SLAM – Quo Vadis?
In Support of Object Oriented and Semantic SLAM

Niko Sünderhauf, Feras Dayoub, Sean McMahon, Markus Eich, Ben Upcroft, and Michael Milford
ARC Centre of Excellence for Robotic Vision, Queensland University of Technology, Brisbane QLD 4001, Australia
niko.suenderhauf@roboticvision.org

Abstract—**Most current SLAM systems are still based on primitive geometric features such as points, lines, or planes. The created maps therefore carry *geometric* information, but no immediate *semantic* information. With the recent significant advances in object detection and scene classification we think the time is right for the SLAM community to ask where the SLAM research should be going[1] during the next years. As a possible answer to this question, we advocate developing SLAM systems that are more object oriented and more semantically enriched than the current state of the art. This paper provides an overview of our ongoing work in this direction.**

I. Introduction and Motivation

Summarized in a single sentence, the problem of *Simultaneous Localization and Mapping* – or SLAM for short – describes the process of a robot building a map of its unknown environment while exploring it. Although [24] considers the first SLAM problem in the scientific literature to date back to the days of Gauss who developed a least-squares method for calculating the orbits of the planets around the sun in 1809 [5], the birth of the modern SLAM problem can be traced back to the ICRA conference in 1986 where the first ideas and thoughts on the subject were discussed. Shortly after that, a number of seminal publications established this new field of robotics research. The acronym SLAM was later coined in [4]. For further references and interesting details on the history of the SLAM problem we refer the reader to [3].

SLAM has been a highly active field of robotics research over the past decades with a myriad of published scientific papers, and a vast number of proposed algorithms and approaches. While in the early days Extended Kalman Filters and later particle filters (see [29] for an overview of such methods) have dominated the field, the community's interest shifted towards smoothing methods based on efficient nonlinear optimization (e.g. [2, 13, 15, 10]) during the past years.

The maps created by most SLAM systems are often based on simple geometric features such as points, line segments [16], or planes [11]. They carry *geometric* but no immediate *semantic* information. An exception is the seminal work by Salas-Moreno et al. [22]. This work proposed a truly object oriented SLAM system by using real-world *objects* such as chairs and tables as landmarks instead of meaningless geometric primitives. [22] detected these objects in RGB-D data by matching 3D models of known object classes.

In the two years after [22] introduced this system, vision-based object detection and recognition has made an impressive performance leap after the re-advent of Convolutional Neural Networks (ConvNets). Starting with [14], several other groups (e.g. [23, 6, 9, 28]) have increased the quality of ConvNet-based methods that have even reached human performance on the standardized ImageNet ILSVRC benchmark [21]. Another area ConvNets currently strive and outperform traditional approaches is the problem of scene categorization [32, 33] that aims at assigning a semantic label such as *living room* or *kitchen* to the whole scene. Even more than the dominance of ConvNet-based approaches in the recent scientific literature, the massive investments of companies like Google, Facebook, Microsoft, and Baidu in advancing these techniques are key indicators of the potential the computer vision and machine learning communities see in these approaches.

Despite these impressive developments, the SLAM community does not seem to have adopted the newly arisen opportunities for their own work. We think the time is right to leverage the recent successes in object detection and scene recognition and work towards more object oriented and semantically enriched SLAM and mapping systems for robotics and autonomous systems.

The paper proceeds as follows: Section II discusses selected areas where we feel more semantic information can be largely beneficial for SLAM. In Section III we introduce some aspects of our currently ongoing work towards a more object oriented SLAM system.

II. More Semantics for SLAM: Thoughts on Some Beneficial Effects

A. What is a Good Landmark for SLAM?

Not all objects in an environment are equally well suited to be used as landmarks for localization and mapping. In fact, what makes a good landmark depends largely on the semantic context and the time frame re-localization should occur in. Knowing the semantics of both the scene and the visible objects allows to estimate or rank the quality of potential landmarks according to learned or pre-defined criteria.

For example, highly dynamic objects such as humans are definitely bad landmarks for localization and SLAM. Other potentially dynamic objects, such as cars, might be useful under certain conditions: Cars in a parking lot are not useful for localization over longer periods of time (e.g. several hours), since they will probably have moved by the time the robot

[1] ”Quo vadis?” Latin for ”Where are you going to?” after the 1896 novel and 1951 motion picture of the same name.

returns. However, over shorter time frames (e.g. 30 minutes) they might provide stable and reliable localization cues. The extent of this time frame also depends on the overall semantic context of the scene and differs between a street scene with cars parked at the side of the road and cars parked in a public parking garage. Also the current daytime might provide further cues, as parked cars during the day tend to be more dynamic than during the night.

How to learn and model which objects provide reliable landmarks, in which semantic contexts, at which time of day, and how to incorporate such knowledge into a working SLAM system are interesting questions for future research. Certainly the recent advances in object and scene classification will play a key role in answering these questions.

B. Providing Scale and Perspective for MonoSLAM / SfM

A common problem in monocular SLAM or Structure from Motion approaches is the unknown scale factor. Typically additional sensors such as IMUs are used to make this missing information accessible. However, when combining purely vision-based monocular SLAM / SfM with an object detection pipeline, classifying the objects in the scene can provide important cues about the distance of these objects and therefore provide overall scale information that are impossible to obtain without additional sensors. This requires knowledge about typical object sizes which can be obtained from training data, modeled with the help of a human expert, or even acquired over time by the combined SLAM / object recognition system itself. Vice versa, knowing the size of an object from SfM or SLAM can provide valuable information for the object classification, and could significantly improve the object detection rate in everyday scenes. We see many possible benefits of combining monocular SLAM / SfM with an object recognition pipeline and will work further towards this goal.

C. High-Level Localization and Place Recognition

Scene categorization or scene classification [32] aims at determining the general semantic class of a scene, such as *office* or *kitchen*. Such semantic information about the currently observed overall scene can provide valuable cues for high-level localization (e.g. localization on scale of individual rooms or larger functional areas in an environment such as a food court or a lobby).

In [26] we showed that using the semantic place category as prior information can help to drastically reduce the search space for place recognition while losing only a small amount of recognition accuracy.

D. Easing Human-Robot Interaction

Semantic information about the environment is an important enabler of more advanced robotic tasks, especially for human-robot collaboration. Humans describe places, goals, and objects using semantic categories and it is natural for them to formulate tasks using these categories [20]: "The kitchen is down the hallway, go there and fetch the milk" [31] or "Pick up this pallet". Semantic knowledge can also modulate the robot's general behaviour, motion primitives, path planning costs, and obstacle avoidance strategies so that it is more compatible with human expectations and requirements. A robot might try to avoid the busy food court during lunch hours, behave differently in a corridor than in an office, or move more carefully through the kitchen than in the living room.

III. Current and Ongoing Work Towards more Object Oriented and Semantic SLAM

In this section we provide a short overview on our current work towards semantically enriched robotic mapping and SLAM. The work described in III-A and III-D has been published in [18] and [27] at the CVPR Workshop on Scene Understanding. The content of III-B and III-C is currently under review [25].

A. Analyzing and Improving Object Proposal Methods

The current state of the art in computer vision for object detection tasks such as the ImageNet [21] challenge (ILSVRC) is to use an *object proposal* step that extracts a number of bounding boxes from an image that *might* contain an object of interest. Each of those bounding boxes is then classified separately by a Convolutional Network. Approaches following this paradigm are for instance [6].

A number of object proposal methods have been proposed, EdgeBoxes [34], BING [1], or Selective Search [30] being among the most prominent. Comparing a variety of such methods, [8] found that EdgeBoxes works best for typical object detection benchmarks. The algorithm mainly relies on the observation that the number of contours that are wholly contained in a bounding box is indicative of the likelihood of the box containing an object. It measures an objectness score by comparing the number of edges within each bounding box with the number of edges passing through it.

Although [8] recommends to use EdgeBoxes for object detection, we found in our own evaluation [18] that the bounding boxes proposed by EdgeBoxes often do not cover meaningful objects well in more realistic real-life scenarios with cluttered scenes. We tested this both on the NYU2 dataset [19] and on a dataset we collected with a robot in our lab and a kitchen environment. Further analysis showed that by refining EdgeBoxes' main parameters `alpha` and `beta`, a significantly better proposal quality can be achieved. For that analysis we defined an objective function that is based on the Intersection-over-Union score (thus measuring how well the proposed bounding boxes cover the ground truth objects) and furthermore weights the proposals by their rank assigned by the core EdgeBoxes algorithm. This follows the intuition that we want the relevant proposals have a high rank, to be able to concentrate further processing on the best few proposals. Fig. 1 illustrates the two dimensional parameter space from that study. We can see that the default parameters (black point) score significantly worse than the best parameter setting (white point) we found through random search. Fig.

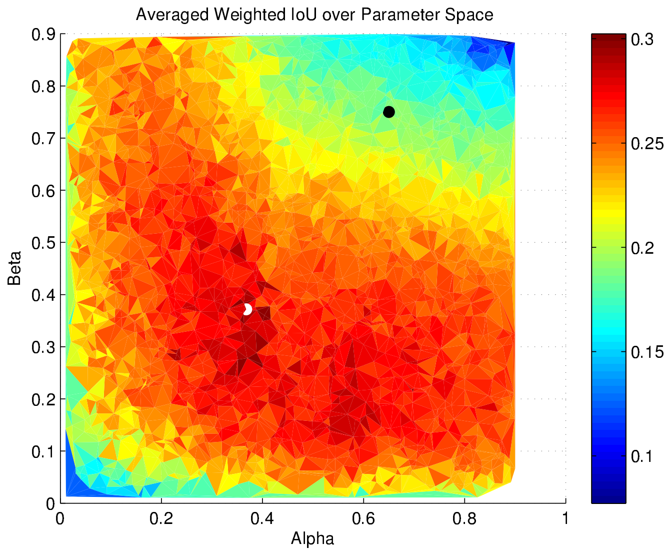

Fig. 1: Parameter space for our performance analysis of EdgeBoxes [34]. Through random search over an IoU-based objective function we found a parameter setting (white) that performs much better on cluttered real-world scenes than the default settings (black).

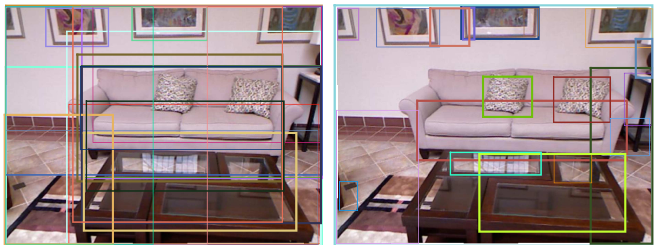

Fig. 2: Left: Object proposal boxes extracted using Edge Boxes' default parameters in a cluttered scene do not cover the relevant objects well. Right: After refining the parameters the quality of the proposals improves both visually and quantitatively.

2 illustrates extracted bounding boxes with the default and refined parameters.

Recently [7] proposed a new object proposal method based on visual saliency that seems to perform much better than EdgeBoxes in cluttered indoor scenes. Since extracting object proposals is the first important step in current object detection pipelines, working towards improvements and generally more robust methods is a worthwhile direction for research.

B. *Creating Semantic Maps*

One aspect of our current research focuses on the question how a standard SLAM system can be combined with a vision-based scene classification system to create semantic maps of the environment. We developed such a system and evaluated it on a real robot. The work is currently under review [25], we therefore only summarize the most important aspects in the following.

The underlying SLAM system we use is the `gmapping` module contained in ROS that maintains a occupancy grid map and uses sensor data from a laser range finder and odometry readings. To classify the currently observed scene, we apply the `Places205` convolutional network [33] that extracts a probability distribution over 205 known class labels (e.g. office, lobby, foodcourt) given a camera image. The probability distribution coming from the convolutional network is then processed by a Bayesian filter for temporal coherence and to incorporate prior domain knowledge about the place categories that can be expected to be observed. The resulting probability distribution is then propagated along the current laser scan and incorporated into the occupancy grid map. We extended this map structure so that it maintains a distribution over scene labels for each map cell. Updates of the scene labels are performed using the usual Bayesian filter update rules for grid maps.

We deployed this system in a variety of places on our university's campus using three very different camera systems mounted on a GuiaBot. Fig. 3 illustrates the resulting variety in visual appearance. By human inspection we evaluated the accuracy of the scene classifier and found that 68% of all camera frames are correctly classified. Fig. 4 illustrates the resulting maps of 9 different places on our campus. Fig. 5 shows a close-up of the map produced in an office environment that also contains a corridor and a kitchenette in the corner.

We could show that a generic ConvNet can be successfully deployed for semantic mapping on a robot without environment-specific training or fine-tuning [25]. We furthermore demonstrated how the created maps could be used to modulate the robot's behavior during simple navigation tasks: The robot was programmed to avoid office areas during work hours to not disturb humans and rather plan longer paths through the corridors. At night time however, the shortest path would always be preferred.

C. *Expandable Place and Object Classification*

A major difference between the computer vision community and robotics is the *closed set* assumption. Most object detection or scene classification benchmarks in computer vision assume that all classes are known during training, and that the classifier is presented only images of one of the known classes during testing [21, 33]. This is called *closed set* classification. However, research in robotics aims at life-long operations and long-term autonomy over extended periods of time. Inevitably, the robot will be faced with scene categories or object classes that were not part of the initial training set, but are important for the robot's mission. Being able to extend the classification framework with new classes during deployment therefore is crucial.

In [25] we show how the place categorization based on the `Places205` network described in the previous section can be expanded by a set of new classes y_i that are not part of the original training set: We propose to train a *one-vs-all* classifier that distinguishes the new class y_i from the already known

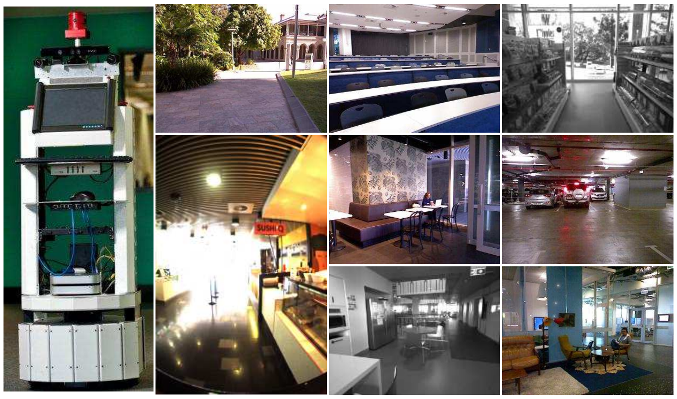

Fig. 3: The Guiabot robot used to evaluate the semantic mapping system and example images from all three cameras in a variety of places: Kinect RGB (color), Grayscale, and Ladybug (portrait format). Notice that all images are resized to a fixed size of 231 × 231 before calculating the ConvNet features. While the change in aspect for the RGB and Grayscale images is minor, the Ladybug image gets squeezed significantly. Also notice the low quality of the Ladybug image.

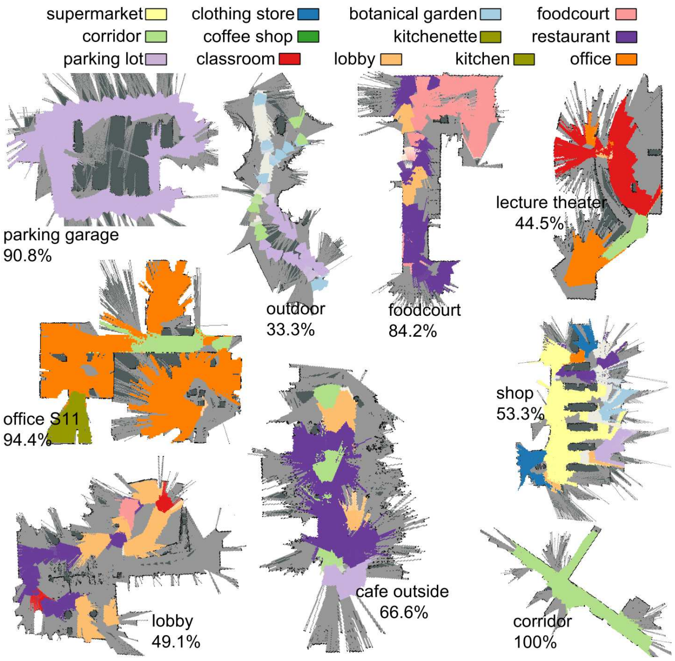

Fig. 4: Maps created by the semantic mapping system in nine different parts of our campus (not drawn to scale). The percentages refer to the fraction of correctly classified *images*, not to the correctly labeled map area.

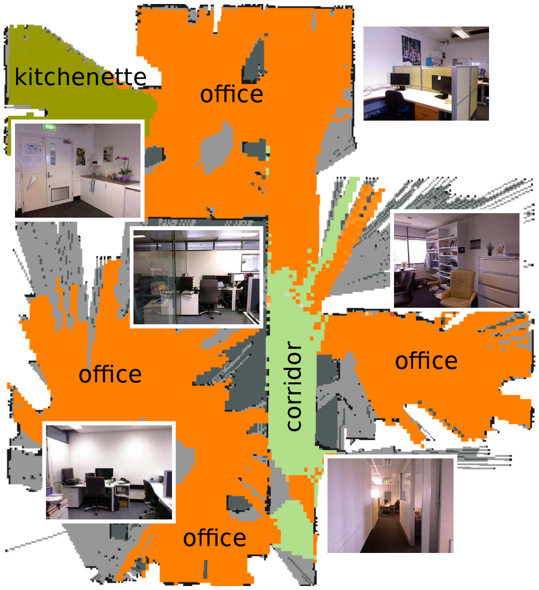

Fig. 5: A semantic map generated by a combination of laser-based `gmapping` and a vision-based scene classifier using the `Places205` ConvNet. The colors encode the semantic categories of different places encountered in the environment. The figure shows a map of an office environment (orange) with a kitchenette (dark green) and a long corridor (light green).

classes $\{x_{0\ldots n}, y_{0\ldots i-1}\}$. The advantage of this approach is that it is not necessary to retrain the ConvNet, which would be computationally expensive (typical training times are in the order of days) and would require a lot of training images (in the order of hundreds or thousands) of the new class. In contrast, a Random Forest one-vs-all classifier can be trained in under a minute using only a few (in the order of 10-100) training images. We let the classifier use the output of the `fc7` layer of the `Places205` network as a feature vector. The `fc7` layer is the last *generic* (i.e. class independent) fully connected layer in the network. The higher layers `fc8` and `prob` have 205 output neurons since they are specifically tailored for the task of recognizing the 205 classes from the training dataset.

As mentioned before, $p(\mathbf{x}_t|\mathcal{I}_t)$ – the discrete probability distribution over $n = 205$ class labels x_i – is the classification result of the `Places205` network, given the current image $\mathcal{I}_t$. Now $p(y_i|\mathcal{I}_t)$ denotes the result of one of the one-vs-all classifiers that is trained to classify the new class y_i. Let

$$\hat{\mathbf{x}} = (x_0, x_1, \ldots x_n, y_0, \ldots, y_m) \tag{1}$$

denote the *combined* vector of class labels. Then we define the combined likelihood $\mathcal{L}(\mathcal{I}_t|\hat{\mathbf{x}}_t)$ as

$$\mathcal{L}(\mathcal{I}_t|\hat{\mathbf{x}}_t) = (p(x_0|\mathcal{I}_t), \ldots, p(x_n|\mathcal{I}_t), p(y_0|\mathcal{I}_t), \ldots, p(y_m|\mathcal{I}_t)) \tag{2}$$

Re-normalization distributes the probability between the n classes known to the ConvNet classifier and the m additional classes known to the one-vs-all classifiers in a natural way. Notice that this assumes independence between the class labels

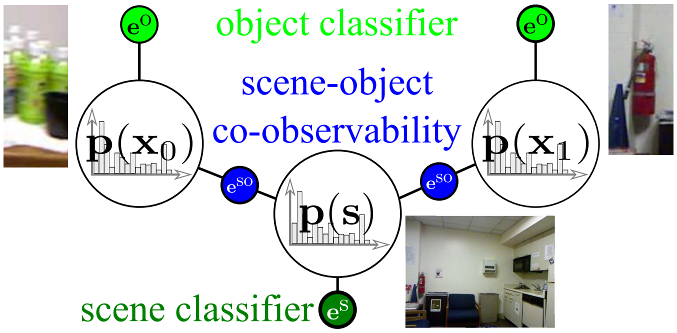
$\mathbf{e}^{\mathrm{O}}$
object classifier
$\mathbf{e}^{\mathrm{O}}$
scene-object
co-observability
$\mathbf{p}(\mathbf{x}_0)$
$\mathbf{p}(\mathbf{x}_1)$
$\mathbf{e}^{\mathrm{SO}}$
$\mathbf{e}^{\mathrm{SO}}$
$\mathbf{p}(\mathbf{s})$
scene classifier
$\mathbf{e}^{\mathrm{S}}$

Fig. 6: We model the scene understanding problem with a factor graph over continuous variables. In contrast to previous work – where the variables are discrete – we can perform exact MAP inference using efficient nonlinear least squares optimization.

$x_{0\ldots n}$ and $y_{0\ldots m}$ as well as pairwise independence between any y_i and y_j.

D. Joint SLAM and Scene Understanding With Continuous Factor Graphs

In the computer vision community, so called *holistic* approaches to scene understanding exploit the rich semantic and spatial relations between individual objects in a scene or between objects and the entire scene to boost the performance of individual object and scene classifiers. *Discrete* graphical models such as conditional random fields (CRFs) are commonly applied to model and solve this problem, e.g. [17, 12].

In contrast to these discrete approaches, *continuous* graphical models dominate in SLAM. Inference in such continuous factor graphs can be conducted via efficient nonlinear least squares optimization. Since we are particularly interested in exploring how SLAM and object detection and scene classification (*scene understanding*) can be combined in one process through joint estimation, we transferred the discrete parts of the joint estimation problem into the continuous domain [27].

Scene understanding aims at finding the optimal discrete label assignment to observed objects x_i and the scene type s, given the observed image and prior semantic knowledge. In order to model and solve this problem with continuous factor graphs, we have to transform it from the discrete into a continuous domain. Instead of creating the graphical model over discrete variables x_i and s, we utilize the *probability distributions* $\mathbf{p}(\mathbf{x}_i)$ and $\mathbf{p}(\mathbf{s})$ as high-dimensional, continuous variables in our formulation: $\mathcal{X} = \{\mathbf{p}(\mathbf{x}_0), \ldots, \mathbf{p}(\mathbf{x}_n), \mathbf{p}(\mathbf{s})\}$. Likewise, we interpret the results of the individual classifiers for the objects and the scene type as *measurements* or *observations*: $\mathcal{Z} = \{\mathbf{z}_0^{\text{object}}, \ldots \mathbf{z}_n^{\text{object}}, \mathbf{z}^{\text{scene}}\}$.

Our formulation corresponds to a probabilistic estimation over probability distributions. The results of the maximum-a-posteriori (MAP) inference therefore are *distributions* over the object and scene classes: $\mathcal{X}^* = \{\mathbf{p}^*(\mathbf{x}_0), \ldots, \mathbf{p}^*(\mathbf{x}_n), \mathbf{p}^*(\mathbf{s})\}$. To retrieve the optimal class label x_i^* for the i-th object, another operation $x_i^* = \operatorname{argmax} \mathbf{p}^*(\mathbf{x}_i)$ is performed, and identically executed for s^*. This is in contrast to the MAP inference step in a CRF where — due to the discrete formulation of the problem — the MAP results are class labels directly [17, 12].

A proof of concept implementation has been recently presented in [27]. The system combined the outputs of a object classification ConvNet, a scene classification ConvNet, and the learned object-scene co-visibility statistics to improve the object recognition rate. At this stage, no SLAM components were implemented. This is ongoing and future work. We are confident that the full model would allow us to jointly estimate the pose and type of objects in the scene, the camera pose, as well as the scene type, while exploiting the semantic and spatial relations between all these variables, building on prior knowledge that is learned from training data or modeled with the help of a human expert.

IV. Conclusions

Our paper gave an overview of our ongoing work towards semantics and object detection for robotic SLAM. We shortly discussed a few selected aspects where we feel a closer connection between object detection, scene classification, scene understanding and SLAM can be highly beneficial. We are convinced that future work into this direction will spawn many interesting research questions and help improve the performance in all of these fields.

Acknowledgements

This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016).

References

[1] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[2] F. Dellaert and M. Kaess. Square Root SAM: Simultaneous Localization and Mapping via Square Root Information Smoothing. *Intl. J. of Robotics Research (IJRR)*, 25(12):1181 – 1203, December 2006. doi: 10.1177/0278364906072768.

[3] Hugh Durrant-Whyte and Tim Bailey. Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. *IEEE Robotics and Automation Magazine*, 13(2):99–110, 2006. doi: 10.1109/MRA.2006.1638022.

[4] Hugh F Durrant-Whyte, D Rye, and E Nebot. Localisation of automatic guided vehicles. In *Proceedings of the 7th International Symposium on Robotics Research*, volume 25, pages 613–625. Springer Verlag, 1995.

[5] Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Hamburg, 1809.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[7] Esther Horbert, Germán Martín García, Simone Frintrop, and Bastian Leibe. Sequence-level object candidates based on saliency for generic object recognition on mobile systems. Proc. of Intl. Conf. on Robotics and Automation (ICRA), 2015.

[8] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. How good are detection proposals, really? In *British Machine Vision Conference, (BMVC)*, 2014.

[9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proc. of ACM International Conference on Multimedia.*, 2014.

[10] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental Smoothing and Mapping with Fluid Relinearization and Incremental Variable Reordering. In *Proc. of Intl. Conf. on Robotics and Automation (ICRA)*, 2011. doi: 10.1109/ICRA.2011.5979641.

[11] Michael Kaess. Simultaneous Localization and Mapping with infinite planes. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2015.

[12] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] K. Konolige, J. Bowman, J. D. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua. View-based maps. *International Journal of Robotics Research (IJRR)*, 29 (10), 2010.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. 2012.

[15] R. Kümmerle, G. Grisetti, and W. Burgard. Simultaneous Calibration, Localization, and Mapping. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 3716 –3721, sept. 2011. doi: 10.1109/IROS.2011.6048393.

[16] T. Lemaire and S. Lacroix. Monocular-vision based SLAM using Line Segments. In *Robotics and Automation, 2007 IEEE International Conference on*, pages 2791 –2796, april 2007. doi: 10.1109/ROBOT.2007.363894.

[17] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2013.

[18] Sean McMahon, Niko Sünderhauf, Ben Upcroft, and Michael Milford. How Good Are EdgeBoxes, Really? In *Workshop on Scene Understanding (SUNw), Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[20] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton CT Lee, Mitchell P Marcus, and Hadas Kress-Gazit. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. In *Robotics: Science and Systems*, 2013.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

[22] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1352–1359. IEEE, 2013.

[23] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[24] Bruno Siciliano and Oussama Khatib, editors. *Springer Handbook of Robotics*. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-23957-4. URL http://dx.doi.org/10.1007/978-3-540-30301-5.

[25] Niko Sünderhauf, Feras Dayoub, Sean McMahon, Ben Talbot, Ruth Schulz, Peter Corke, Gordon Wyethand Ben Upcroft, and Michael Milford. Place Categorization and Semantic Mapping on a Mobile Robot. *under review, arXiv preprint*, 2015.

[26] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the Performance of ConvNet Features for Place Recognition. In *Proc. of IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[27] Niko Sünderhauf, Ben Upcroft, and Michael Milford. Continuous Factor Graphs For Holistic Scene Understanding. In *Workshop on Scene Understanding (SUNw), Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[29] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. The MIT Press, 2005.

[30] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[31] Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. Learning semantic maps from natural language descriptions. Robotics: Science and Systems, 2013.

[32] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, pages 1–20, 2014.

[33] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*, 2014.

[34] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*. European Conference on Computer Vision, September 2014.