Semantic 3D Octree Maps based on Conditional Random Fields

Dagmar Lang, Susanne Friedmann, Dietrich Paulus
Active Vision Group, University of Koblenz-Landau,
Universitätsstr. 1, 56070 Koblenz, Germany
{dagmarlang, scfriedmann, paulus}@uni-koblenz.de

Abstract

In this paper we present a 3D semantic outdoor mapping system with multi-label and resolution octree maps based on the OctoMap mapping framework. The semantic labeling of point clouds uses conditional random fields. Speeding up the conditional random field, we use an adaptive graph downsampling method based on voxel grids and the histogram-of-oriented-residuals operator to describe the local point cloud distribution. We validate the proposed classification and map representation approach on real-world 3D point cloud data. The presented classification approach achieves an overall precision about 96 %. The integration of the classification results into the map data structure offers the opportunity to solve complex task settings. Furthermore, the runtime of the presented approach allows an integration of the classification into a real-time 3D semantic outdoor mapping system.

1 Introduction

In recent years, simultaneous localization and mapping (SLAM) algorithms for environment mapping with cameras and 2D or 3D laser range finders have been developed. Most of the created maps are represented as a mixture of metrical and topological data structures. For outdoor environments, mainly object classification approaches for camera, laser range, and fused data based on probabilistic graphical models (PGMs) have been established to create semantic labels as presented in Section 2. Only few approaches combine the semantic labeling and map representation into one mapping system. Such a combination can be used in urban and rural environments for example to adapt the robot's behavior, while inferring from objects in the vicinity.

With this work, we present a reliable and fast 3D semantic mapping system. The semantic labeling task is solved by a pairwise conditional random field (CRF) classifying 3D point clouds without loss of context information. This probabilistic classification method is integrated into an octree-based map data structure. Prior to classification, the CRF parameter vectors have to be learned. Then, a wheeled robot gathers successive point clouds. Creating the map the following steps are performed:

- Each incoming 3D point cloud is classified by the CRF.
- Simultaneously, corresponding point clouds are registered by a SLAM frontend based on the ICP algorithm and a graph is constructed. This graph can be optimized continuously by a SLAM backend especially after loop closing.
- The registered and classified point clouds can now be converted into a multi-label and resolution octree map.

The SLAM frontend and backend are out of the scope of this paper. In our experimental section, we show the result of the mapping system on a real-world dataset.

The remainder of this paper is organized as follows. First, we discuss the related work for point cloud classification and map representation in Section 2. Then, we present in Section 3 an extension of the OctoMap data structure (available at http://octomap.sf.net) modeling label probabilities determined with a CRF. The experiments and results are depicted in Section 4. We draw our conclusion and point out future work in Section 5.

2 Related Work

Two different tasks has to be fulfilled to create a semantic mapping system. On the one hand, a semantic classification has to be developed. On the other hand, a suitable data structure for map representation has to be developed, which is suitable for additional task settings based on the map and allows the integration of the semantic labels.

There are different approaches to model the environment based on 3D clouds. Popular representations include point clouds, voxel grids, octrees and surfels. For semantic mapping, a data structure has to afford the opportunity to model occupied, free and unknown space. Furthermore, it should be memory-efficient, allow multi-resolutions and the integration of labels. In the following, only approaches that partially meet the requirements will be presented. Marton et al. [5] presented an fast and robust triangulation algorithm for unorganized 3D point clouds, which can deal with different labels. An approach called OctoMap for modeling large scale 3D environments based on octrees using a probabilistic occupancy estimation was introduced by Wurm et al. [10]. An extension to hierarchies of octrees was presented by Wurm et al. [11] with object labels in different resolutions. The approach cannot handle different labels in one voxel.

One of the first approaches to classify 3D point clouds was based on associative Markov networks and was presented by Anguelov et al. [1]. An extension of this approach was proposed by Triebel et al. [9], where the authors showed that adaptive data reduction not necessarily influences the classification results. A contextual classification of 3D point clouds or camera data using a linear associative max-margin Markov network approach was presented by Munoz et al. [6]. The authors adapted a functional gradient approach to learn high-dimensional parameters of random fields in order to perform discrete, multi-label classification.

Lim and Suter [4] proposed an adaptive data reduction method and used discriminative CRFs for 3D point cloud classification. They showed that smaller sets of data samples containing relevant information within the support region of super-voxels produce similar results as using the whole point cloud for classification. A classification approach based on pairwise CRFs to segment terrestrial LIDAR point clouds was proposed by Niemeyer et al. [7]. The approach provides the opportunity to incorporate contextual information and learning of specific relations of label classes.

3 Semantic 3D Octree Maps

In the following, we first explain the OctoMap approach of Wurm et al. [10] and then our extension to a multi-class CRF based OctoMap. A octree is representing a hierarchical data structure dividing the 3D space into spatial subdivisions. The OctoMap mapping framework is based on octrees and creates a voxelized 3D map for registered 3D point clouds.

Each node of the octree is a cubic volume named voxel. The whole volume is recursively subdivided into eight partial voxels with the same size until a minimum size for each voxel is reached. This minimal size defines the resolution of the octree. Depending on the robotic application, the octree of the OctoMap data structure can be traversed to a coarser resolution.

The occupancy probability $P(n|z_{1:t})$ of each node n is estimated using a Bayes filter as

$$P(n|z_{1:t}) = \left[1 + \frac{1 - P(n|z_t)}{P(n|z_t)} \frac{1 - P(n|z_{1:t-1})}{P(n|z_{1:t-1})} \frac{P(n)}{1 - P(n)}\right]^{-1}, \quad (1)$$

where $z_{1:t}$ are all sensor measurements from time point 1 to t. The calculation of the occupancy probability, adding new sensor data, is performed only for the maximum node resolution of the tree.

To obtain a map with different resolutions, the inner nodes n_{in} of the octree can be updated by the calculation of the mean occupancy probability $\hat{l}_\mu$ or the maximum occupancy probability $\hat{l}_{\max}$ as

$$\hat{l}_\mu(n_{\text{in}}) = \frac{1}{8}\sum_{i=1}^{8} L(n_i) \quad \text{or} \quad \hat{l}_{\max}(n_{\text{in}}) = \max_i L(n_i).$$

Although, $L(n_i)$ is the *logOdds* of $P(n_i|z_{1:t})$, where n_i is the child of n_{in}.

The presented approach of Wurm et al. [11] incorporate different labels into the OctoMap is based on hierarchies of octrees. For each label and object, a new OctoMap is created, such that multi-resolution object maps can be created and objects can be represented in a finer resolution than less interesting ones. The different OctoMaps are connected by a tree. The authors assume a precise classification of the point cloud and do not offer a solution for 3D points with different labels in one voxel. A solution for this multi-label problem in one voxel, with loss of different resolutions for different objects is presented in the following.

Probability based Multi-Label Octree The definition of the multi-label octree is based on the assumption that we can apply any arbitrary PGM to classify 3D points of a point cloud into different semantic labels. We use a pairwise CRF for classification and the inference method for super-voxels of Lim and Suter [4] can be applied to the octree voxels. Here, the probability $P(n, y_{\max})$ of the likeliest class $y_{\max}$ of each node n can be calculated by

$$P(n, y_{\max}) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{\boldsymbol{x}_i \in n} P(\boldsymbol{y}|\boldsymbol{x}_i), \quad (2)$$

where $P(\boldsymbol{y}|\boldsymbol{x}_i)$ is the conditional probability modeled by the CRF. The probability is estimated for a sequence of random variables $\boldsymbol{y}$ of object labels, given all random variables $\boldsymbol{x}_i$ representing 3D points in node n of the point cloud. To determine the likeliest label for the inner nodes n_{in} of the octree, the following update rule $\hat{p}(n_{\text{in}}, y_{\max})$ is applied:

$$\hat{p}(n_{\text{in}}, y_{\max}) = \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{j=1}^{8} \prod_{\boldsymbol{x}_i \in n_j} P(\boldsymbol{y}|\boldsymbol{x}_i) \quad (3)$$

The conditional probability for each child node n_j of n_{in} is calculated and propagated upwards in the octree.

CRF based Semantic Object Classification We use a pairwise multi-label CRF to calculate $P(\boldsymbol{y}|\boldsymbol{x})$. CRFs are discriminative undirected graphical models often applied to sequence labeling problems. The input consists of a fully observed data sequence, e.g. the features for every 3D point in a point cloud. The CRF models the relationship between these observations and assigns a label from a finite set of learned classes for each given feature. The pairwise multi-label CRF model is based on [7] and is defined as

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_{i \in n} \boldsymbol{w}_l^T \Phi(f_i(\boldsymbol{x})) + \sum_{i \in n} \sum_{j \in \text{MB}_i} \boldsymbol{w}_{l_i, l_{\text{MB}_i}}^T \Phi(f_{ij}(\boldsymbol{x}))\right), \quad (4)$$

with the partition function $Z(\boldsymbol{x})$. In contrast to all features f of the data sequence $\boldsymbol{x}$, the association potential $\Phi(f_i(\boldsymbol{x}))$ determines the likeliest label $\boldsymbol{y}_i$ for each node in the graph. $\Phi(f_i(\boldsymbol{x}))$ and the label sequence depending parameter vector $\boldsymbol{w}_l^T$ will be calculated for the number of nodes n in the graph and its elements i. The interaction potential $\Phi(f_{ij}(\boldsymbol{x}))$ and the corresponding parameter vector $\boldsymbol{w}_{l_i, l_{\text{MB}_i}}^T$ model the relationship between node i and node j in the graph. The neighborhood of node $\boldsymbol{x}_i$ is defined by the Markov blanket MB_i. For node labels y_i and y_j of MB_i, the edge feature vector $\boldsymbol{\mu}_{ij}$ is determined by the difference of the feature vectors f_i and f_j depending on $\boldsymbol{x}$. Similar labels are preferred by the interaction potential $\Phi(f_{ij}(\boldsymbol{x})) = \delta_{ij}\boldsymbol{\mu}_{ij}$ by the Kronecker's delta δ_{ij}. We train the CRF using pseudo log-likelihood training with an optimization by the L-BFGS algorithm [8]. For inference, we run loopy belief propagation with residual message update schedule as proposed in [2] until convergence.

Graph Downsampling In the literature, point based CRFs, where each 3D point in the cloud gets connected to its k-nearest neighbors, yield good and robust classification results. One drawback is the complexity of the graph structure which leads to expensive

and slow computations for large point clouds. In this case, a reduction of the cost was achieved by downsampling the point cloud, which however leads to a loss of information, especially in the presence of small objects.

In order to keep as much information as possible, we downsample the point based CRF graph by using a voxel grid with an adaptive cell size. The basic voxel grid consists of metrically equidistant voxels in each dimension and the nodes are integrated into the voxels. For each voxel, we compute the center of mass for all points in the voxel. The center of mass then becomes a new node in the CRF graph and the other nodes in the voxel are removed. Since the structure of the voxel grid is fix, we loose a lot of information, if the boundary of the grid passes through small objects. Therefore, we perform a merge step, if the Euclidean distance between neighboring voxel nodes is smaller than the distance between their geometric voxel centers. We recompute a new node and the center of mass for the merged voxels. Now, each voxel gets connected to its k-nearest neighbor voxel. After the adaptive downsampling, the graph is reduced by about 20 % of its original size.

Voxel Descriptor As main feature, we use the "histogram-of-oriented-residuals" (HOR) operator introduced by Krückhans [3] for facade and ornament detection. For each node $\boldsymbol{x}_i$ of the downsampled graph and the corresponding 3D point $\boldsymbol{p}_s$, the descriptor determines all points $\boldsymbol{p}_{\mathcal{N}_j}$, $j = 1, \ldots, n$, in a local neighborhood $\mathcal{N}$, by searching for all neighbors in a given radius in the point cloud. The search radius is initially fixed, but will be adapted by the adaptive graph downsampling method such that the merged voxel is enclosed by the volume created by the radius.

The HOR operator uses the difference between points and planes, so called residuals, to characterize local regions in a point cloud. Therefore, m planes $\boldsymbol{q}_{ai} = \boldsymbol{R}_a\left(i\frac{2\pi}{m}\right)\boldsymbol{e}_{a_k}$ are defined by rotation $\boldsymbol{R}_a\left(i\frac{2\pi}{m}\right)$ around axis a with incrementally increased step sizes $i = 0, \ldots, m$ and the unit vector $\boldsymbol{e}_{a_k}$ of a corresponding axis a_k. The residual $r_{ai\mathcal{N}_j}$ for axis a, step size i and point $\boldsymbol{p}_{\mathcal{N}_j}$ of the neighborhood are calculated as

$$r_{ai\mathcal{N}_j} = \left\langle \begin{pmatrix} \boldsymbol{p}_{\mathcal{N}_j} \\ \|\boldsymbol{q}_{ai}\|^{-1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{q}_{ai} \\ -\langle \boldsymbol{p}_s, \boldsymbol{e}_{a_k} \rangle \end{pmatrix} \right\rangle \qquad (5)$$

The first vector of the scalar product is representing one point of the neighborhood, which is density invariant according to the fourth entry and the second vector is representing the Hesse normal form of the plane. The residuals $r_{ai\mathcal{N}_j}$ are calculated for all planes around axis a, number of steps m and points $\boldsymbol{p}_{\mathcal{N}_j}$. They are summarized into histogram $\boldsymbol{h}_a$.

The original descriptor is based on rotation $\boldsymbol{R}_z$ with $\boldsymbol{e}_x$, summarized in $\boldsymbol{h}_z$, and performs well in separating the label *ground* from *building*. Improving its discriminative strength, we added two additional rotation axes $\boldsymbol{R}_x$ with $\boldsymbol{e}_y$ and $\boldsymbol{R}_y$ with $\boldsymbol{e}_z$ to the original descriptor, such that $f(\boldsymbol{x}_i) = (\boldsymbol{h}_z, \boldsymbol{h}_y, \boldsymbol{h}_x)$.

4 Experiments and Results

We tested our semantic multi-label mapping algorithm on the Freiburg dataset.[1] The dataset was captured using a wheeled robot equipped with a SICK LMS laser range finder mounted on a pan-tilt unit and consists of 77 3D scans capturing an area of 292 m × 167 m × 28 m. Each 360° scan was acquired in a stop-and-go fashion and consists of 150,000-200,000 points.

[1] http://ais.informatik.uni-freiburg.de/projects/datasets/fr360/

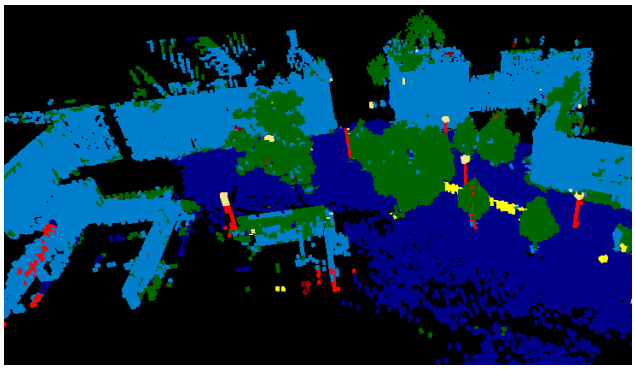

Figure 1. A part of the CRF based octree map with a voxel resolution of 0.4 m (best viewed in color).

Results for CRF based Semantic Classification The dataset was classified into the semantic labels *ground*, *building*, *vegetation*, *column*, *street lamp* and *bollards*. The results for each scan of the dataset were achieved using the adaptive graph downsampling method with an initial voxel size of 1 m × 1 m × 1 m and a Markov blanket with 6 neighbors. The search radius for the extended HOR operator for all neighboring 3D points of the voxel center was set to 2.5 m and the histogram for each rotation axis was calculated with 15 bins, leading to a final region descriptor of size 45.

We evaluated the performance of our classifier using small subsets of the point clouds, representing each class, to train the CRF model. Point clouds not involved in training where used for classification evaluation. The ground truth was annotated by hand for the dataset. *Ground* was classified with a precision >99 %, *building* with 96 %, *bollards* with 89 %, *vegetation* with 86 % and *columns* with 72 %. *Street lamps* reached only a precision of 27 % and are often confused with the semantic label *building*. Less frequently misclassifications of *columns* as *building* or *building* as *street lamp* are present. The approach reaches an overall precision for the Freiburg dataset of 96 %.

Results for Multi-Label Octree Maps We evaluated the performance of our CRF based semantic mapping approach by measuring the runtime for all presented algorithms. Therefore, we run the classification process for the entire dataset five times. A comparison of the runtimes for one point cloud is presented in Table 1. The HOR calculation includes the downsampling time as well as the time for feature calculation. Creating an original OctoMap data structure (Equation 1) needs 25.25 MB memory for the Freiburg dataset with a voxel resolution of 0.1 m and for the CRF multi-label octree map (Equation 2) 151.25 MB. A part of the CRF based octree map is presented in Figure 1 and shows the above mentioned classification results. The color-coding is wrt. to the ground truth (light blue = *building*, dark blue = *ground*, green = *vegetation*, red = *column*, violet = *bollard*, yellow = *street lamp*). The entire CRF multi-label octree map

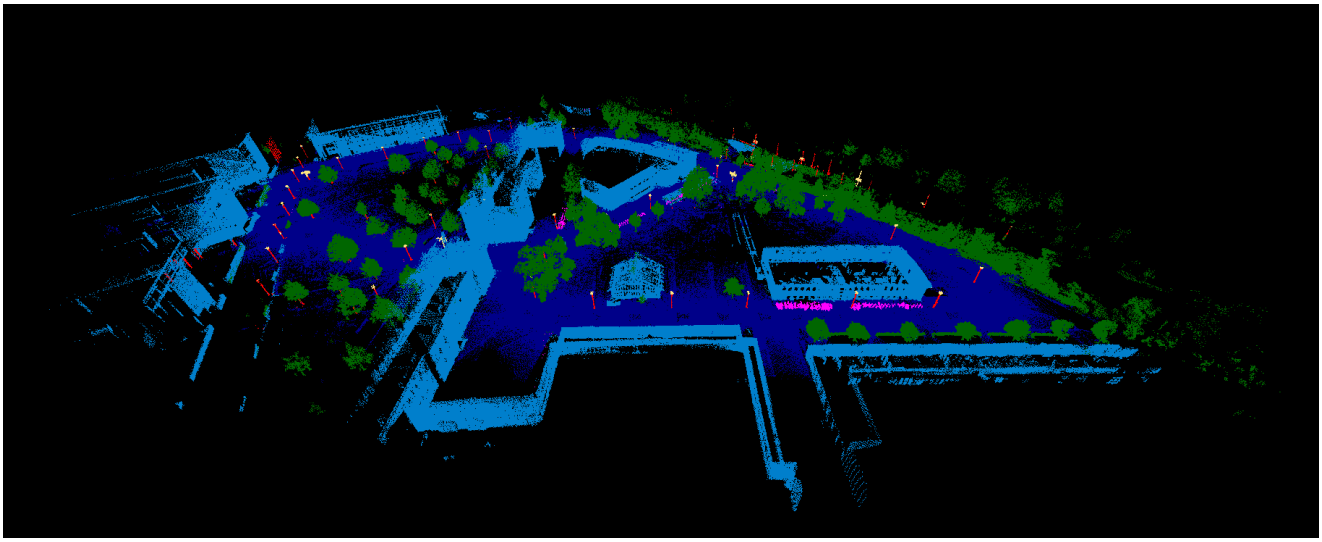

Figure 2. The entire CRF based octree map with a voxel resolution of 0.1 m (best viewed in color).

Method	Mean	Max
HOR calculation	566 ms	795 ms
Inference	804 ms	841 ms
Insertion of one point into the multi-label map	$1.9{\cdot}10^{-4}$ ms	0.065 ms
Creation the complete multi-label map	2030 ms	2163 ms

Table 1. Runtime results.

for the Freiburg dataset is shown in Figure 2.

5 Conclusion and Future Work

In this paper we presented an extension of the OctoMap data structure, which offers the opportunity to create large scale 3D semantic maps in a representative compact data structure. The semantic classification was based on a pairwise CRF with an adaptively downsampled graph and the scale and density invariant HOR operator as feature. We have shown that the runtime allows the integration of the semantic classification into a robotic mapping system in a stop-and-go fashion and also achievs a high precision in the classification. The classification results provide a good start to incorporate the mapping results for geometrical scene interpretation and e.g. learning new robot behaviors or developing a path planning based on the created map.

Acknowledgement

This work was partially funded by Wehrtechnische Dienststelle 51 (WTD), Koblenz, Germany.

References

[1] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz and A. Y. Ng, *Discriminative Learning of Markov Random Fields for Segmentation of 3D Scan Data*, In Proc. of CVPR, pages 169–176, 2005

[2] G. Elidan, I. McGraw and D. Koller, *Residual Belief Propagation: Informed Scheduling for Asynchronous Message Passing*, In Proc. of UAI, 2006

[3] M. Krückhans, *Ein Detektor für Ornamente auf Gebäudefassaden auf Basis des "histogram-of-oriented-gradients"-Operators*, Master's thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 2010

[4] E. H. Lim and D. Suter, *3D Terrestrial LIDAR Classifications with Super-Voxels an Multi-Scale Conditional Random Fields*, Computer-Aided Design, 41(10):701–710, 2009

[5] Z. C. Marton, R. B. and M. Beetz, *On Fast Surface Reconstruction Methods for Large and Noisy Datasets*, In Proc. of ICRA, 2009

[6] D. Munoz, J. A. Bagnell, N. Vandapel and M. Hebert, *Contextual Classification with Functional Max-Margin Markov Networks*, In Proc. of CVPR, 2009

[7] J. Niemeyer, F. Rottensteiner and U. Soergel, *Conditional Random Fields for LIDAR Point Cloud Classification in Complex Urban Areas*, In Proc. of ISPRS, vol. I-3, pages 263–268, 2012

[8] N. Okazaki. *libLBFGS: A Library of Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)*, 2010.

[9] R. Triebel, K. Kersting and W. Burgard, *Robust 3D Scan Point Classification using Associative Markov Networks*, In Proc. of ICRA, pages 2603–2608, 2006

[10] K. M. Wurm, A. Hornung, M. Bennewitz, C. Stachniss and W. Burgard, *OctoMap: A Probabilistic, Flexible, and Compact 3D Map Representation for Robotic Systems*, In Proc. of ICRA, Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation, 2010

[11] K. M. Wurm, D. Hennes, D. Holz, R. B. Rusu, C. Stachniss, K. Konolige and W. Burgard, *Hierarchies of Octrees for Efficient 3D Mapping*, In Proc. of IROS, pages 4249–4255, 2011