

Urban Scene Segmentation with Laser-Constrained CRFs

Charika De Alvis

Lionel Ott

Fabio Ramos

Abstract—Robots typically possess sensors of different modalities, such as colour cameras, inertial measurement units, and 3D laser scanners. Often, solving a particular problem becomes easier when more than one modality is used. However, while there are undeniable benefits to combine sensors of different modalities the process tends to be complicated. Segmenting scenes observed by the robot into a discrete set of classes is a central requirement for autonomy as understanding the scene is the first step to reason about future situations. Scene segmentation is commonly performed using either image data or 3D point cloud data. In computer vision many successful methods for scene segmentation are based on conditional random fields (CRF) where the maximum a posteriori (MAP) solution to the segmentation can be obtained by inference. In this paper we devise a new CRF inference method for scene segmentation that incorporates global constraints, enforcing the sets of nodes are assigned the same class label. To do this efficiently, the CRF is formulated as a relaxed quadratic program whose MAP solution is found using a gradient-based optimisation approach. The proposed method is evaluated on images and 3D point cloud data gathered in urban environments where image data provides the appearance features needed by the CRF, while the 3D point cloud data provides global spatial constraints over sets of nodes. Comparisons with belief propagation, conventional quadratic programming relaxation, and higher order potential CRF show the benefits of the proposed method.

I. INTRODUCTION

Scene segmentation is a core competency for many robotic tasks. It provides the foundation which allows a robot to understand and reason about its environment. For navigation in urban environments such information is critical for safety, as it allows the robot to predict which areas pose a risk due to the presence of dynamic objects. Typically robots carry many different sensors, such as cameras, laser scanners, RGB-D cameras, etc, which typically observe the environment from slightly different angles. This variation in view point and modality makes the optimal combination of sensors very challenging. In this paper we propose a model which effectively combines multiple modalities. The method is applied to image segmentation using camera and laser scan data but is general in nature and applicable to a wide variety of sensor combinations.

Our method is based on a relaxed quadratic program formulation of CRFs for scene segmentation which enforces a set of global constraints. Image data is used to build the CRF graph and potential functions while the depth data is used to formulate global constraints over sets of nodes in the CRF. These constraint sets contain all nodes belonging to the same object, as determined by the depth data and ensure they take the same label during the inference process. The

method finds the MAP solution using an efficient gradient based algorithm, based on [27]. The main contributions of the paper are:

- Novel CRF formulation using global constraints capable of enforcing label consistency;
- Experimental evaluation of the proposed method for scene segmentation using image and 3D laser data gathered by a robotic platform.

The remainder of the paper is structured as follows. In Section II we give an overview of work related to ours, before we introduce our method in Section III. In Section IV we provide experimental evaluation of our method before concluding in Section V.

II. RELATED WORK

In computer vision many successful image segmentation methods are based on graph cuts [5] and refinements such as normalised cuts [22, 4]. Graph cuts represent the image as a graph and attempt to find the set of edges with minimal cost, that when cut results in a segmentation of the image. There are other approaches that work on a similar representation but use a different way of solving the problem. Felzenszwalb and Huttenlocher [7] propose a method that uses greedy local segmentation decisions to obtain accurate global results. A novel graphical model, associative hierarchical random fields, with applications to scene segmentation has proposed in [16]. Stereo vision based scene segmentation is another common method. For example He and Upcroft [9] present a method to build a dense 3D semantic occupancy map of an environment based on semantic labels obtained using a Markov random field which are used to update the semantic labels of the map cells. A similar approach is taken in [21] using a CRF for the segmentation task and creating a triangulated mesh of the environment rather than a voxel grid. All of these approaches use only image data without any additional outside information.

In robotics there has been a lot of work on scene segmentation using multiple modalities, such as camera and 3D laser data. Douillard et al. [6] propose a spatial-temporal CRF method integrating measurements from a conventional 2D laser scanner with images from a calibrated camera. Munoz et al. [18] extract features from image and laser data and use these in a classifier to segment the scene. A method that accumulates image based segmentation results in a 3D point cloud was presented by Hermans et al. [10]. An extension of [7] to RGBD data is presented in [23], taking advantage of distance and normal information. In [3] a link-chain clustering method operating on a super voxel representation of RGB-D data was presented. Xu et al. [25]

present a method using multiple independent classifiers with a sophisticated fusion framework. A method that exploits both colour and depth information with the help of a CRF is presented in [26]. This method makes predictions separately on the depth and colour data and fuses the results using a CRF.

Higher order potentials (HOP) allow encoding additional information which the unary and pairwise potentials of a CRF cannot represent. This enables modelling longer range dependencies within the model. These HOP act as soft constraints during the optimisation. Kohli et al. [13] use a P^n Potts model-based CRF with HOP for the task of image segmentation and use a graph cut based algorithm to solve the optimisation problem. Tarlow et al. [24] proposed a method with HOP models and belief propagation, adopting a set of potentials for which efficient message passing rules exist. In [14] a dual decomposition based master-slave framework is presented to solve generic higher order Markov random fields. While HOP can be created from the same information as the constraints, they only form soft constraints and as such can be violated in the final solution. Our approach, in contrast, ensures that the imposed constraints from depth information are satisfied by the solution. Further we exploit much simpler features compared to the state of the art methods while providing higher accuracy for even tricky classes such as pedestrians. Our method also has the potential to be implemented in real time.

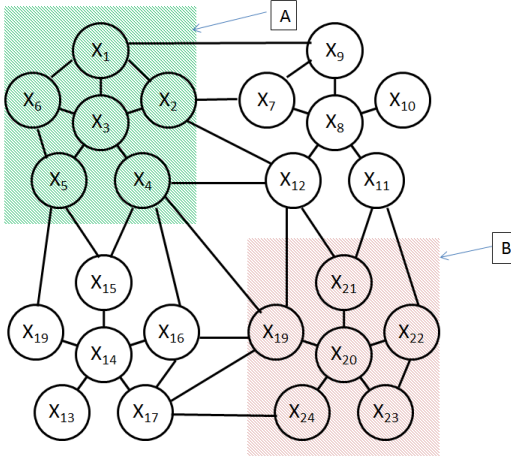


Fig. 1: Example of the type of CRF graph used in this paper. Pairwise potentials are indicated by the edges, while the additional constraints are indicated by the two shaded areas, A and B. These areas encode sets of nodes which are required to be assigned the same label.

III. GLOBALLY CONSTRAINED CRF

Our segmentation method is based on conditional random fields (CRF) with unary and pairwise potentials. The additional *a priori* information about sets of points which belong to the same group is encoded as constraints on the CRF. A graphical representation of this structure is shown in Figure 1, where nodes are denoted by circles while edges indicate connections between nodes. The two sets of nodes coloured identically represent sets of nodes constrained to

take the same label. The unary and pairwise potentials are based on information extracted from the image while the information about groups of nodes is extracted from 3D laser data. Our goal is to find the best label assignment for each node, i.e. the MAP solution of the CRF. To do this efficiently we represent the CRF as a quadratic program.

A. Conditional Random Field

The log likelihood model of a conditional random field is given by:

$$\log P(X | S) = \sum_{i \in S} \phi_i(X_i) + \sum_{i \in S, j \in \mathcal{N}(i)} \psi_{ij}(X_i, X_j) - Z(S), \quad (1)$$

where $Z(S)$ is the normaliser, $X = \{X_1, X_2, \dots, X_N\}$ is the set of discrete random variables associated with the super pixels [1] set S in the input image. Each super pixel X_i is assigned one of the output labels $L = \{1, \dots, K\}$. The potential functions of the CRF are denoted by $\phi_i(X_i)$ for the unary potential and $\psi_{ij}(X_i, X_j)$ for the pairwise potential defined for each super pixel i and each of its neighbours $\mathcal{N}(i)$.

B. Quadratic Program Formulation

The goal is to find the best assignment of labels to the nodes (MAP assignment) considering local and global information. As finding the MAP solution to Eq. (1) is NP hard we start by representing it as a quadratic integer program of the following form:

$$\begin{aligned} \text{maximise} \quad & \sum_{i \in S} \sum_{p \in L} \phi_i(x_i^p) \mu_i(x_i^p) \\ & + \sum_{i \in S} \sum_{j \in \mathcal{N}(i)} \sum_{p, q \in L} \psi_{ij}(x_i^p, x_j^q) \mu_i(x_i^p) \mu_j(x_j^q) \end{aligned} \quad (2a)$$

$$\text{subject to} \quad \sum_{p \in L} \mu_i(x_i^p) = 1 \quad \forall i \quad (2b)$$

$$\mu_i(x_i^p) \in \{0, 1\} \quad \forall i, p, \quad (2c)$$

with the indicator function:

$$\mu_i(x_i^p) = \begin{cases} 1 & \text{if } (X_i = p) \wedge (x_i^p = 1) \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where x_i^p encodes if node X_i has been assigned label p . This quadratic program formulation penalises disagreements between the data via the indicator function, which guides the model to obtain coherent segmentations. Additionally, Equations (2b) and (2c) enforce that exactly one label is selected for each node. Relaxing the integer requirement of

the quadratic program [27] we obtain:

$$\begin{aligned} \text{maximise} \quad & \sum_{i \in S} \sum_{p \in L} \phi_i(x_i^p) \mu_i(x_i^p) \\ & + \sum_{\substack{i \in S \\ j \in N(i)}} \sum_{p, q \in L} \psi_{ij}(x_i^p, x_j^q) \mu_i(x_i^p) \mu_j(x_j^q) \end{aligned} \quad (4a)$$

$$\text{subject to} \quad \sum_{p \in L} \mu_i(x_i^p) = 1 \quad \forall i \quad (4b)$$

$$0 \leq \mu_i(x_i^p) \leq 1 \quad \forall i, p. \quad (4c)$$

Optimising Eq. (4) yields an approximation to the MAP solution for the segmentation problem. However, it does not yet include the global constraints on sets of nodes. Adding these constraints we obtain:

$$\begin{aligned} \text{maximise} \quad & \sum_{i \in S} \sum_{p \in L} \phi_i(x_i^p) \mu_i(x_i^p) \\ & + \sum_{\substack{i \in S \\ j \in N(i)}} \sum_{p, q \in L} \psi_{ij}(x_i^p, x_j^q) \mu_i(x_i^p) \mu_j(x_j^q) \end{aligned} \quad (5a)$$

$$\text{subject to} \quad \sum_{p \in L} \mu_i(x_i^p) = 1 \quad \forall i \quad (5b)$$

$$\sum_{i, j \in C_k} \sum_{p \in L} \mu_i(x_i^p) - \mu_j(x_j^p) = 0 \quad \forall C_k \in \mathcal{C} \quad (5c)$$

$$0 \leq \mu_i(x_i^p) \leq 1 \quad \forall i, p \quad (5d)$$

where Eq. (5c) enforces that all pairs of points i and j in a constraint set $C_k \in \mathcal{C}$ are assigned the same label.

In order to solve Eq. (5) efficiently we follow [15] and rewrite it in matrix notation:

$$\text{maximise} \quad \frac{1}{2} A^T Q A + b^T A \quad (6a)$$

$$\text{subject to} \quad E A = 0 \quad (6b)$$

$$0 \leq A \leq 1, \quad (6c)$$

where Q encodes the quadratic coefficients (pairwise potentials) and b the linear coefficients (unary potentials). A is the indicator matrix representing the $\mu_i(x_i^p)$ variables and E encodes the global constraints from Eq. (5c). The solution to Eq. (6) can be found by introducing Lagrange multipliers as follows:

$$\text{maximise} \quad \frac{1}{2} A^T Q A + b^T A + \lambda E A \quad (7a)$$

$$\text{subject to} \quad 0 \leq A \leq 1, \quad (7b)$$

We can achieve the same maximum as in Eq. (6) by making $\lambda E A$ equal to zero. To this end we introduce new variables:

$$E Z = 0 \quad (8)$$

$$Z R = A, \quad (9)$$

where R has the dimension $\dim(A) - \dim(E)$, while solving Eq. (8) implies that Z is the null space of E . Substituting

these two equations back into Eq. (7) we obtain:

$$\text{maximise} \quad \frac{1}{2} R^T (Z^T Q Z) R + (Z^T c)^T R \quad (10a)$$

$$\text{subject to} \quad 0 \leq R \leq 1 \quad (10b)$$

This transformation has two benefits: First, the dimensionality of R is reduced compared to that of A based on the number of constraints. This means that a large number of constraints makes the optimisation problem easier to solve. Second, the optimisation problem is now unconstrained which again makes it easier to solve.

Similar to the transformation from Eq. (5) to Eq. (6) we can rewrite Eq. (10) using element wise notation as follows:

$$\begin{aligned} \text{maximise} \quad & \sum_i \sum_p \rho_i(y_i^p) \mu_i(y_i^p) \\ & + \sum_{i, j} \sum_{p, q} \tau_{ij}(y_i^p, y_j^q) \mu_i(y_i^p) \mu_j(y_j^q) \end{aligned} \quad (11a)$$

$$\text{subject to} \quad 0 \leq \mu_i(y_i^p) \leq 1, \quad (11b)$$

with the unary potential $\rho_i = -Z^T c$ and the pairwise potential $\tau_{ij} = -Z^T Q Z$ and y_i^p denotes if label p has been assigned to node Y_i . We optimise Eq. (11) using gradient ascent which can be done efficiently as the gradient can be computed in closed form [27]:

$$\begin{aligned} q_i(y_i^p) &= \frac{\partial B}{\partial \mu_i(y_i^p)} \\ &= \rho_i(y_i^p) + 2 \sum_{i, j} \sum_q \tau_{ij}(y_i^p, y_j^q) \mu_j(y_j^q) \end{aligned} \quad (12)$$

$$\mu_i^{t+1}(y_i^p) = \frac{\mu_i^t(y_i^p) q_i(y_i^p)}{\sum_q \mu_i^t(y_i^q) q_i(y_i^q)}, \quad (13)$$

with B standing for Eq. (11a).

This allows us to implement a highly efficient gradient ascent based algorithm as the gradient can be evaluated directly in closed form. Once the algorithm has converged we can extract the values of original indicator variables $\mu(x_i^p)$ and thus the MAP label assignments to the X_i variables. To this end we transform the solution for $\mu(y_i^p)$ obtained from Eq. (11) back into the form of Eq. (5) using $A = Z R$. A is a column vector whose entries correspond to the values of the $\mu(x_i^p)$. The optimal assignment to each node X_i is found by selecting the label $p \in L$ for which $\mu(x_i^p) = 1$ holds. This is summarised in Algorithm 1. The required inputs are the values of the potentials over the possible R value settings. Then the gradient (Eq. (12)) is computed and used to update the solution iteratively until convergence is achieved. Finally, the solution is extracted and returned.

IV. EXPERIMENTS

In this section we present experimental evaluation of our proposed framework on the task of image-based scene segmentation. We use the KITTI dataset [8] as it provides typical urban data. The dataset was captured by driving around the

Algorithm 1: Globally Constrained CRF

Input: Potential values ρ and τ
Output: Assignment of X
 // Perform gradient descent

```

1 repeat
2   foreach  $i \in \{1, \dots, N\}$  do
3     foreach  $p \in \{1, \dots, L\}$  do
4        $q_i(y_i^p) \leftarrow$ 
5          $\rho_i(y_i^p) + 2 \sum_{i,j} \sum_q \tau_j(y_i^p, y_j^q) \mu_j^t(y_j^q)$ 
6        $\mu_i^{t+1}(y_i) \leftarrow \frac{\mu_i^t(y_i) q_i(y_i)}{\sum_i \mu_i^t(y_i) q_i(y_i)}$ 
7     end
8   end
9   until convergence;
  // Extract final solution
10   $A \leftarrow ZR$ 
11  foreach  $i \in \{1, \dots, N\}$  do
12    foreach  $p \in \{1, \dots, L\}$  do
13       $X_i \leftarrow p$  if  $\mu(x_i^p) = 1$ 
14    end
15  end
16 return  $X$ 

```

Type	Description	Dimensionality
Texture	RGB gradient magnitude histogram	$50 \times 3 = 150$
	RGB gradient orientation histogram	$50 \times 3 = 150$
Colour	RGB mean	3
	RGB std	3
	HSV histogram	$50 \times 3 = 150$
Location	Super pixel image coordinates	200

TABLE I: Features used for the unary potential of the CRF based on a discriminant analysis classifier applied to super pixels.

city of Karlsruhe. Importantly, the data contains both colour images and Velodyne depth data. The image information is used to build the CRF model structure and potential functions while the Velodyne data is used to construct global constraint sets.

A. Model Building

We start by extracting super pixels from the image using SLIC [2] which forms an over segmentation of the original image. From each 375×1242 image we extract roughly 1600 super pixels, shown in Figure 2b. Each of these represents a node in our CRF and the goal is to label them with one of the seven different classes: vehicle, pedestrian & cyclist, buildings, road & paved area, sky, vegetation, and unknown. Due to the low sample size of pedestrians and cyclists in the dataset they are assigned to same class. The edges between nodes are defined by their distance within the image, i.e.:

$$E(i, j) = \begin{cases} 1 & \text{if } \text{dist}(i, j) < \Theta \\ 0 & \text{otherwise} \end{cases}, \quad (14)$$

where Θ is the distance threshold and $\text{dist}(i, j)$ is the Euclidean distance in image coordinates between centres of two super pixels. All super pixels closer than the user defined distance Θ are connected. In our experiments Θ was set such that each node is connected to roughly ten neighbouring nodes, which results in a roughly grid like structure.

The unary potentials ϕ_i are obtained from the posterior of a pseudo linear discriminant analysis classifier [17]. The classifier is trained on 120 manually labelled images from the KITTI dataset using colour, texture, and location features, shown in Table I. The pairwise potentials ψ_{ij} are derived based on their dissimilarity using colour, texture, and location information of the super pixels, i.e.:

$$\begin{aligned} \text{dis}(i, j) = & D(\text{colour_hist}(i), \text{colour_hist}(j)) \\ & + \theta_c \|\text{mean_colour}(i) - \text{mean_colour}(j)\|_2 \\ & + \theta_l \|\text{com}(i) - \text{com}(j)\|_2 / 3, \end{aligned} \quad (15)$$

where $\text{mean_colour}(i)$ is the mean colour of the i -th super pixel normalised to 1 by θ_c , $\text{com}(i)$ is the centre of mass of the super pixel in pixel coordinates, normalised to 1 with θ_l , and $\text{colour_hist}(i)$ is the colour histogram of the i -th super pixel whose difference is computed using the Bhattacharya distance:

$$D(a, b) = \sqrt{1 - \frac{1}{\sqrt{\sum_i a_i \sum_i b_i N^2}} \sum_i \sqrt{a_i b_i}}, \quad (16)$$

where a and b are two histograms and N is the number of bins in the histograms. This results in a similarity value between 0 and 1, with 0 encoding identical super pixels. As the constraint function requires a value of 1 for identical super pixels we use the following final pairwise potential function:

$$\psi_{i,j}(x_i^p, x_j^q) = \begin{cases} 1 - \text{dis}(i, j)^2 & \text{if } p = q \\ \text{dis}(i, j)^2 & \text{otherwise} \end{cases}. \quad (17)$$

We obtain the global constraints on sets of super pixels by extracting groups of connected points, or objects, from the Velodyne point cloud. This is facilitated by the KITTI dataset providing time synchronised camera images and Velodyne point clouds. To this end we first perform a simple ground plane removal step using RANSAC to find the largest plane aligned with the ground. The remaining points are then grouped using Euclidean distance based clustering [19]. This results in a collection of clusters, of which we only consider those that contain more than 150 points which ensures that each clusters contains only points belonging to a single class. The ground plane as well as the retained segments of this process can be seen in Figure 2c. The 3D coordinates of the points contained in the selected clusters are then translated into image space coordinates using the extrinsic calibration provided by the KITTI dataset and then associated with super pixels. Based on this mapping we create the constraint sets \mathcal{C} used in the optimisation. All super pixels that correspond to the same laser segment are constrained to be assigned the same label. Super pixels which do not belong to any of the extracted laser segments are kept unconstrained.

B. Segmentation Quality

In the following we present image only CRF solutions obtained using loopy belief propagation (LBP) and quadratic programming (QP) to showcase the quality of the results obtained by these methods without using any additional constraints. Thereafter, we introduce the constraints obtained

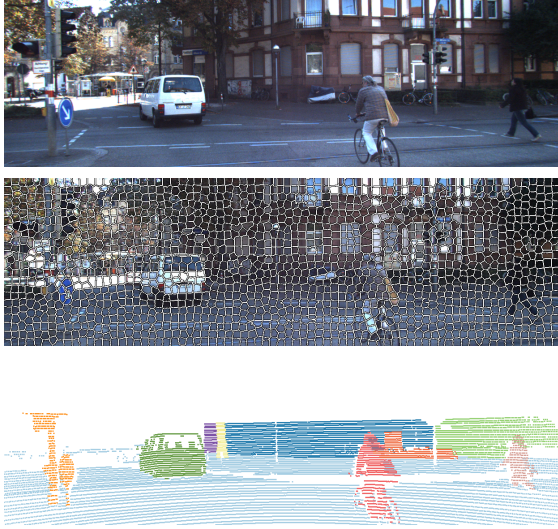


Fig. 2: Display of a typical scene from the KITTI dataset. top image shows the raw image while middle overlays the super pixels extracted from the image. In bottom image the global constraints extracted from the 3D laser point clouds are shown projected into the image space, each colour represents a single segment. Labels of the segments are unknown at this stage.

from the Velodyne and compare the results obtained using a graph-cut based HOP method [12] with our hard constraint based method.

Visual Information Only Segmentation

We present results from three methods, (i) discriminant analysis classifier which provides the unary potentials of the CRF, (ii) loopy belief propagation using the UGM toolbox [20], and (iii) quadratic programming solution [27]. Exemplary results together with the original image and ground truth labels are shown in Figure 3. The first row shows the original colour images while the second row shows the most likely class of the discriminant analysis classifier which is used as the unary potentials of the CRF. As to be expected the classifier output is noisy and incorrect in several places. Both the LBP and QP based CRF solutions produce a much cleaner and consistent result compared with the raw classifier result. However, there are still segmentation errors present due to effects such as shadowing and illumination changes. The quantitative evaluation results from 100 manually labelled images, shown in Table II, further demonstrates the improvements and also indicates that the QP based solution outperforms the LBP one. This demonstrates that the basis on which our method is built is capable of producing high quality segmentation results before any additional constraints are added, which will be evaluated next.

Laser Constrained Segmentation

In this section we explore the impact additional constraints, extracted from Velodyne data, have on segmentation results by comparing our method to a HOP based method by Kohli et al. [12]. The higher order potentials penalise label inconsistencies between nodes identified to be part of a single segment in the 3D data. Both methods use uniform weight parameters for the unary, pairwise, and higher order

potentials, where applicable.

Some exemplary results are shown in Figure 4 with the original image shown on the far left, followed by the result of the HOP based method in the second column, then our method, and finally the hand labelled ground truth. Inspecting the results we can see that the HOP based method struggles to correctly identify distant objects, especially when cars or walls are involved. Additionally, the results our method obtains appear more uniform with less spurious classifications. This difference in behaviour is explained by the way the additional 3D information is used. While our method enforces the constraints the HOP based method is allowed to violate them. The examples in Figure 5 show the benefit of using the hard constraints rather than soft constraints. The first two rows showcase this for a single wall while the third row shows the result of this in a scene populated by pedestrians. The first two columns show the original image and the segment extracted from the Velodyne data. Due to the visual appearance of these areas the classifier fails to pick the correct class in some parts of the 3D segment. The HOP based method fixes some classification errors, however, cannot fix every single one. In the case of the pedestrian scene the HOP method even misclassifies all pedestrians. Our method on the other hand is forced to assign a single class to the entire segment and as such the correct class is assigned even to the areas where the classifier makes mistakes.

For a quantitative analysis we compute average precision, recall, accuracy, and F1-score for the different methods on 100 labelled images. As we can see in Table II the addition of global constraints in our method allows it to significantly outperform the other methods lacking this information and even the HOP method, using the same information, does not provide the same benefits. This shows that adding constraints based on simple information about which areas belong to a single object allows the segmentation to be more accurate. This is good news, as this type of information is readily available in robotic systems. Looking at the performance of the individual classes in Table III we can see that “cyclist & pedestrian” class is the hardest one. This is explained by the fact that instances of this class occur infrequently and as such the classifier has a harder time at classifying them correctly. Furthermore, this class has the smallest appearance in the Velodyne data and as such will only be detected at close range. The other classes exhibit similar performance, which is not surprising, given that they occur frequently in the data and cover larger areas of the scene.

The performance of both constrained QP and HOP can be improved by training the weight parameters of the potential functions, which encodes knowledge about class relationships and object co-occurrence statistics. The advantage of our method is, that it only requires unary and pairwise potentials while HOP has additional higher order potentials, which can be harder and time consuming to learn. This makes the proposed method easier to fine tune as there are fewer parameters involved.

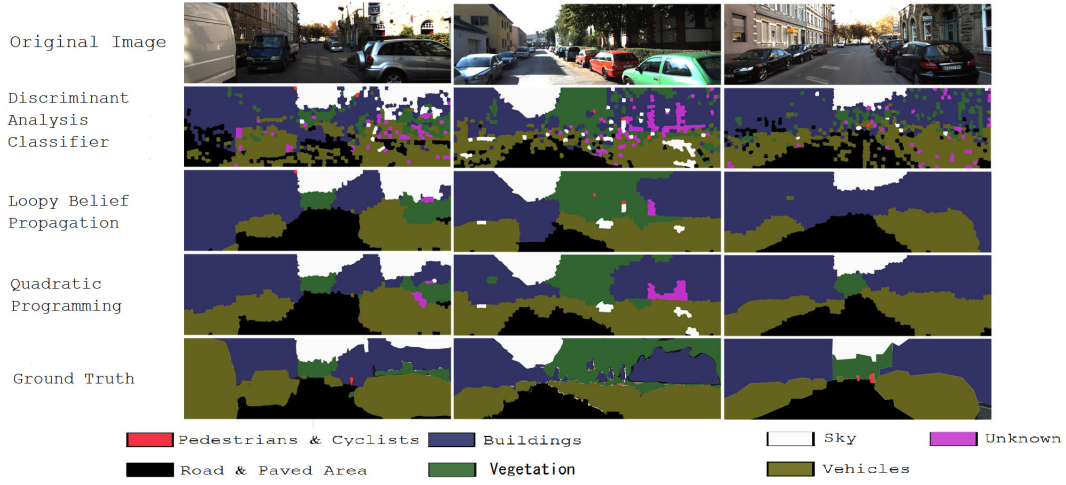


Fig. 3: This figure shows exemplary scene segmentation results obtained on several images. From top to bottom we have: the original image processed by each method, discriminant analysis classifier, loopy belief propagation, quadratic programming, and ground truth.

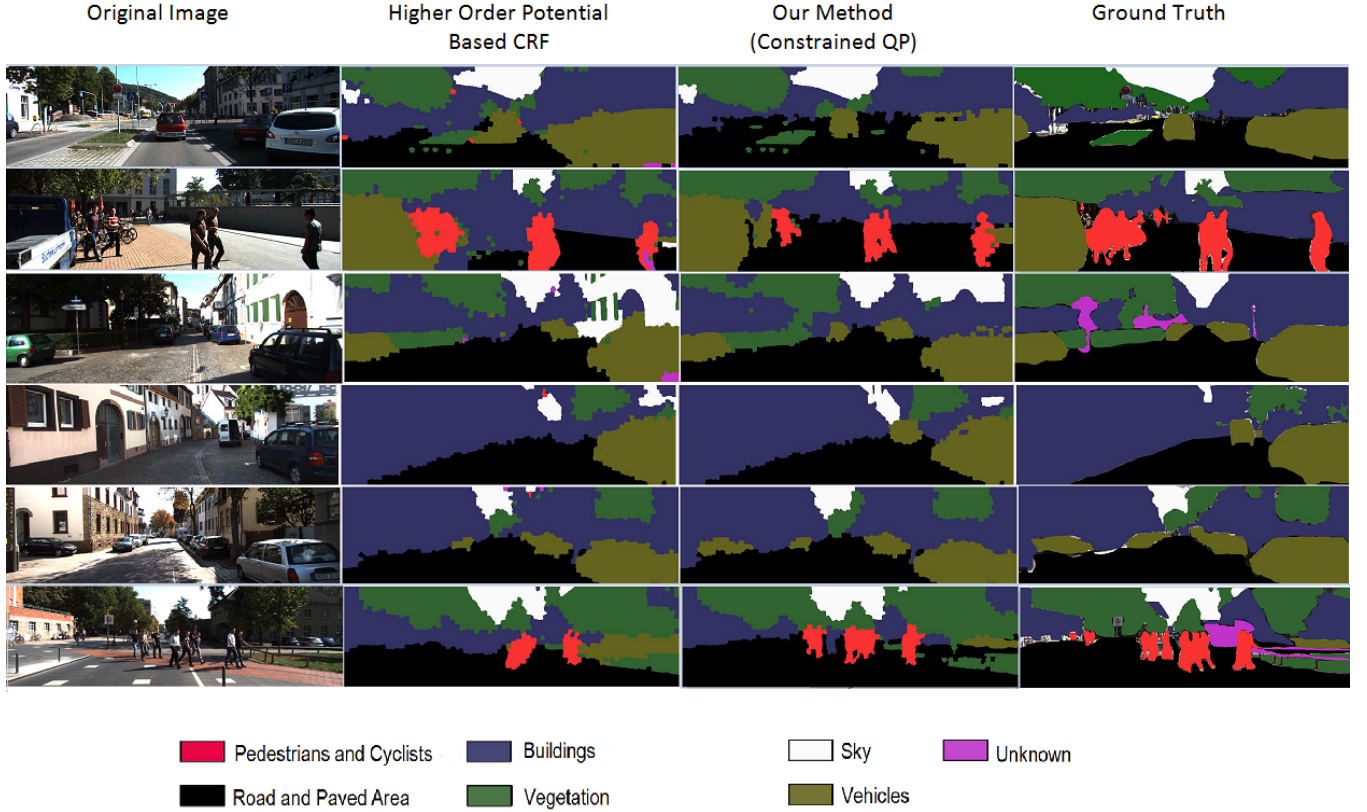


Fig. 4: This figure shows results obtained on various scenes using the HOP based method and our proposed method together with the original image and the hand labelled ground truth. We can see that our method (Constrained QP) performs better than the HOP based method at segmenting distant and objects cast in shadows.

C. Runtime Comparison

We start by comparing the runtime required to solve the constrained quadratic program of Eq. (5) directly using NLOPT BOBYQA [11] compared to our proposed framework. As we can see in Figure 6, directly solving the quadratic program is not feasible for problems of interesting size. On the other hand, our method scales very favourably with the problem size. Additionally, while typically increas-

ing the number of constraints makes the problem harder and thus slower to solve, our method becomes faster with more constraints. This is caused by the fact that constraints reduce the size of the actual problem we solve. This means that adding more domain knowledge allows us to improve the quality of the result as well as speed up the computation.

A typical CRF derived from the images used in the experiments consists of 1600 nodes, each of which can have

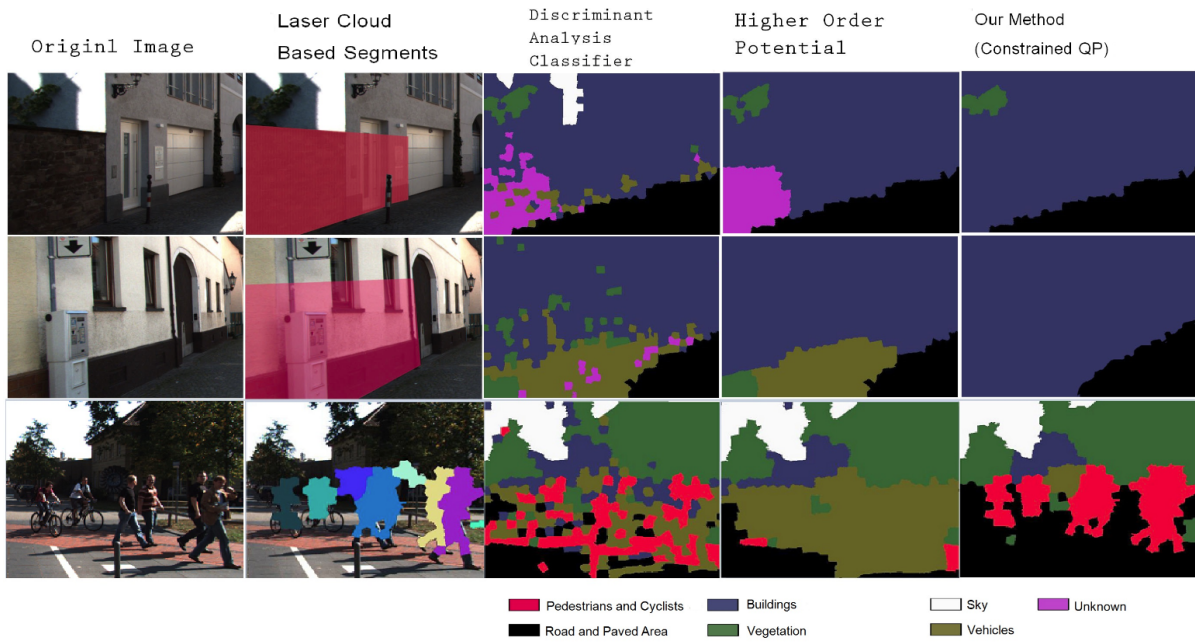


Fig. 5: Examples of the benefits that enforcing hard constraints provide. The highlighted areas in the image show continuous 3D segments extracted from Velodyne data. The classifier output in these areas is noisy and wrong due to visual ambiguities. While the HOP based method fails to correct this our method succeeds in classifying the entire area correctly, as it is forced to assign a single class to each of the laser based segments.

Method	Average Precision	Average Recall	Average Accuracy	F1 Score
Discriminant Analysis Classifier	0.7027 ± 0.045	0.5127 ± 0.061	0.8826 ± 0.045	0.5927 ± 0.057
Loopy Belief Propagation	0.7435 ± 0.051	0.7197 ± 0.081	0.9024 ± 0.053	0.7314 ± 0.067
Quadratic Programming Relaxation	0.8001 ± 0.032	0.7645 ± 0.048	0.9150 ± 0.053	0.7818 ± 0.040
Higher Order Potentials [12]	0.8319 ± 0.073	0.8143 ± 0.067	0.9278 ± 0.022	0.8230 ± 0.070
Constrained Quadratic Programming	0.8549 ± 0.079	0.8424 ± 0.078	0.9507 ± 0.025	0.8482 ± 0.076

TABLE II: Quantitative evaluation of various segmentation methods. The first three rows represent standard CRFs using only image based information. The last two rows show the results for methods using additional information obtained from 3D Velodyne scans. The HOP method incorporates this information as an additional potential, while our method (Constrained Quadratic Programming) enforces the validity of this additional information as constraints. We can see that the addition of the 3D information improves the performance compared to the image only based solutions. However, actively enforcing the constraints allows our method to outperform the HOP based method.

Quality Measure	Average Precision		Average Recall		Average Accuracy		F1 Score	
Method	HOP	CQP	HOP	CQP	HOP	CQP	HOP	CQP
Cyclists & Pedestrians	0.7689	0.7700	0.5134	0.5334	0.9670	0.9772	0.6153	0.6302
Roads & Paved Area	0.8431	0.8554	0.9747	0.9775	0.9569	0.9789	0.9032	0.9124
Vegetation	0.8284	0.8440	0.5161	0.5359	0.9468	0.9473	0.5931	0.6192
Buildings	0.8431	0.8420	0.8448	0.8838	0.8652	0.9103	0.8382	0.8568
Sky	0.7519	0.7877	0.7690	0.7564	0.9723	0.9780	0.7265	0.7461
Vehicles	0.8485	0.9089	0.7101	0.8058	0.9031	0.9413	0.7614	0.8543

TABLE III: Quantitative evaluation of the performance on a per class for HOP method and our method CQP. All the major 6 classes excluding the unknown class are separately evaluated for the quality of segmentation.

one of seven different labels, which means we have on the order of 11 200 random variables. Solving this CRF using the quadratic program formulation Eq. (4) (with no laser based constraints) takes around 2s while the belief propagation based solution takes 0.5s. Including the constraints we can reduce the number of nodes to around 400 which results in a much smaller number of variables, around 2800. Solving this problem using gradient based method takes around 0.07s. All computations were performed on an Intel Core i5 3.20 GHz processor with C++ implementations of the algorithms.

Besides the reduction of the number of variables involved our method also requires fewer iterations to converge, around 25, compared to 70 for the purely image based quadratic program. These two advantages, reduction in number of variables and faster convergence gives our method a significant computational advantage.

V. CONCLUSION

In this paper we presented a novel image segmentation method based on a conditional random field with additional

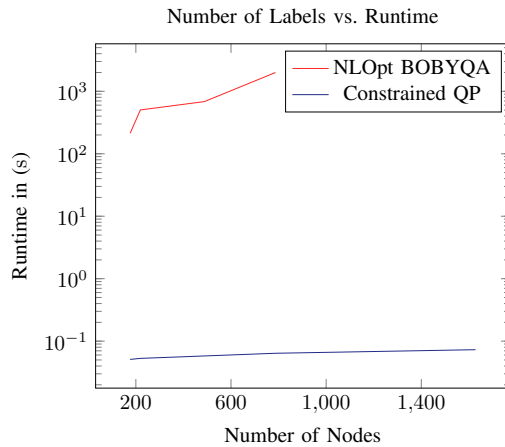


Fig. 6: The plot shows the time (log scale) needed to find a solution as a function of the number of nodes in the CRF. NLOpt BOBYQA solving the problem directly scales very poorly while our proposed method is scaling much more favourably.

global constraints which encode *a priori* information about groups of nodes having the same label obtained from a secondary sensor. This CRF is formulated as a relaxed quadratic program whose MAP solution is found using gradient descent based optimisation. We evaluate our method on data from the KITTI project. Each image is pre-processed into super pixels which provide the unary and pairwise potentials of the CRF. The global constraints on sets of super pixels are obtained from Velodyne data. The results show that the addition of these hard constraints significantly improves on the solution obtained without constraints. Runtime comparisons show how black box solvers do not scale for this problem and how our formulation exploits constraints in a way which simplifies the problem. Finally, the proposed method is general and capable of encoding other forms of constraints, such as relative positioning of classes with respect to each other.

REFERENCES

- [1] Superpixel: Empirical studies and applications. <http://ttic.uchicago.edu/~xren/research/superpixel/>.
- [2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Sabine. SLIC Superpixels. Technical report, EPFL, 2010.
- [3] A. Aijazi, P. Checchin, and L. Trassoudaine. Segmentation Based Classification of 3D Urban Point Clouds: A Super-Voxel Based Approach with Evaluation. *Remote Sensing*, 2013.
- [4] Y. Boykov and G. Funka-Lea. Graph Cuts and Efficient ND Image Segmentation. *International Journal of Computer Vision*, 2006.
- [5] Y. Boykov and M. Jolly. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in ND Images. In *IEEE International Conference on Computer Vision*, 2001.
- [6] B. Douillard, D. Fox, and F. Ramos. A Spatio-Temporal Probabilistic Model for Multi-Sensor Object Recognition. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2007.
- [7] P. Felzenszwalb and D. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 2004.
- [8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets Robotics : The KITTI Dataset. *The International Journal of Robotics Research*, 2011.
- [9] H. He and B. Upcroft. Nonparametric semantic segmentation for 3d street scenes. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2013.
- [10] A. Hermans, G. Floros, and B. Leibe. Dense 3D Semantic Mapping of Indoor Scenes from RGB-D Images. In *Proc. of the IEEE Int. Conf. on Robotics & Automation*, 2014.
- [11] S. Johnson. The NLOpt nonlinear-optimization package. <http://ab-initio.mit.edu/nlopt>.
- [12] P. Kohli, L. Ladicky, and P. Torr. Graph cuts for minimizing robust higher order potentials. In *International Conference on Computer Vision and Pattern Recognition*, 2008.
- [13] P. Kohli, L. Ladicky, and P. Torr. Robust Higher Order Potentials for Enforcing Label Consistency. *International Journal of Computer Vision*, 2009.
- [14] N. Komodakis and N. Paragios. Beyond Pairwise Energies: Efficient Optimization for Higher-Order MRFs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
- [15] E. Krogstad. *Optimeringsteori Quadratic Programming Basics*. PhD thesis, NTNU, 2012.
- [16] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative Hierarchical Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- [17] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher Discriminant Analysis with Kernels. In *Proc. of the IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, 1999.
- [18] D. Munoz, J. Bagnell, and M. Hebert. Co-Inference for Multi-Modal Scene Analysis. In *European Conference on Computer Vision*, 2012.
- [19] Bogdan R. and S. Cousins. 3D is here: Point Cloud Library (PCL), May 9-13 2011.
- [20] M. Schmidt. UGM: A Matlab toolbox for probabilistic undirected graphical models, 2007. URL <http://www.cs.ubc.ca/~schmidtm/Software/UGM.html>.
- [21] S. Sengupta, E. Greveson, A. Shahrokni, and P. Torr. Urban 3d semantic modelling using stereo vision. In *Proc. of the IEEE Int. Conf. on Robotics & Automation*, 2013.
- [22] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.
- [23] J. Strom, A. Richardson, and E. Olson. Graph-Based Segmentation for Colored 3D Laser Point Clouds. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2010.
- [24] D. Tarlow, I. Givoni, and R. Zemel. HOP-MAP: Efficient Message Passing with High Order Potentials. In *International Conference on Artificial Intelligence and Statistics*, 2010.
- [25] P. Xu, F. Davoine, J. Bordes, Z. Huijing, and Thierry Denœux. Multimodal Information Fusion for Urban Scene Understanding. *Machine Vision and Applications*, 2014.
- [26] R. Zhang, S. Candra, K. Vetter, and A. Zakhor. Sensor Fusion for Semantic Segmentation of Urban Scenes. In *Proc. of the IEEE Int. Conf. on Robotics & Automation*, 2015.
- [27] Y. Zhang and T. Chen. Efficient inference for fully-connected CRFs with stationarity. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.