only the performance record and the location of the robot (required by LOC), or using the performance record and the incoming image feed (required by APP).

In order to evaluate the accuracy of the estimates of performance given by LOC and APP, we analyse the number of wrong decisions the robot takes while employing them. Each image frame that the robot records while driving a test trajectory is used in order to take one of the two decisions: to offer autonomy or to deny it, as described in Section 3. What we refer to as *mistakes* are the outcomes of the following two cases:



(a) (b)

Fig. 5: (a) Figure showing the percentage of total mistakes made by the robot with varying hyperparameter τ. For almost all values of τ, APP has a lower total percentage of mistakes than LOC. (b) Figure showing the percentage of the route that the robot offers autonomy on. The shaded regions in both plots indicate one standard deviation from the mean.

- Choosing to deny autonomy when there are no false positive and no false negative detections in an image (detector performance is perfect but the robot asks for help). These errors are of type I.
- Choosing to offer autonomy when there is at least one false detection in an image (detector performance is not perfect but the robot decides to drive autonomously). These errors are of type II.

Figure 5 shows the results obtained in an evaluation of all traversals in a leave-one-out fashion and an equal cost ($L_{\text{offer}} = L_{\text{deny}}$) for each type of mistake. We show APP having a lower total number of mistakes than LOC (Figure 5(a)) and offering autonomy in a lot more frames (Figure 5(b)), for high values of τ which are the most desirable to use in operation. We attribute this to the fact that APP selects similar observations based on both appearance and proximity, while LOC selects observations based on proximity only. Note that at lower values of τ, both methods are more permissive of driving which leads to more false negative mistakes (failing to recongnise that the perception system is operating poorly), while at higher values of τ, they deny autonomy more often which leads to more false positive mistakes (stopping the vehicle from driving despite it having good performance). Figure 5(a) also shows the total percentage of mistakes produced by always offering autonomy (*Always-yes*) and always denying autonomy (*Always-no*), which are both considerably higher than the methods proposed. This encourages us to believe that if we allow the robot to deny autonomy occasionally, rather than demanding it at times, the overall performance on a task is improved.

Figure 5(b) shows that for an equal cost on type I and type II errors APP is less conservative than LOC and prompts the robot to offer autonomy more often. This is an important advantage as encouraging the robot to take either action can be achieved by adjusting the $L_{\text{offer}}/L_{\text{deny}}$ ratio such that the action which incurs a lower cost will be selected more often (as demonstrated by Figure 3).

	$\mathrm{L_{offer}} = \mathrm{L_{deny}}$			$\mathrm{L_{offer}} = 3 \times \mathrm{L_{deny}}$		
	Type I (%)	Type II (%)	A(%)	Type I (%)	Type II (%)	A(%)
LOC	39.01	**2.27**	11.70	42.75	**0.78**	6.47
APP	**17.28**	15.94	**47.10**	**30.39**	8.26	**26.41**

Table 1: Percentage of mistakes (Type I, Type II) and percentage of route driven autonomously (A) shown for the two methods proposed when 2 different loss matrices are used. The value of τ (hyperparameter at which the action is taken) is set to 0.5. In bold we show that APP has a better outcome than LOC in all cases except for type II errors, which we discuss in the text.

	30% autonomy		50% autonomy		70% autonomy	
	Type I (%)	Type II (%)	Type I (%)	Type II (%)	Type I (%)	Type II (%)
LOC	25.88	12.99	19.13	18.39	4.6	33.07
APP	**21.73**	**12.57**	**17.28**	**15.94**	**4.2**	**29.37**

Table 2: APP makes fewer type I and type II errors than LOC for an equal percentage of the route driven autonomously (30%, 50% and 70%).

Table 1 shows that by increasing the cost of $\mathrm{L_{offer}}$, type II errors for both methods are reduced. Note that in this comparison it appears that LOC makes fewer type II errors. This is because type II errors are computed strictly on the frames on which the decision taken was to offer autonomy, which is to begin with lower for LOC. The percentage of autonomy offered is shown in the table as $A(\%)$. When instead we compute the mistakes made by the two methods for the same percentage of the route driven autonomously (set to 30%, 50% and 70% respectively) APP makes both fewer type I and type II errors. This result is shown in Table 2 for the case of $\mathrm{L_{offer}} = \mathrm{L_{deny}}$.

5 Conclusions

This work proposes a framework for estimating the robot's perception performance at test time based on its performance at previous visits of the same place. Through a classical decision making scenario, we demonstrate that it is possible to reduce the number of mistakes the robot makes by denying autonomy when the performance is predicted to be poor. Selecting past observations from similar environmental conditions further improves our estimates. We believe that performance records can improve with more experience in the same workspace and represent a step towards reliable vision systems operating in the real world.

6 Acknowledgements

The authors gratefully acknowledge the support of this work by the European Communitys Seventh Framework Programme under grant agreement no FP7-610603 (EU-

ROPA2) and by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J012017/1. The authors would also like to thank Dushyant Rao for his helpful suggestions.

References

1. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Neural Information Processing Systems (NIPS)*, 2015.
2. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.
3. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.
4. J. Hawke, C. Gurau, C. H. Tong, and I. Posner, "Wrong today, right tomorrow: Experience-based classification for robot perception," in *Field and Service Robotics (FSR)*, June 2015.
5. C. Gurau, J. Hawke, C. H. Tong, and I. Posner, "Learning on the job: Improving robot perception through experience," in *Neural Information Processing Systems (NIPS) Workshop on "Autonomously Learning Robots"*, Montreal, Quebec, Canada, 12 December 2014.
6. T. Peynot, J. Underwood, and S. Scheding, "Towards reliable perception for unmanned ground vehicles in challenging conditions," in *IROS*, October 2009.
7. T. Peynot, S. Scheding, and S. Terho, "The marulan data sets: Multi-sensor perception in a natural environment with challenging conditions," *The International Journal of Robotics Research*, vol. 29, no. 13, pp. 1602–1607, 2010.
8. A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR'11*, June 2011.
9. A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *European Conference of Computer Vision (ECCV)*, 2012.
10. W. Churchill, C. H. Tong, C. Gurau, I. Posner, and P. Newman, "Know Your Limits: Embedding Localiser Performance Models in Teach and Repeat Maps," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015.
11. J. Dequaire, C. H. Tong, W. Churchill, and I. Posner, "Off the beaten track: Predicting localisation performance in visual teach and repeat," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, May 2016.
12. H. Grimmett, R. Triebel, R. Paul, and I. Posner, "Introspective Classification for Robot Perception," *International Journal of Robotics Research (IJRR)*, 2015.
13. P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh, "Predicting failures of vision systems," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
14. P. Furgale and T. D. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, vol. 27, no. 5, 2010.
15. W. Churchill and P. Newman, "Experience-based Navigation for Long-term Localisation," *The International Journal of Robotics Research (IJRR)*, 2013.
16. C. Linegar, W. Churchill, and P. Newman, "Work Smart, Not Hard: Recalling Relevant Experiences for Vast-Scale but Time-Constrained Localisation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, USA, May 2015.
17. P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection."
18. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 2005.
19. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.
20. D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection," in *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.