Autonomous Semantic Perception in Uncertain Environments

Yuri Feldman and Vadim Indelman
Autonomous Navigation and Perception Lab (ANPL), Technion - Israel Institute of Technology, Israel
yurif@cs.technion.ac.il vadim.indelman@technion.ac.il

I. Overview

Semantic perception is the process of acquiring and maintaining knowledge of the environment of a robot (or more generally - embodied agent) beyond geometric structure, i.e. capturing meaning - such as classes and other high-level properties of visible scene elements - as opposed to pure geometry. Semantic perception is essential in allowing robots to operate in diverse, low-structured and dynamic environments and alongside humans [7, 15, 37]. In the past decade semantic information has become increasingly available for robotics applications thanks to advances in processing of streams of raw data such as images and text, primarily using Machine Learning-based methods, as well as a persistent increase in compute power [7]. However, while established methods exist for estimating and maintaining the geometric structure of a partially observable environment from local measurements (SLAM) [15], these do not readily adapt to treating semantics: Firstly, semantic measurements behave differently from commonly used geometric ones - for example, violating the common assumption of measurement independence [12, 43, 44, 2]. Further, Machine-Learning based semantic measurements and observation models commonly depend on the training data, and cannot be safely assumed correct [12, 41, 36, 22, 26]. Second, semantics often being categorical in nature leads to mixed - continuous and discrete - inference problems that are intractable in their precise form. Thirdly, there is a need in novel expressive state and environment representations to capture the additional rich semantic information in a way enabling high-level scene understanding for downstream tasks [7, 37].

In my research I seek to address the above gaps in the context of autonomous semantic perception and mapping, focusing on "object-centric" [38, 6, 40, 5] perception, with semantic measurements taking the form of detections of objects (or more general elements of the environment) observed by the robot. In this setting, a viewpoint-dependent model can be used to capture spatial variations in semantic measurement vector s for for relative viewpoint $x^{(rel)}$ to object of class c

$$\mathbb{P}(s \mid c, x^{(rel)}). \tag{1}$$

Such models couple between inference of geometry and semantics, permitting mutual disambiguation. They can be represented using e.g. a Gaussian Process (GP) [43, 44] or a Neural Network [19] fit offline to classifier responses for a set of objects representative of each class.

In this context, the contributions of the current work are

- An approach [11, 12] for classification of an object from multiple views under localization uncertainty, accounting for statistical dependence among semantic measurements and shift in data distribution w.r.t. training set (model uncertainty [14, 17]).
- An approach and framework for data-association aware object-centric semantic localization and mapping [42] via maintaining a mixed belief utilizing viewpoint-dependent models.

Subsequently, it turned out that defining a viewpoint-dependent model to be conditioned on a *continuous* object representation rather than on a discrete class could both resolve the need in the intractable mixed inference of semantics and reduce the separate per-class models required previously - to a single observation model, which can be used as a continuous nonlinear factor in inference [8]. In this formulation, each object is represented with an "object embedding" - a continuous vector in a learned latent space, and the predictive viewpoint-dependent model is learned directly from object viewpoints. I am currently aiming to show that a model learned in this way can be used for inference of semantics occurring in the learned latent space, jointly with localization / geometry. Thus, an additional (currently ongoing) contribution is

- A novel formulation [13] of object-centric mapping with inference over a learned continuous semantic representation.

In the following, I provide a brief problem definition, then additional detail on the directions described above, as well as projected future directions.

II. Problem Formulation and Contributed Approaches

The general inference problem I address can be stated as maintaining the posterior, or belief, at time k

$$b[k] \doteq \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O}, \mathcal{C} \mid \mathcal{H}_k), \tag{2}$$

with $\mathcal{X}_{0:k}$ the robot trajectory at time steps $0 \dots k$, object (semantic feature) poses and associated semantic properties (e.g. classes) $\mathcal{O}, \mathcal{C}$ respectively and $\mathcal{H}_k \doteq \{\mathcal{U}_{0:k-1}, \mathcal{Z}_{0:k}\}$ history of user controls and raw observations (e.g. images) respectively. Note that Eq. (2) can be split into:

$$b[k] = \mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O}, \mid \mathcal{C}, \mathcal{H}_k) \cdot \mathbb{P}(\mathcal{C} \mid \mathcal{H}_k), \tag{3}$$

which can be interpreted as a product of a hypothesis over continuous variables $\mathcal{X}_{0:k}, \mathcal{O}$, multiplied by hypothesis weight. Generally the two terms need to be maintained over time for each combination of object classes $\mathcal{C}$ to keep track of possible perceptual aliasing, i.e. of all different hypotheses that may have produced the measurements.

A. Bayesian Viewpoint-Dependent Classification

While semantic measurements are often distinct across viewpoints (e.g. an object is detectable from different directions), they are generally viewpoint-(and inter-viewpoint-) dependent - two identical or similar views do not generally contribute information, violating standard assumptions in (geometric) observation models of statistical independence among measurements, and leading to over-confident inference [43, 12]. In addition, detector/classifier responses may differ across viewpoints. This variation, which in viewpoint-independent models is essentially modeled as noise [1, 30] can actually benefit robot localization if captured by the model [19]. However, related methods generally ignore viewpoint dependence, and the few that do model it ([44], [43]) do not handle partial observability, in particular uncertainty in robot pose. Further, as semantic information is commonly extracted from raw measurements (e.g. images) using Machine Learning based algorithms, difference in deployment environment w.r.t. the training set may lead to out-of-distribution measurements and algorithm failure. This phenomenon known as dataset shift [35, 3] is closely related to AI safety in robotics. It can be addressed by considering model uncertainty in semantic measurements, as provided by Bayesian Deep Learning methods [14, 18, 20, 31, 21, 23, 28], which however is not done by previous semantic mapping methods.

In an initial work [11, 12] we address classification of a single object under model and localization uncertainty. We assume semantic measurements to carry information of model uncertainty, and use Gaussian Processes as class models, to capture spatial inter-dependence of class measurements. For the resultant classification scheme, both synthetic simulation results and subsequent experiments with rendered images show a marked reduction in confident mis-classifications compared to not taking said sources of uncertainty into account.

B. Data-Association Aware Semantic Mapping

In inference over discrete semantics (e.g. classification), belief (Eq. (2)) becomes mixed - over continuous and discrete variables. Inference over such a belief, particularly when attempting to simultaneously address data association, produces mixture models which can quickly become intractable [32, 33]. In subsequent work [42], we addressed semantic mapping in a scene containing multiple objects with unknown data association, expanding over previous work that either assumed data association is given [11, 6, 38], or resorted to approximations such as E-M and similar [27, 5] and the later [9] or max-mixture [10] - the latter being dependent on initialization and limited in their ability to cope with perceptual aliasing, i.e. ambiguity due to different states producing similar observations. In contrast, we show how to maintain the full joint hybrid belief over robot state and object localization, classification and data association, directly controlling approximation accuracy through pruning. In initial experiments on synthetic data we found that utilizing semantic measurements to jointly infer classes and data associations indeed reduced the number of non-negligible components w.r.t. [33], quickening data association disambiguation.

C. Continuous Learned Semantic Representation through a Viewpoint-Dependent Observation Model

Replacing the discrete class variable c in the viewpoint-dependent model Eq. (1) with a per-object continuous embedding vector e, the mixed posterior of Eq. (2) becomes a continuous distribution over geometry and continuous semantic description vectors, $\mathbb{P}(\mathcal{X}_{0:k}, \mathcal{O}, \mathcal{E} \mid \mathcal{H}_k)$, with $\mathcal{E} = \{e\}$ the set of object semantic representations. As shown in [13] this can be developed to allow joint learning of the model $\mathbb{P}(\mathcal{Z}_k \mid e, \mathcal{X}_k^{(rel)})$ and the representations $\mathcal{E}$ of the training set objects. Following the work of Pirk et al. [34], an N-Pairs loss could be used to encourage organization of the latent space according to photometric (and thus emergently, semantic) similarity. Since the learning of the viewpoint-dependent model does not in itself require ground truth semantic information, the implication would be that semantic mapping can be performed without (semantic) ground truth, in particular with no candidate class hierarchy defined in advance or a labeled dataset - solely using detections based on objectness [34, 13]. Semantics could then ideally be recovered from the latent representation using a few user-tagged examples and a simple classification scheme, such as nearest-neighbors.

Bloesch et al. [4] and Sucar et al. [39] equally perform inference in a learned latent space, however, in both the latent space represents shape, or depth measurements - no attempt is done to infer semantics. In particular, the closer related [39] relies on shape prediction to be able to keep pose transformations outside of the deep model, whereas our approach aims to directly model viewpoint-dependent variations in measurements. Thus the proposed representation of semantics is novel w.r.t. to the commonly used per-object classification vectors [29, 10] or dense (surface or volumetric) representations [25, 24] which are the alternative (however it's not immediately clear how a dense representation will be adapted to a non-static environment).

III. Future Directions

As long as no two object instances have similar semantic properties, an assumption of locally correct odometry suffices to learn the model defined in Sec. II-C with no (additional) ground truth data. Further, as described in [13], a re-formulation of the model for conditioning on *change* in relative pose could further relax ground-truth requirements by eliminating the need for an origin to be defined for every object instance. A subsequent research goal therefore is to demonstrate the ability to learn the viewpoint-dependent model from data collected online with no ground-truth then employ it for semantic SLAM. A still longer-term goal, and an important application, is the use of the semantic object descriptors to resolving data association, especially in the context of a dynamic environment. An additional related interesting avenue could be active semantic disambiguation, or Belief Space Planning [16] using the learned model.

References

[1] N. Atanasov, B. Sankaran, J.L. Ny, G. J. Pappas, and K. Daniilidis. Nonmyopic view planning for active object classification and pose estimation. *IEEE Trans. Robotics*, 30:1078–1090, 2014.

[2] Israel Becerra, Luis M Valentín-Coronado, Rafael Murrieta-Cid, and Jean-Claude Latombe. Reliable confirmation of an object identity by a mobile robot: A mixed appearance/localization-driven motion approach. *Intl. J. of Robotics Research*, 35(10):1207–1233, 2016.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[4] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018.

[5] S. Bowman, N. Atanasov, K. Daniilidis, and G. Pappas. Probabilistic data association for semantic slam. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1722–1729. IEEE, 2017.

[6] S. Choudhary, A. Trevor, H. I. Christensen, and F. Dellaert. Slam with object discovery, modeling and mapping. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1018–1025, 2014.

[7] Andrew J Davison. Futuremapping: The computational structure of spatial ai systems. *arXiv preprint arXiv:1803.11288*, 2018.

[8] Frank Dellaert and Michael Kaess. Factor graphs for robot perception. *Foundations and Trends in Robotics*, 6(1-2):1–139, 2017.

[9] Kevin Doherty, Dehann Fourie, and John Leonard. Multimodal semantic slam with probabilistic data association. In *2019 international conference on robotics and automation (ICRA)*, pages 2419–2425. IEEE, 2019.

[10] Kevin J Doherty, David P Baxter, Edward Schneeweiss, and John J Leonard. Probabilistic data association via mixture models for robust semantic slam. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1098–1104. IEEE, 2020.

[11] Y. Feldman and V. Indelman. Bayesian viewpoint-dependent robust classification under model and localization uncertainty. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2018.

[12] Y. Feldman and V. Indelman. Spatially-dependent bayesian semantic perception under model and localization uncertainty. *Autonomous Robots*, 2020.

[13] Y. Feldman and V. Indelman. Towards self-supervised semantic representation with a viewpoint-dependent observation model. In *Workshop on Self-Supervised Robot Learning, in conjunction with Robotics: Science and Systems (RSS)*, July 2020.

[14] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2017.

[15] Sourav Garg, Niko Sünderhauf, Feras Dayoub, Douglas Morrison, Akansel Cosgun, Gustavo Carneiro, Qi Wu, Tat-Jun Chin, Ian Reid, Stephen Gould, et al. Semantics for robotic mapping, perception and interaction: A survey. *arXiv preprint arXiv:2101.00443*, 2021.

[16] V. Indelman, L. Carlone, and F. Dellaert. Planning in the continuous domain: a generalized belief space approach for autonomous navigation in unknown environments. *Intl. J. of Robotics Research*, 34(7):849–882, 2015.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[18] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[19] D. Kopitkov and V. Indelman. Robot localization through information recovered from cnn classificators. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, October 2018.

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6402–6413, 2017.

[21] Keuntaek Lee, Ziyi Wang, Bogdan I Vlahov, Harleen K Brar, and Evangelos A Theodorou. Ensemble bayesian decision making with redundant deep perceptual control policies. *arXiv preprint arXiv:1811.12555*, 2018.

[22] Björn Lütjens, Michael Everett, and Jonathan P How. Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE, 2019.

[23] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 7047–7058, 2018.

[24] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017.

[25] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *2018 International Conference on 3D Vision (3DV)*, pages 32–41. IEEE, 2018.

[26] Dimity Miller, Niko Sünderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: a distance-based loss for training open set classifiers. *arXiv preprint arXiv:2004.02434*, 2020.

[27] Beipeng Mu, Shih-Yuan Liu, Liam Paull, John Leonard, and Jonathan How. Slam with objects using a nonparametric pose graph. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2016.

[28] Pavel Myshkov and Simon Julier. Posterior distribution analysis for bayesian inference in neural networks. In *Workshop on Bayesian Deep Learning, NIPS*, 2016.

[29] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object

detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters (RA-L)*, 4(1):1–8, 2018.

[30] Shayegan Omidshafiei, Brett T Lopez, Jonathan P How, and John Vian. Hierarchical bayesian noise inference for robust real-time probabilistic object classification. *arXiv preprint arXiv:1605.01042*, 2016.

[31] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4026–4034, 2016.

[32] S. Pathak, A. Thomas, and V. Indelman. Nonmyopic data association aware belief space planning for robust active perception. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2017.

[33] S. Pathak, A. Thomas, and V. Indelman. A unified framework for data association aware robust belief space planning and perception. *Intl. J. of Robotics Research*, 32(2-3):287–315, 2018.

[34] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online object representations with contrastive learning. *arXiv preprint arXiv:1906.04312*, 2019.

[35] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT press, Cambridge, MA, 2009.

[36] Charles Richter and Nicholas Roy. Safe visual navigation via deep learning and novelty detection. In *Robotics: Science and Systems (RSS)*, 2017.

[37] David M Rosen, Kevin J Doherty, Antonio Terán Espinoza, and John J Leonard. Advances in inference and representation for simultaneous localization and mapping. *Annual Review of Control, Robotics, and Autonomous Systems*, 4, 2021.

[38] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. Kelly, and A. J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, 2013.

[39] Edgar Sucar, Kentaro Wada, and Andrew Davison. Nodeslam: Neural object descriptors for multi-view shape reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 949–958. IEEE, 2020.

[40] Niko Sünderhauf, Trung T Pham, Yasir Latif, Michael Milford, and Ian Reid. Meaningful maps with object-oriented semantic mapping. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 5079–5085. IEEE, 2017.

[41] V. Tchuiev and V. Indelman. Inference over distribution of posterior class probabilities for reliable bayesian classification and object-level perception. *IEEE Robotics and Automation Letters (RA-L)*, 3(4): 4329–4336, 2018.

[42] V. Tchuiev, Y. Feldman, and V. Indelman. Data association aware semantic mapping and localization via a viewpoint-dependent classifier model. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[43] WT Teacy, Simon J Julier, Renzo De Nardi, Alex Rogers, and Nicholas R Jennings. Observation modelling for vision-based target search by unmanned aerial vehicles. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1607–1614, 2015.

[44] Javier Velez, Garrett Hemann, Albert S Huang, Ingmar Posner, and Nicholas Roy. Modelling observation correlations for active exploration and robust object detection. *J. of Artificial Intelligence Research*, 2012.