Semantic Mapping of Large-Scale Outdoor Scenes for Autonomous Off-Road Driving

Fernando Bernuy[1]
[1]Dept. of Electrical Engineering
Universidad de Chile
Santiago, Chile
fbernuy@ing.uchile.cl

Javier Ruiz del Solar[1,2]
[2]Advanced Mining Technology Center
Universidad de Chile
Santiago, Chile
jruizd@ing.uchile.cl

Abstract

Semantic mapping is a very active and growing research area, with important applications in indoor and outdoor robotic applications. However, most of the research on semantic mapping has focused on indoor mapping and there is a need for developing semantic mapping methodologies for large-scale outdoor scenarios. In this work, a novel semantic mapping methodology for large-scale outdoor scenes in autonomous off-road driving applications is proposed. The semantic map representation consists of a large-scale topological map built using semantic image information. Thus, the proposed representation aims to solve the large-scale outdoors semantic mapping problem by using a graph based topological map, where relevant information for autonomous driving is added using semantic information from the image description. As a proof of concept, the proposed methodology is applied to the semantic map building of a real outdoor scenario.

1. Introduction

In robotics, a semantic map is defined as “a map that contains, in addition to spatial information about the environment, assignments of mapped features to entities of known classes” [9]. Thus, a semantic map contains labels associated to objects (e.g., “tree”, “traffic sign”) and places (e.g., “road”, “building”). Semantic mapping is the process of building semantic maps. Semantic perception is defined as the process of converting sensory observations to this kind of abstractions [4].

Currently, semantic mapping is a very active and growing research area. The semantic representation of the environment is considered essential in the forthcoming unmanned cars, agriculture robotics, and any robotic application requiring human-robot interaction [5]. Some of the reasons for the high interest of the community in those areas are the need for a more robust operation in unconstrained environments and more efficient task execution [12].

The autonomous generation of semantic maps aims to combine the strengths of SLAM techniques and object categorization to recover semantic-spatial knowledge about the environment, which can be used for task planning and inference about non-sensed areas [2]. Current methods are based on map building and application of real-time object categorization techniques over captured data for attaching object categories into the map, usually as a hierarchical structure, generating maps with semantic significance.

However, to the best of our knowledge, most of the research on semantic mapping has focused on indoor mapping or outdoor mapping of restricted outdoor areas. In [3], hierarchies related to spatial and conceptual information are represented using a spatial hierarchy map and a conceptual hierarchy map, in which depth represents different levels of abstraction in semantic content. Nodes between both trees are related via anchoring, which is provided through a user interface. In [6, 8, 18], a multi-layered representation based on a metric map, a navigation map, a topological map, and a conceptual map is used for representing the environment. The metric map is based on an EKF laser-based SLAM, while place classification is based on simple geometrical features extracted from laser scans, and object recognition capabilities are added through the use of SIFT descriptors [7]. Authors use an ontology-based system, and a commonsense OWL ontology of an indoor environment that describes taxonomies (is-a relations) of room types, and typical objects found therein (has-a relations). These conceptual taxonomies are handmade. In [1], Conditional Random Fields (CRF) are used for modeling the map, which is represented as a set of nodes with a hidden state that corresponds to a position plus a category. Categories are related to the kind of objects normally found in street views, mixing visual and laser information. Laser features are computed by generating local descriptors based in angles between laser measurements, and image features are generated by composing different types of features, like local gradient directions, color information, and Haar features. In [16], a dense semantic map is built in an urban environment, using a vehicle equipped with six cameras with different orientations. Each image is segmented with an unsupervised method and labeled into one of the thirteen categories, using CRF. Then, the labeled images are projected assuming a flat world model, creating a labeled metric map, called dense semantic map. In a later work [14], stereo cameras are used to reconstruct the environment surface, replacing the flat world assumption. In a more recent work [15], an octree based 3D map is built using the stereo cameras, and a CRF method is used over the map to label it, creating a 3D occupancy semantic map. The semantic maps proposed by Sengupta et al.[14, 15, 16] archive a great description of the environment but the semantic information included is limited to the labels obtained. Also, due to its construction on a metric map the results are strongly dependent on the localization. The semantic map proposed by Douillard et al. [1] contains a similar amount of semantic knowledge, as it only contains the labels of the different objects detected, but its graph structure allows faster localization tasks.

According to Kostavelis & Gasteros [5], "a challenge for the upcoming endeavors constitutes the semantic mapping of large scale outdoors scenarios", as most of the existing methods aim to solve the problem for indoor topological mapping environments rather than outdoors, and an important aspect in indoor semantic mapping is place and object recognition, which is still underdeveloped for outdoors.

In order to address this challenge, a novel semantic mapping methodology for large-scale outdoor scenes in autonomous off-road driving applications is proposed in this work. The semantic map representation consists of a large-scale topological map built using semantic image information. Thus, the proposed representation aims to solve the large-scale outdoors semantic mapping problem by using a graph based topological map, where relevant information for autonomous driving is added using semantic information from the image description. The proposed mapping methodology is restricted to off-road driving applications where a limited number of object categories is found. Its extension to other driving applications, such as driving in highways or city driving, will depend on a proper definition of the object categories to be stored in the semantic map.

Thus, the main contribution of this work is a graph based topological semantic mapping method suitable for large scale off-road autonomous driving.

This paper is organized as follows. In Section 2, the proposed mapping methodology is described. In Section 3, experiment conducted to test the methodology is presented. Finally, in Section 4 main conclusions are drawn.

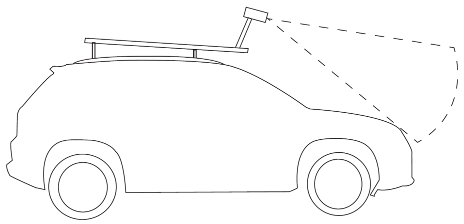

Figure 1. Camera location on the vehicle.

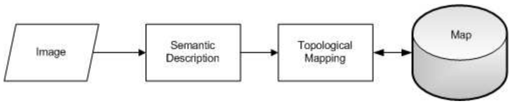

Figure 2. Block diagram of the proposed method

2. Proposed Methodology

The objective is to build a semantic map based on a consistent topological map constructed from images taken with a camera looking to the front of the vehicle (see Figure 1), and fed with high-level information obtained from an online built semantic description of the image (see Figure 2). The proposed method has two main stages: Semantic Description and Topological Semantic Mapping.

In the Semantic Description stage, each image is processed in order to obtain a semantic description of the scene, including the road shape, vegetation and soil around the road, as well as obstacles and objects of interest (e.g. trees, posts, pedestrians, etc.). In the Topological Semantic Mapping stage, the semantic description of the image is used to generate a topological map. This topological map is either added to the global topological map in case that the vehicle is driving for the first time in this area, or used for the vehicle self-localization.

2.1. Semantic Description

This stage aims to create a semantic representation of the images content using a graph structure based on the semantic segmentation of the image as well as context information. Figure 3 shows an example of the semantic description of three different images, and the resulting graph structure for each one. For demonstrative purposes, sketch versions of the images are used in this figure instead of the real images.

Each image is semantically segmented using the TextonBoost method proposed by Shotton et al. [17], a well known multi-class segmentation method that incorporates shape, texture, color, location, and edge cues in a single unified Conditional Random Fields (CRF) model. For this model, the conditional probability for class label c given the image x can be defined as

$$logP\left(c \mid x,\theta\right) = \sum_{i} \psi_i\left(c_i,x;\theta_\psi\right) + \pi\left(c_i,x_i;\theta_\pi\right)$$
$$+\lambda\left(c_i,i;\theta_\lambda\right) + \sum_{(i,j)\in\xi} \phi\left(c_i,c_j,g_{ij}\left(x\right);\theta_p hi\right) \qquad (1)$$
$$-logZ\left(\theta,x\right)$$

ξ is the set of edges in the 4-connected grid, $Z\left(\theta,x\right)$ is a partition function, $\theta = \{\theta_\psi,\theta_\pi,\theta_\lambda,\theta_\phi\}$ are the model parameters, and i,j represent coordinates on the image.

The ψ term in the equation represents the shape-texture potentials, and is based on a boosted combination of texton features. The π term represents the color potentials, and is based on a Gaussian Mixture Model (GMM). The λ represents the Location potential and it works as a look-up table. Finally, the ϕ term represents the edge potential, based on a contrast sensitive Potts model.

As the system is intended to be used in off-road driving applications, the following kinds of objects/labels are used: dirt, grass, road bushes, foliage, sky, tree trunk, post, and pedestrian.

Then, each ground type segment/object (e.g. dirt, grass, and road) is represented by a node in the semantic description graph, and the edges of the graph represent the neighborhood of the segments. Every node contains the following information: (i) type, (ii) spatial position, (iii) objects list, and (iv) traversability index. Additionally, road nodes include (v) curvature index, and (vi) odometry.

- The type of a node is the label acquired from the semantic segmentation, and it defines the kind of ground being represented by the node.

- The spatial position represents the relative spatial position of the node to the road. The node of the first ground segment in front of the vehicle is labeled as ‘Road’, and each other node is labeled as either ‘Left’ or ‘Right’.

- The objects list corresponds to a list of the objects detected inside the segment being represented by the node, which are considered as important for the vehicle (e.g. trees, posts, pedestrians, etc.)

- The traversability index (TI) estimates how dangerous can be for the vehicle to drive over that region being represented by the road [13]. Each label from the semantic segmentation has associated a TI. The nodes TI is calculated as the most dangerous value between the TI of the objects in the node’s object list, and the TI associated to the node’s type.

- The curvature index is added for the road node only, and it is a coarse classification of the observed curvature of the road. This index can take 5 possible values ‘Closed Left’, ‘Left’, ‘Straight’, ‘Right’, and ‘Closed Right’ [10]. The road curvature is obtained by using a Hugh Transform based detection with a set of road curvatures on a flat world model.

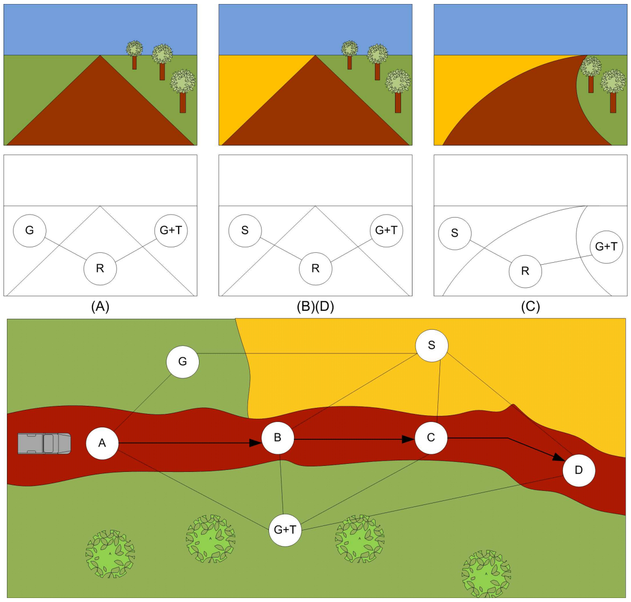

Figure 3. Example of a topological map. The first row shows three different images. The second row shows the semantic description. The third row shows the topological semantic map. For demonstrative purposes, sketch versions of the images are used instead of the real images. In this figure G stands for grass and S for sand. The node G+T has a grass type of ground with trees over it. A, B, C, and D are road nodes. B and D nodes have the same semantic description, the one second one, but they represent different regions of the road.

- Finally, the vehicle’s odometry at the time of the image acquisition is included. This information will be later used to estimate the length of a section in the topological map.

In addition, each node includes pointers to the ’main’ Left and Right nodes, where ‘main’ refers to the closest segment to the vehicle that is bigger than a given threshold in that side.

We define SD_k as the semantic description for the kth processed image, and SD_k.Road the road node of SD_k. Also, $SD_k.Left$ and $SD_k.Right$ are defined as the list of nodes in SD_k, tagged as *Left* and *Right*, respectively.

2.2. Topological Semantic Map

The Topological Semantic Map (TSM) is a graph structure that describes the road and its environment, based on the semantic description of the images described in the former section. The TSM is organized as a sequence of consecutive road nodes and two surrounding sequences of so called *environmental nodes*. A road node represents a region of the road defined by its semantic information. Each environmental node represents the main left or right neighboring node of a group of road nodes. Thus, environment nodes correspond to non-road nodes of the semantic description that are neighbors to the road.

Each road node in the TSM represents a section of the road where the semantic descriptions obtained from consecutive images are the same in terms of same main left node, main right node, and road curvature.

The environment nodes are located correspondently in the TSM using the spatial label of the semantic description (Left or Right) and are connected to its neighboring road nodes, creating a consistent description of the neighborhood of the road in the topological map.

The TSM is built incrementally from the semantic description for the observed image. Let us define M_k the TSM in time step k. M_k summarizes the information of k semantic descriptions SD_i, with $i = 1, , k$. The nodes of M_k have the same basic structure than the nodes of SD_i, but each road node in the TSM includes a *length* value, calculated as the difference between the first odometry observed in that node, and the first odometry from the next road node in the TSM.

Algorithm 1 Map update evaluation

```
function MAPPINGUPDATE($SD_k$,$M_{k-1}$)
    new_road_node ← $false$
    $M_k \leftarrow M_{k-1}$
    $M_k.Road.last.length \leftarrow SD_k.Road.odometry-$
    $first\_odom$
    if $SD_k.Road.curv \neq M_{k-1}.Road.curv$ then
        new_road_node ← $true$
    end if
    if $SD_k.Left.main \neq M_{k-1}.Left.last$ then
        $M_k.add_left(SD_k.Left.main)$
        new_road_node ← $true$
    end if
    if $SD_k.Right.main \neq M_{k-1}.Right.last$ then
        $M_k.add_right(SD_k.Right.main)$
        new_road_node ← $true$
    end if
    if new_road_node then
        $M_k.add_node(SD_k.Road.curv)$
        $first\_odom \leftarrow SD_k.Road.odometry$
    end if
end function
```

The TSM update criterion is detailed in Algorithm 1. Every new map node is added to one of the three lists of nodes: *Road, Left, or Right*, accordingly to its spatial label. The last pointer of each list points to the last added element to that list. The *add_left*, and *add_right* functions add a new environment node to the TSM, adding it to the *Left* or *Right* list, and updating their *last* pointer. The *add_node* function adds a new road node to the TSM, sets its Left and Right nodes as the last from the *Left* and *Right* lists, and updates the *Road* list *last* pointer.

For every new semantic description SD_k, the *MappingUpdate* function updates the current road node length and evaluates if the semantic description's road curvature, main left node, and main right node, are the same as the TSMs last road node curvature, the last left node, and the last right node. If the semantic description's main left node is different to the TSM last left node, then a new environment node is added to the TSM, and the same applies to the right nodes. If a new environmental node was added or if the road curvature is different, then a new road node is added to the map, using the semantic description's road node.

Figure 4 shows an example of a topological map built from the images of the road shown in the map below. The semantic descriptions for the images are presented in the second row, and the topological map is shown in the third row, over a virtual satellite image of the road.

When using this map structure, the vehicle's localization is reduced to the problem of finding the TSM road's node corresponding to the observations obtained by the vehicle. Given that, a vehicle position is defined as the one of the corresponding road's node. The proposed localization method is detailed in Algorithm 2.

Algorithm 2 Proposed localization method.

```
function LOCALIZATION($SD_k$, $globalTSM_k$,
                      $localTSM_k$)
    $localTSM_k$.MappingUpdate($SD_k$)
    if $is\_lost$ then
        for i=1...length($hypothesisList$) do
            if $D(localTSM, globalTSM,$
                    $hypothesisList[i]) \geq T$ then
                $hypothesisList$.remove($i$)
            end if
            new_road_node ← $true$
        end for
        if length($hypothesisList$) $= 1$ then
            $is\_lost \leftarrow false$
            $pose\_estimation \leftarrow hypothesisList(1)$
        end if
        if length($hypothesisList$) $= 0$ then
            $localTSM$.reset()
            $hypothesisList \leftarrow globalTSM.Road$
        end if
    else
        if $D(localTSM, globalTSM,$
              $hypothesisList[i]) \geq T$ then
            $localTSM$.reset()
            $hypothesisList \leftarrow globalTSM.Road$
            $is\_lost \leftarrow true$
        else
            $pose\_estimation \leftarrow localTSM.Road$.last
        end if
    end if
end function
```

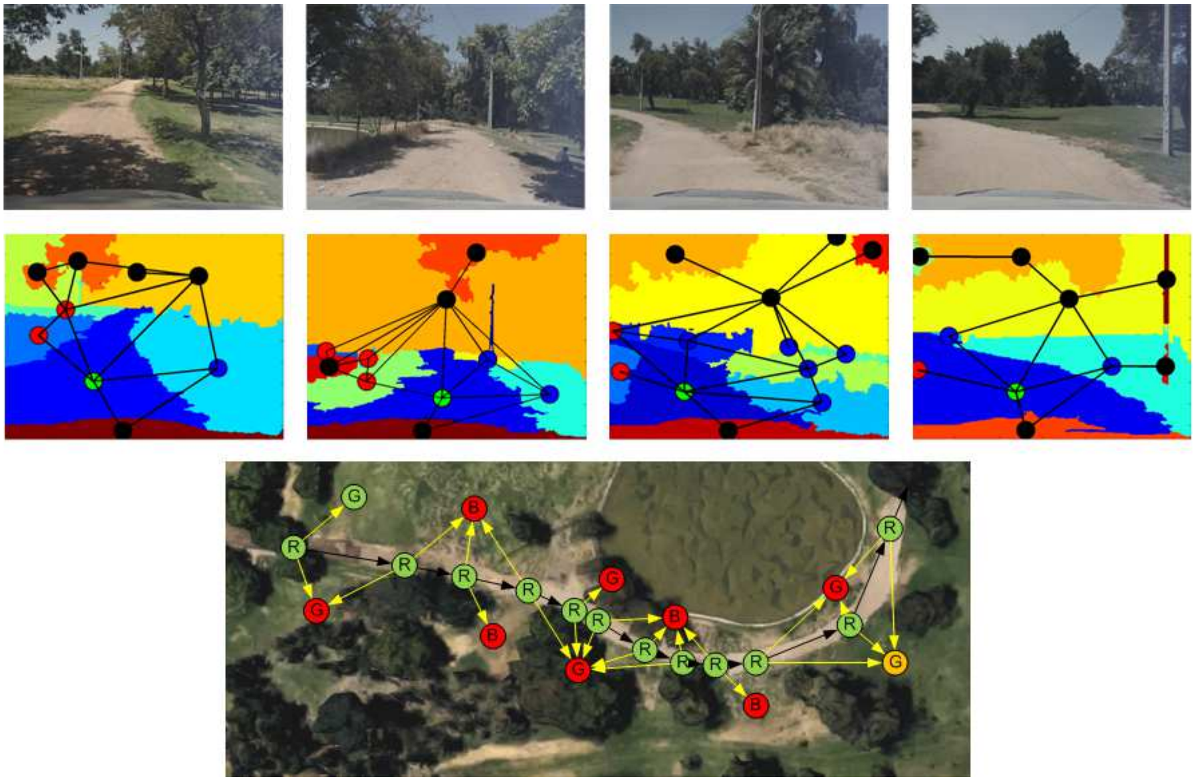

Figure 4. Example of a topological map generated with the proposed method. First row: Exemplar images from the used database. Second row: corresponding semantic description of these images. Third row: semantic map obtained displayed over the satellite image of the road (Source: Google Maps, satellite image).

The vehicle can be in two localization states: (i) located, or (ii) lost. While the vehicle has a TSM pose estimate consistent with the incoming observations, the vehicle is located. The vehicle is lost either if the pose estimate is inconsistent with the observations, or if there is no initial pose estimate.

Two TSM are used for localization purposes: a global TSM, which is a previously built map of the environment, and local TSM, built locally based on the last observations, and using the previously described mapping method.

When a lost condition is detected, the local semantic map is set as empty and the hypothesis list is filled with all the road nodes from the global TSM.

In lost state, the local TSM is compared with every candidate in the hypothesis list using the distance D (see equation 2), in each new observation (a.k.a. image) from the vehicle. If the distance is greater than a given threshold, that hypothesis is removed from the list. If after an observation, the list ends up with only one hypothesis in the list, then the vehicle's pose estimate is set as the hypothesis, and the vehicle is no longer lost. If the hypothesis list ends up empty, then the local TSM is reset, and the hypothesis list is refilled.

In located state, the local TSM only keeps track of the current road node. Here, the local TSM is compared with the estimated pose in the global TSM with the distance D. If the distance is greater than a second threshold, then the vehicle is set as lost, otherwise the pose estimation is updated.

$$D\left(M_a, M_b, k\right) = k_p \cdot P + k_q \cdot \sum_i Q_i. \qquad (2)$$

The distance function D (see equation 2) compares two TSM using the weighted sum of two terms: a P and Q. The P term depends on the length of the current road node of both TSM, and the Q term is proportional to the differences in information for each equivalent node between the TSMs.

3. Experiment

The proposed method is tested using the same database used by Parra-Tsunekawa et al. [11]. The database was recorded with the Advanced Mining Technology Center's (AMTC) autonomous vehicle (Volkswagen Tiguan 2010) inside O'Higgins Public Park, located near the downtown area of Santiago, Chile. The database was captured while the vehicle was driven on unpaved, rough terrain at low

(<10 [m/s]) and medium speeds (<20[m/s]). The unpaved area is formed by a very irregular road having a length of about 800 [m] with positive and negative slopes. The road is a track made of dirt surrounded of grass and some trees. The height difference between the lowest and the highest point of the track is about 4 meters.

For this work, only one segment of the database was used (starting at coordinates (-33.468614,-70.662464)). The image labeling and the object detection were manually annotated. The traversability index considers three levels: low, medium, and high. Grass and dirt regions have high traversability level (TL), while bushes have medium TL. Any obstacle detected close to the road border, at a distance smaller than the road width, has a low TL. Obstacles located at distances between 1 and 2 road widths of the road border have medium TL. Other obstacles have a high TL.

As a proof of concept, Figure 4 shows the results of the application of the proposed method over the used database. Four images of the dataset are shown in the first row. In the second row the semantic descriptions of the images are shown. The third row shows the topological map obtained, over a satellite image of the road. Each node is marked with a letter and color: 'R' stands for Road node, 'G for Grass node, and 'B' for Bushes, while the color represents the associated Traversability Index: green for high, orange for medium, and red for low.

The semantic map obtained is shown over the satellite image of the road to make a qualitative comparison of the topological map. The location of the environment nodes is just a reference. Naturally, only the topologic representation of the environment is the one that needs to be stored.

4. Conclusions

A novel outdoor semantic mapping method based on visual information and topological maps is presented and applied on a complex database.

Figure 4 shows that the resulting topological map is consistent to the satellite image of the road. The resulting map is composed by 12 road nodes and 10 environment nodes, which means that a road segment of about 160 meters is represented by a lightweight structure of only 22 nodes containing high level information of the road and its surroundings.

The results of the proposed method are consistent to the road environment, while keeping a lightweight and computationally efficient structure, making the proposed method suitable for large scale outdoor semantic mapping.

The high level interpretation of the road allows easy and fast ways to include high level deductions to the semantic map, as shown in the results of the experiment in Figure 4. By adding simple rules and expert information to the conceptual map, it is possible to give new kinds of information to an autonomous driving scheme, such as how dangerous can be to take the vehicle to the sides of the road in case of an emergency or an obstruction on the road.

Future work for this method includes automatic image labeling and object detection, the inclusion of range sensors for object detection and scene understanding. Another important challenge to solve for the proposed method is the closed loop case, and how can it be implemented keeping the method free from any global localization systems.

References

[1] B. Douillard, D. Fox, F. Ramos, and H. Durrant-Whyte. Classification and semantic mapping of urban environments. *The international journal of robotics research*, 30(1):5–32, 2011.

[2] C. Galindo, J.-A. Fernández-Madrigal, J. González, and A. Saffiotti. Robot task planning using semantic maps. *Robotics and Autonomous Systems*, 56(11):955–966, 2008.

[3] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. Fernández-Madrigal, J. Gonzalez, et al. Multi-hierarchical semantic maps for mobile robotics. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 2278–2283. IEEE, 2005.

[4] C. Henson, A. Sheth, and K. Thirunarayan. Semantic perception: Converting sensory observations to abstractions. *Internet Computing, IEEE*, 16(2):26–34, 2012.

[5] I. Kostavelis and A. Gasteratos. Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, 66:86–103, 2015.

[6] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen. Situated dialogue and spatial organization: What, where... and why. *International Journal of Advanced Robotic Systems*, 4(2):125–138, 2007.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[8] O. M. Mozos, P. Jensfelt, H. Zender, G.-J. M. Kruijff, and W. Burgard. From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots. In *ICRA Workshop: Semantic Information in Robotics*, 2007.

[9] A. Nüchter and J. Hertzberg. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems*, 56(11):915–926, 2008.

[10] J. W. Park, J. W. Lee, and K. Y. Jhang. A lane-curve detection based on an lcf. *Pattern Recognition Letters*, 24(14):2301–2313, 2003.

[11] I. Parra-Tsunekawa, J. Ruiz-del Solar, and P. Vallejos. A kalman-filtering-based approach for improving terrain mapping in off-road autonomous vehicles. *Journal of Intelligent & Robotic Systems*, 78(3-4):577–591, 2015.

[12] A. Pronobis, A. Aydemir, K. Sjöö, and P. Jensfelt. Exploiting semantics in mobile robotics. In *ICRA 2012 Workshop on Semantic Perception, Mapping and Exploration*, 2012.

[13] A. L. Rankin, A. Huertas, and L. H. Matthies. Stereo-vision-based terrain mapping for off-road autonomous navigation. In *SPIE Defense, Security, and Sensing*, pages

733210–733210. International Society for Optics and Photonics, 2009.

[14] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. Torr. Urban 3d semantic modelling using stereo vision. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 580–585. IEEE, 2013.

[15] S. Sengupta and P. Sturgess. Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order mrf. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1874–1879. IEEE, 2015.

[16] S. Sengupta, P. Sturgess, P. H. Torr, et al. Automatic dense visual semantic mapping from street-level imagery. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 857–862. IEEE, 2012.

[17] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision–ECCV 2006*, pages 1–15. Springer, 2006.

[18] H. Zender, P. Jensfelt, Ó. M. Mozos, G.-J. M. Kruijff, and W. Burgard. An integrated robotic system for spatial understanding and situated interaction in indoor environments. In *AAAI*, volume 7, pages 1584–1589, 2007.