2012 IEEE International Conference on Robotics and Automation
RiverCentre, Saint Paul, Minnesota, USA
May 14-18, 2012

Acquiring Semantics Induced Topology in Urban Environments

Gautam Singh and Jana Košecká

Abstract— **Methods for acquisition and maintenance of an environment model are central to a broad class of mobility and navigation problems. Towards this end, various metric, topological or hybrid models have been proposed. Due to recent advances in sensing and recognition, acquisition of semantic models of the environments have gained increased interest in the community. In this work, we will demonstrate a capability of using weak semantic models of the environment to induce different topological models, capturing the spatial semantics of the environment at different levels. In the first stage of the model acquisition, we propose to compute semantic layout of the street scenes imagery by recognizing and segmenting buildings, roads, sky, cars and trees. Given such semantic layout, we propose an informative feature characterizing the layout and train a classifier to recognize street intersections in challenging urban inner city scenes. We also show how the evidence of different semantic concepts can induce useful topological representation of the environment, which can aid navigation and localization tasks. To demonstrate the approach, we carry out experiments on a challenging dataset of omnidirectional inner city street views and report the performance of both semantic segmentation and intersection classification.**

I. INTRODUCTION

The problem of robot localization and mapping constitutes one of the basic capabilities of autonomous robotics systems and has attracted a lot of attention in the community. There have been a large variety of maps proposed ranging from metric, topological and hybrid maps. Lot of progress has been made in the Simultaneous Localization and Mapping (SLAM) approach for acquisition of metric models with a single reference frame. The topological models represent environments as graphs and differ in how the nodes and connections between them are defined [17], [4], [16]. An example of an on-line topological model acquisition by means of place recognition was tackled in [5] and was instrumental in solving the loop detection problem.

Despite the progress in acquiring different types of representations of the environment, the connections between different types of models were explored in more limited settings. Examples of hybrid metric and topological maps have been introduced in indoors environments, where topology was induced by Voronoi tessellation of the occupancy map or polygonal model of the environment [1], [4]. Numerous approaches to topological mapping although providing the discrete representation of the environment, typically focus on the representation which would maximize the performance of loop detection [5] or place recognition tasks [18] without considerations of how well the topology captures the spatial layout of the environment.

This work was supported by NSF Grant IIS-0347774. G. Singh and J. Košecká are with the Department of Computer Science, George Mason University, Fairfax, VA. {gsinghc,kosecka}@cs.gmu.edu

In earlier works, one of the reasons for endowing models of environments with a topological map, was to attain discrete representation of the continuous space, typically represented as a graph and to enable efficient path planning. The idea of different types of representations of the environment related in a hierachical manner capturing different types of spatial semantic hierarchy has been introduced by Kuipers in [13]. While earlier proposals were more of the conceptual nature, the later works instantiated them in the context of indoors environments, where several methods for building topological maps from metric representations of space were explored [1]. More recent trends in mapping focus on endowing the environments in addition to geometry and topology, with additional semantics. The semantic labels have been either associated with individual locations [24], such as kitchen, corridor, printer room or individual image regions as in [20].

Outline. In the presented work, we address the problem of semantic labeling of street scenes and explore different types of topologies induced by the semantic layout. In the first stage of our approach, we propose to compute the semantic layout by recognizing and segmenting images into buildings, roads, sky, cars and trees. We then propose an informative feature characterizing the layout, which can be used for clustering different locations based on their semantic layout as well as training a classifier to recognize street intersections in challenging inner city scenes. We will show how the evidence of different semantic concepts and intersections can induce useful topological representation of the environment, which can aid navigation and localization tasks. The attained semantic labels can be further used as priors for more refined semantic labeling and object detection or more detailed scene classification. To demonstrate the approach, we carry out experiments on a large-scale dataset of omnidirectional street views reporting the performance of both semantic segmentation and intersection classification.

A. Related work

There is a large body of related work which differs in the choice of the representation, sensing modality used for model acquisition and experimental evaluation. Since in our case we deal with visual sensing, we mention few representative works in metric and topological modeling and semantic understanding. In many instances, it has been demonstrated that visual sensing is a feasible alternative to previous approaches based on laser range data. Existing models acquired by means of visual sensing consider different features, such as sparse set of point features [11] or line

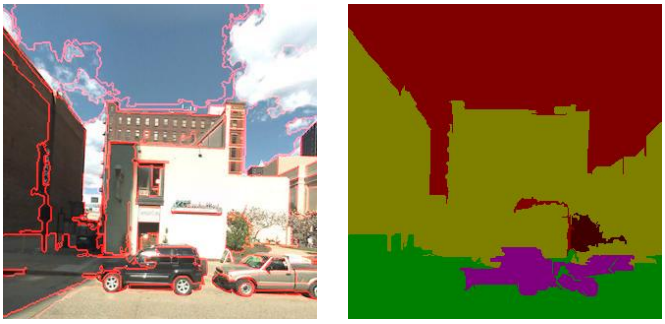
segments [28], [3] acquired by monocular [6] or binocular systems [23]. While maps comprised of dense or sparse cloud of points are often suitable for accurate localization, they are often insufficient for more advanced planning and reasoning tasks. Alternative topological maps often acquired topology in an ad-hoc manner, either grouping neighboring places with similar appearance into a node [12], [26] or considered each location/frame as a node in the graph [5]. In the semantic maps, labels have been either associated with individual locations [24], such as kitchen, corridor, printer room or individual image regions as in [20]. Similar supervised learning strategies to classify the notion of a place/location have been made by [21], along with extensive experiments in how these representations generalize across different environments. In majority of the approaches the features used to infer different semantic categories were derived from both 3D range data and photometric cues. In outdoors settings, the final semantic labeling problem has been formulated as MAP assignment of labels to image regions in the Markov Random field framework [30], where the example labels include road, building, pedestrians, sky, and trees. Prior works in semantic segmentation differ in the choice of primitives they try to label (images, 3D clouds of points), number of semantic categories and the approach. More recent scalable approaches to semantic labeling include works of [27], [10], [14]. The image based approaches typically differ in the choice of the regions, choice of features/statistics computed over these regions and methods for learning the likelihoods of individual semantic labels given the observations as well as pairwise region coherency terms. While the final labeling is often formulated as a MAP problem in random field, the graph structure also differs. The approach of classifying regions of an image into semantic concepts and then using this information for image retrieval tasks has been explored by [29].

B. Semantic Segmentation

In the first stage of our approach, we propose to compute the semantic layout by recognizing and segmenting images into buildings, roads, sky, cars and trees. For the street scene imagery, we use StreetView™ panoramas acquired by a 360^o field of view LadyBug multi-camera system. Our sequence consists of 12,000 panoramas acquired from a run in an urban environment. A single panorama is obtained by warping the radially undistorted perspective images onto the sphere assuming one virtual optical center. The sphere is back-projected into a quadrangular prism to get a piecewise perspective panoramic image, see Fig. 1. Our panorama is composed of four perspective images covering 360^o horizontally and 127^o vertically. The system includes a top camera as well, but it is discarded as it does not provide much information. The panorama is represented by 4 views (front, left, back and right) each covering 90^o horizontal FOV as seen in Fig. 1. We discard the bottom part of all views, containing parts of the car acquiring the panoramas.

The semantic labels we consider are *ground, sky, building, car, tree* and our semantic segmentation approach is most closely related to [9] and [27]. As an elementary region which we will try to classify, we choose the superpixels obtained by color based over segmentation scheme proposed in [8]. This segmentation algorithm typically generates large irregular regions of different sizes see Fig. 2.

Fig. 1. A panoramic piecewise perspective image used in our experiments; four parts (front, back, left and right side) are merged.



(a) (b)

Fig. 2. (a) Example of the color-based over segmentation using method of [8]. Superpixel boundaries are marked by red color; (b) Semantic labelling result for the given over segmentation. Note that due to the crude initial segmentation, a few image regions are misclassified. The color code is the following: building: yellow, car:pink, ground: green, sky: red, tree: brown and void: black.

Since we are interested in learning the coarse semantic layout of the urban environment, we use both geometric as well as appearance features to capture the statistics of individual regions. The choice of features has been adopted from [9] where each superpixel is characterized by location and shape (position of the centroid, relative position, number of pixels and area in the image), color (color histograms of rgb and hsv values and saturation value), texture (mean absolute response of the filter bank of 15 filters and histogram of maximum responses) and perspective cues computed from long linear segments and lines aligned with different vanishing points. Details of the features and pointers to the code for their computation can be found in [9]. In addition to the above features, we endow each superpixel region with a histogram of SIFT descriptors computed densely at each image location and quantized into 100 clusters. The entire feature vector is of 194 dimensions. In order to compute the label likelihood for individual superpixels, we use boosting [22]. Within the boosting framework, we use decision trees as the weak learners since they automatically provide feature selection. We learn separate classifiers for each of the five classes and this is done in a one vs. all fashion. During testing, the separate classifiers are run on the individual feature vectors of the superpixels of an image and output confidence scores. The class with the maximum confidence score is assigned to

be the superpixel's label. In our implementation, each strong classifier has 15 decision trees and each of the decision trees has 6 nodes. An example of the obtained semantic layout is shown in Fig. 3.

(a) (b)

Fig. 3. (a) A side view (b) Semantic labeling obtained for the image using the boosting classifier.

We have annotated a dataset of 320 side and 90 frontal views where each pixel of an image is assigned one of the five classes or *void* if it does not fall into any of the categories. Two separate models are learned, one using the dataset of side views and the other using the frontal views. This is because while the classes may have similar appearance across side or frontal views (e.g. trees are generally green in color), they may not necessarily share the same geometric properties in the two different views. As an example, in a frontal view the buildings are generally observed on the sides with the ground/road in the middle while that is not the case in the side views where they appear in fronto-parallel views. To evaluate the performance of the boosting classifier and compare it to state of the art systems, we use the dataset of 320 side views. The classifier was trained using a randomly selected half of the dataset similar to [32] and the other half of the dataset is used for testing. The results for the boosting classifier and its comparison to the approach of supervised label transfer [32] and non-parametric scene parsing [31] methods on this dataset can be seen in the Table I and II. It is observed that the boosting classifier outperforms the other state of the art systems on this dataset and therefore, we use this classifier through all our experiments for the semantic labeling of an image. While we compare our approach to only two existing methods, these have been shown to be superior to many other systems, such as on the CamVid street scene dataset introduced in [2]. The best performing approach of [25] on CamVid dataset considers more detailed labeling of a total of 11 object and non-object categories. This includes training a more complex likelihood and CRF model and leads to a computationally more expensive inference stage due to the proposed higher order MRF. The appeal of our approach is the simplicity and efficiency of the method, making it applicable to large scale datasets.

Some examples of the results of semantic segmentation can be found in Figure 4.

TABLE I

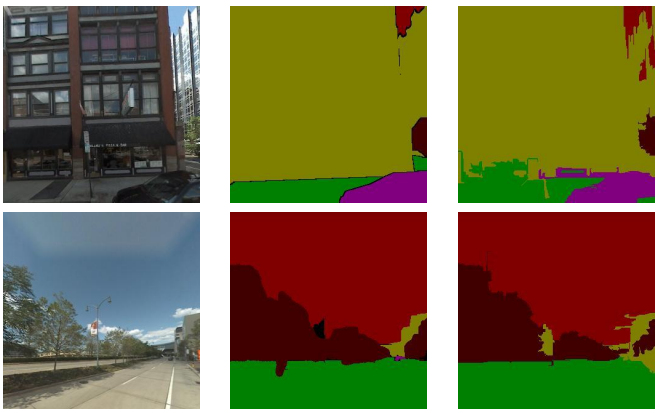
CATEGORY WISE ACCURACY OF BOOSTING CLASSIFIER

System	building	car	ground	sky	tree
[32]	89.1	56.4	89.6	97.1	69.7
[31]	95.3	40.5	96	92.5	41.4
Boosting	96.4	68.3	94.4	97.2	48.9

TABLE II

GLOBAL AND CATEGORY AVERAGE ACCURACY

System	Global	Average
[32]	88.4	80.4
[31]	93.2	73.1
Boosting	94.4	81



(a) (b) (c)

Fig. 4. Examples of semantic labeling of the images. The top row is of a side view while the bottom row is that of a frontal view. From left to right, (a) actual image (b) ground truth labeling (c) predicted labeling

C. Semantic Label Descriptor

To summarize the semantic information in the labeled image, we introduce the *semantic label descriptor*. This descriptor captures the basic underlying structure of the image and can help divide images into sets of visually and semantically similar images. For example, streets inside a city have high rise buildings on the side while highways generally have trees and plants besides the roadside.

For a given image I, we divide I into a uniform $n_k \times n_k$ grid. Within each grid cell, we compute the distribution for each of the five classes using the number of individual pixels in that grid cell which have been assigned that class. This results in a five bin histogram for a single grid cell. The class distribution values for each cell are normalized so that they sum to one. The histograms for the n_k^2 grid cells are concatenated together resulting in a feature vector of length $5 \times n_k^2$. A high value for n_k will capture the details of the layout more precisely but be prone to classification errors while a low value for n_k would be less sensitive to errors in the labeling. In the experiments of this paper, we use n_k = 4 resulting in a 80-dimensional semantic label descriptor.

1) Clustering Topology: We use the semantic label descriptor to cluster the locations of the sequence. While evaluating the performance of our boosting classifier *(Table I and II)*, we had used half of the 320 labelled images for

training and second half for testing. When computing the semantic layout of the entire sequence of 12,000 views the classifier is trained using all 320 side views and run on the left and right side views of each location. The classifier trained using the 90 frontal views is run on the front and back view of each location. Locations for which the ground truth labels are available were excluded from the sequence labeling exercise. The resulting semantic layouts for the four views are then converted into the semantic label descriptor as described above. They are then concatenated together to form a location descriptor for each individual location. Since each individual semantic layout results in a 80 dimensional descriptor, the dimensionality of the location descriptor is 320 (using the four views).

The location descriptors are then used for clustering the sequence. We perform k-means clustering and use cosine distance between the descriptors instead of euclidean distance. Fig. 6 visualizes the average frontal view for a cluster when k=6. It can be noted that the different clusters capture distinct semantic structures. For example, the top row has clusters for areas on highways or with buildings on only one side of the road. In the bottom row, there is a difference in the height of the buildings indicating that some areas have taller buildings than others. A visualization of the results of clustering using the location descriptor for the entire dataset and its non-highway portion can be found in Figure 7 and 8 respectively.

We check the robustness of the clustering by analyzing the cluster assignments of revisited locations. Each location is provided with GPS coordinates specifying the latitude and longitude of that location. Using the GPS coordinates, the individual distance between all locations can be calculated. Any location which has a past location within a threshold distance of 10m is considered a revisited location. In order to avoid considering immediately preceding locations as revisits, we discard the previous 25 frames for a location so that views taken within short time of each other are not considered. Following this, we obtained a set of 3362 revisited locations. For each of these 3362 locations, we obtain its nearest neighbor location from the past. The cluster assignment for a revisited location and its closest past location are checked against each other. The matching rates for the set of revisited locations for different number of clusters is provided in Fig. 5. As can be seen, the cluster assignments maintain a matching rate of more than 75% for a large number of clusters.

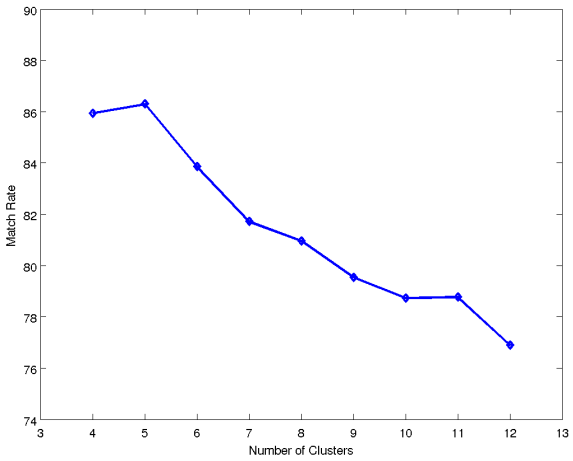


Fig. 5. Match rate between revisited locations and their nearest past neighbors based on cluster assignments

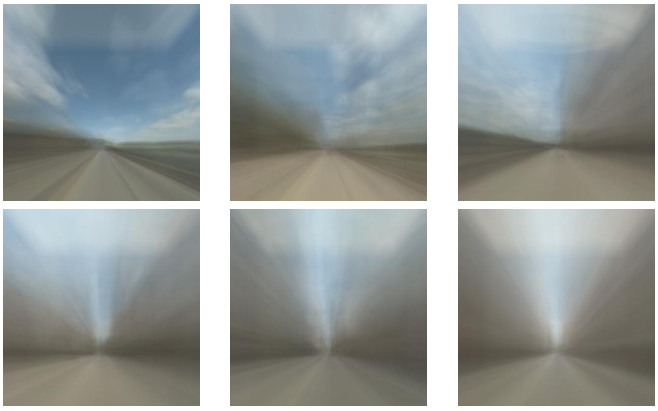


Fig. 6. Visualization of the average frontal view for each cluster(shown for 6 clusters)

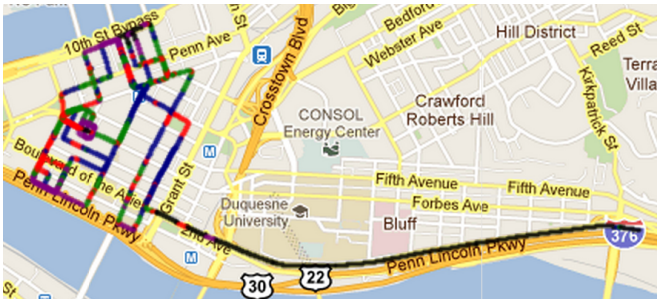


Fig. 7. Clustering visualization for the entire dataset(shown for 5 clusters). Different colors distinguish the cluster assignments for individual locations

D. Intersection Classification

The semantic label descriptor introduced in the previous section was instrumental in grouping different urban regions together based on the presence and layout of different semantic categories in the scene. In this section we show how to infer additional semantic concepts from the attained image representation. In urban environments which can be described as networks of roads and intersections, it is useful to be able to classify a particular view as an intersection or not. The capability of detecting intersections often provides useful prior information of presence of additional semantic concepts, such as pedestrian crossings, stop lights, traffic lights etc. Intersections also correspond to locations where navigations decisions can be made and hence are of interest for automated driving systems. Previous works explored scene classification using either global gist descriptor [19] or spatial pyramid matching [15] and considered more general scene categories like coast, mountain, forest, inside city and highway. In our setting we consider subordinate categories

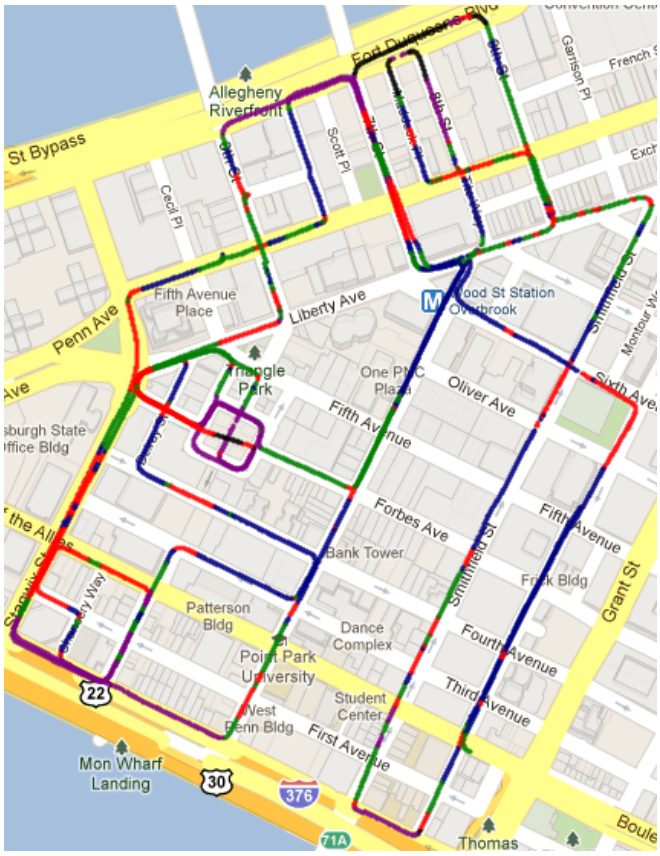


Fig. 8. Clusters for non-highway section of the dataset(shown for 5 clusters). The highway section has been removed in this visualization. Notice the black color assigned to the highway in Figure 7 is missing here except for an area next to the riverfront which is similar to the highway and lacks buildings on the sides

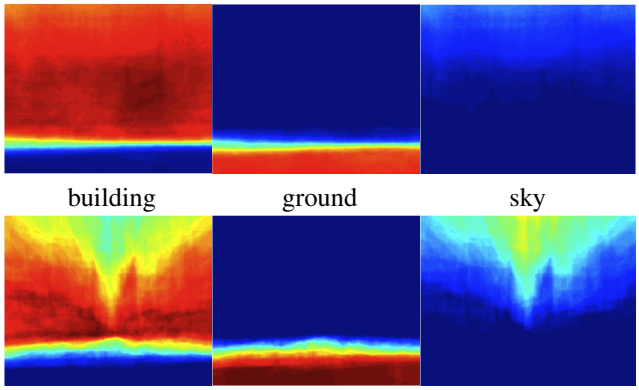


Fig. 9. Top: Probability maps for each label occurring at a pixel at non-intersection side images. Bottom: Probability maps for each label occurring at a pixel at intersection side images. Red indicates a high probability while blue indicates a low probability.

of intersection and non-intersection, which belong to urban scenes but vary in finer spatial semantic layout. More closely related to our approach is the work by [7], where they compute informative features over a grid of patches and train separate classifiers for 13 categories of semantic labels of patches as well as 8 categories of semantic labels of entire scenes, applicable to inner city street scene understanding.

To recognize intersections, we compute an additional normalized histogram of the five semantic labels over the middle part of the image width for side views. This additional histogram is concatenated with the side view's semantic label descriptor to yield a 85-dimensional descriptor and used to train a boosting classifier to classify the side views as intersections or non-intersections. This very simple approach is effective partly due to the 360 degree field of view and availability of the high quality of the semantic labels. The choice of integrating the label statistics from the middle of the side view is motivated by the distinguished appearance of intersections in inner city environments and also the fact that they typically appear at an angle from the main direction of travel. To visualize this intuition, we have computed for the side views (perpendicular to the direction of travel), for each pixel, the probability of a label occurring at that pixel at intersections and non-intersections in Fig. 9. Based on this observation, the extra histogram is computed over 70% of the middle part of each side view.

The 320 side views dataset was annotated for the intersection classifier experiments. Each of the 320 images is manually labelled as an intersection or a non-intersection. This resulted in a set of 250 non-intersection and 70 intersection views. The intersection descriptor is computed for all the 320 side views and another boosting classifier is trained using the resultant 320 descriptors. This boosting classifier has 5 decision trees and each of the decision trees has 4 nodes. This boosting classifier is now run on only the side views of the entire dataset. Locations which contributed images to the training of the intersection classifier were excluded from the test stage. If both the left and right side views of a location are classified as an intersection by the classifier, the location is categorized as an intersection. Otherwise the location is categorized as a non-intersection.

A visualization of this experiment can be seen in Fig. 10. It can be observed that our intersection classifier successfully predicts intersection at many of the major intersections. A human annotator marked 79 unique areas of the sequence as intersections in the city. The intersection classifier correctly predicted an intersection for 63 of the 79 marked intersections for a recall rate of 79.7% indicating the effectiveness of our approach. A successful detection implies that at least two locations within 10m of an intersection were classified as an intersection by the classifier.

E. Conclusions

We have demonstrated an approach for semantic parsing of outdoor urban street scenes acquired by an omnidirectional camera. We have shown how the attained coarse semantic labels (*building, sky, ground, trees, cars*) and their spatial layout can be used to further understanding of street scenes and classifying them as intersections and non-intersections. We have carried out the experiments on a dataset of 12,000 ominidirectional views of urban scenes. The accuracy of semantic parsing has been evaluated on a dataset of 320 ground truth images (with a 50-50 split between training and test sets) yielding a global accuracy of 94%. Intersection

Fig. 10. Visualization of the intersection recognition experiment. Green points mark the locations classified as intersections.

recognition has been tested, by correctly labeling 63 intersections achieving a 79% recognition rate. The visualization of the intersection detection demonstrates that the induced topological model, where majority of nodes marked by intersections are places where different navigation decisions can be made. Recent efforts in world wide scale development of technologies for automated mapping and road network graph construction can benefit from methods of automated street scene understanding. In the future work, we plan to explore strategies for learning more refined semantic concepts as well as an incorporation of stronger temporal constraints between the locations. These advancements will aid understanding of urban scenes for autonomous or semi-autonomous driving applications.

REFERENCES

[1] P. Beeson, J. Modayil, and B. Kuipers. Factoring the mapping problem: Mobile robot map-building in the hybrid spatial semantic hierarchy. *International Journal of Robotics Research*, 29(4):428–459, 2010.

[2] G. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[3] M. Chandraker, J. Lim, and D. Kriegman. Moving in stereo: Efficient structure and motion using lines. In *ICCV*, pages 1741–1748, 2009.

[4] H. Choset and K. Nagatani. Topological simultaneous localization and mapping (slam): toward exact localization without explicit localization. *IEEE Transactions on Robotics*, 17(2):125–137, 2001.

[5] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *International Journal of Robotic Research*, 27(6):647–665, 2008.

[6] A. Davison, I. Reid, N. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 29(6):1052–1067, 2007.

[7] A. Ess, T. Mueller, H. Grabner, and L. van Gool. Segmentation based urban traffic scene understanding. In *BMVC*, 2009.

[8] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[9] D. Hoeim, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

[10] Q. Huang, M. Han, B. Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In *CVPR*, pages 1953–1960, 2011.

[11] K. Konolige, J. Bowman, J. Chen, P. Michelich, M. Calonder, V.Lepetit, and P. Fua. View-based maps. In *RSS*, 2009.

[12] J. Košecká and F. Li. Vision based topological Markov localization. In *ICRA*, pages 1481–1486, 2004.

[13] B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1-2):191–233, 2000.

[14] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. H. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. In *BMVC*, pages 1–11, 2010.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[16] J. Lim, J.-M. Frahm, and M. Pollefeys. Online environment mapping. In *CVPR*, pages 3489–3496, 2011.

[17] M. Mataric. Integration of representation into goal driven behavior-based robots. *IEEE Transactions on Robotics and Automation*, 8(3):304–312, 1992.

[18] A. C. Murillo and J. Košecká. Experiments in place recognition using gist panoramas. In *9th IEEE Workshop OMNIVIS, held with ICCV*, pages 2196–2203, 2009.

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[20] I. Posner, M. Cummins, and P. Newman. Fast probabilistic labeling of city maps. In *RSS*, 2008.

[21] A. Pronobis, O. Mozos, B. Caputo, and P. Jensfelt. Multi-modal semantic place classification. *International Journal of Robotic Research*, 29(2-3):298–320, 2010.

[22] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[23] S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant features. *International Journal of Robotics Research*, 21(8):735–760, 2002.

[24] C. Stachniss, O. Martinez-Mozos, A. Rottmann, and W. Burgard. Semantic labeling of places. In *ISRR*, 2005.

[25] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.

[26] A. Tapus and R. Siegwart. Incremental robot mapping with fingerprints of places. pages 2429–2434, 2005.

[27] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV (5)*, pages 352–365, 2010.

[28] M. Tomono. Robust 3d slam with a stereo camera based on an edge-point icp algorithm. In *ICRA*, pages 4306–4311, 2009.

[29] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72(2):133–157, 2007.

[30] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *ICCV*, pages 686–693, 2009.

[31] H. Zhang, T. Fang, X. Chen, Q. Zhao, and L. Quan. Partial similarity based nonparametric scene parsing in certain environment. In *CVPR*, pages 2241–2248, 2011.

[32] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. In *ECCV (5)*, pages 561–574, 2010.