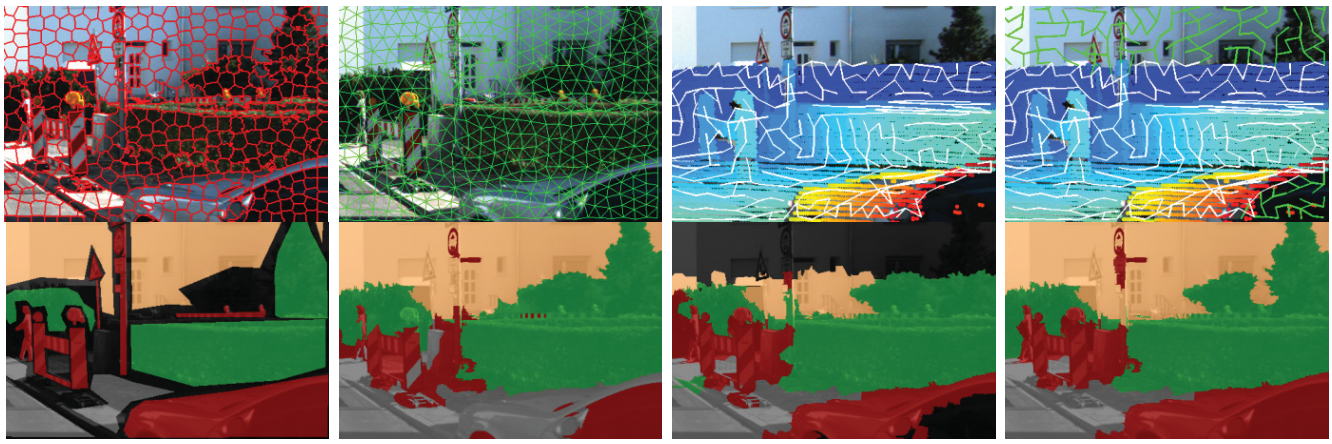
Semantic Segmentation with Heterogeneous Sensor Coverages

Cesar Cadena and Jana Košecká



(a) (b) (c) (d)

Fig. 1: Our novel semantic parsing approach can seamlessly integrate evidence from multiple sensors with overlapping but possibly different fields of view and account for missing data, while predicting semantic labels over the spatial union of sensors coverages. The semantic segmentation is formulated on a graph, in a manner which depends on sensing modality. First row: (a) over-segmentation on the image; (b) graph induced by superpixels; (c) 3D point cloud re-projected on the image with a tree graph structure computed in 3D, and (d) the full graph as proposed here for full scene understanding. In the second row is the semantic segmentation (a) ground truth and results of (b) using the image graph and only visual information; (c) using the 3D graph and visual and 3D information, and finally (d) the result from using a graph for full coverage and all the information. Note the best semantic segmentation achieved over the union of the spatial coverage of both sensors. Color code: ■*ground*, ■*objects*, ■*building* and ■*vegetation*.

***Abstract*— We propose a new approach to semantic parsing, which can seamlessly integrate evidence from multiple sensors with overlapping but possibly different fields of view (FOV), account for missing data and predict semantic labels over the spatial union of sensors coverages. The existing approaches typically carry out semantic segmentation using only one modality, incorrectly interpolate measurements of other modalities or at best assign semantic labels only to the spatial intersection of coverages of different sensors. In this work we remedy these problems by proposing an effective and efficient strategy for inducing the graph structure of Conditional Random Field used for inference and a novel method for computing the sensor domain dependent potentials. We focus on RGB cameras and 3D data from lasers or depth sensors. The proposed approach achieves superior performance, compared to state of the art and obtains labels for the union of spatial coverages of both sensors, while effectively using appearance or 3D cues when they are available. The efficiency of the approach is amenable to real-time implementation. We quantitatively validate our proposal in two publicly available datasets from indoors and outdoors real environments. The obtained semantic understanding of the acquired sensory information can enable higher level tasks for autonomous mobile robots and facilitate semantic mapping of the environments.**

I. Introduction

In recent years numerous advances have been made in semantic mapping of environments. Associating semantic concepts with robot's surroundings can enhance robot's autonomy and robustness, facilitate more complex tasks and enable better human robot interaction. One important component of semantic understanding is so called semantic segmentation, which entails simultaneous classification and segmentation of the sensory data.

In the computer vision community large variety of approaches have been proposed using only appearance features computed from RGB images or sparse 3D geometric features computed by structure from motion techniques. The most successful approaches for semantic segmentation typically use Markov or Conditional Random Fields (MRFs or CRFs) framework [6,12,16,23,26] and vary in the types of features used, the local classifiers and the graph structure defining the random field [16,23,26]. With the exception of few [6,12], most commonly the graph structure is induced by super-pixels obtained by over-segmentation algorithms on individual pixels.

In robotics community an active research topic is semantic segmentation of 3D point clouds from laser range finders [9,25]. The method of Triebel et al. [25] implements an online clustering for the point cloud and carries out the inference in a CRFs setting. While Hu et al. [9] carry out the segmentation through a cascade of classifiers from fine to coarse point cloud clusters. Nowadays many robotic platforms use simultaneously both cameras and 3D lasers for outdoor navigation, and RGB-D cameras for indoors, which are calibrated with respect to each other. The use of these modalities creates opportunities for using richer set of features as well as obtaining better coverage. In the common approaches for semantic segmentation proposed [3,20], the features and the graphical model are constrained to the

César Cadena and Jana Košecká are with the Computer Science Department, Volgenau School of Engineering at George Mason University, Fairfax, VA 20030, US. cesarcadena.lerma@gmail.com, kosecka@.cs.gmu.edu. This research has been funded by the US Army Research Office Grant W911NF-1110476.

area that is covered simultaneously by both sensors. Areas covered by only one of the sensors are not considered. For example in the case of cameras and lasers commonly used in outdoors setting, due the small FOV of laser range sensors the upper portion of the image is typically discarded from evaluation, see Fig. 1(c).

In the case of RGB-D sensors for indoors environments, while the depth data cover almost the same area as the RGB sensor, many image regions have missing corresponding 3D data due to specular or transparent surfaces or large depth values which cannot be estimated reliably with the sensor, see Fig. 7 (second column). The missing depth values are typically filled by using different *in-painting* strategies to obtain dense depth images [10,22], resulting in a full overlap between the RGB and depth channels. The in-painted depth is used by [4,5,8,22] for RGB-D in their semantic parsing approaches. Both Silberman et al. [22] and Gupta et al. [8] also exploit computationally demanding boundary detectors and expensive engineered features. The work of Cadena and Košecká [4] does not require dense depth but it is able to obtain labels only in regions with common sensor coverage. While the in-painted depth images provide a partial solution to a missing data problem, the interpolated depth is often incorrect and available algorithms are computationally expensive (15s/frame), making them unsuitable for real-time operation.

Contribution: In this work we tackle these issues by proposing an effective and efficient strategy for inducing the graph structure of Conditional Random Field used for inference and a novel method for computing the sensor domain dependent potentials. We achieve superior semantic segmentation for the regions in the union of spatial coverage of the sensors, while keeping the computational cost of the approach low. The problem is illustrated in Fig. 1. For example with an image sensor note how in column (b) one portion of the car is confused with the ground because their colors are similar. When we combine the visual sensing with the evidence from a 3D laser sensor, we are able to mitigate sensor specific perceptual confusers, column (c), but now we are only able to explain a subset of the scene, the spatial intersection coverage, leaving us without output for the car glass and the building in the top portion of the image. With the strategy proposed in the remaining of the paper, we take advantage of both sensor modalities without discarding the non-overlapping zones, column (d) in Fig. 1.

Related work: This problem was also addressed by Muñoz et al. [18], as a co-inference problem, where a cascade of classifiers are used over a hierarchy of over-segmentations in each modality, image and 3D. The propagation of evidence between modalities is achieved by using the partial classification results from a modality, at each level of the hierarchy, as input for the classification in the other modality in the overlapping regions. They validate the approach on their own urban dataset containing images and off-line processed 3D point clouds. Later, re-projecting the 3D on the camera poses to emulate an on-line acquisition. Some issues like, e.g., no handle of occluded structures from each point of view, make the emulation a non-realistic one. In terms of timing, the computational cost increases linearly with the number of sensors as they are solving a classification problem for each one. After an expensive (image and 3D) feature extraction, the co-inference takes 0.46s to classify one scene. Beyond the framework and features, our approach differs in the use of an augmented feature space whenever possible and our computational complexity depends on the size of the scene rather than on the number of sensors. In our case we propose to augment CRFs framework by proposing a unified graphical model and using simple, and efficient to compute, features in each modality to define proper energy potentials.

In the next section, we recall the general formulation of Conditional Random Fields, then in Section III we describe the application of CRFs to the semantic segmentation problem for images, intersection of image and 3D, and at last our proposal for the union of image and 3D. In Section IV we describe the experiments in outdoors and indoors datasets. We show the advantages of our approach using all the information available, and compare the results with state of the art methods. Finally, in Section V we present discussions and conclusions of the presented work.

II. General Formulation

CRFs directly model the *conditional* distribution over the hidden variables given observations $p(\mathbf{x}|\mathbf{z})$. The nodes in a CRF are denoted $\mathbf{x} = \langle \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n \rangle$, and the observations are denoted $\mathbf{z}$. In our framework the hidden states correspond to the m possible classes: $\mathbf{x}_i = \{ground, objects, \cdots\}$. A CRF factorizes the conditional probability distribution over hidden states into a product of *potentials*, as:

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \prod_{i \in \mathcal{N}} \phi(\mathbf{x}_i, \mathbf{z}) \prod_{i,j \in \mathcal{E}} \psi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z}) \prod_{c \in \mathcal{H}} \eta(\mathbf{x}_c, \mathbf{z}) \tag{1}$$

where $Z(\mathbf{z})$ is the normalizing partition function, $\langle \mathcal{N}, \mathcal{E} \rangle$ are the set of nodes and edges on the graph, and $\mathcal{H}$ represents the set of high order cliques. The computation of this function is exponential in the size of $\mathbf{x}$.

The unary, or data-term, and pairwise potentials are represented by $\phi(\mathbf{x}_i, \mathbf{z})$ and $\psi(\mathbf{x}_i, \mathbf{x}_j, \mathbf{z})$, respectively, as their domains span over one and two random variables or nodes in the graphical model. The domain for higher order potentials $\eta(\mathbf{x}_c, \mathbf{z})$ span over cliques of three or more fully connected nodes. In the remainder we consider only the data and pairwise terms, choice commonly referred to as pairwise CRFs. The potentials are functions that map variable configurations to non-negative numbers capturing the agreement among the involved variables: the larger a potential value, the more likely the configuration.

Potentials are described by log-linear combinations of *feature functions*, $\mathbf{f}$ and $\mathbf{g}$, i.e., the conditional distribution in Eq. 1 can be rewritten as:

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \exp\left(\sum_{i \in \mathcal{N}} \mathbf{w}_u^T \mathbf{f}(\mathbf{x}_i, \mathbf{z}) + \sum_{i,j \in \mathcal{E}} \mathbf{w}_p^T \mathbf{g}(\mathbf{x}_{i,j}, \mathbf{z}) \right) \tag{2}$$

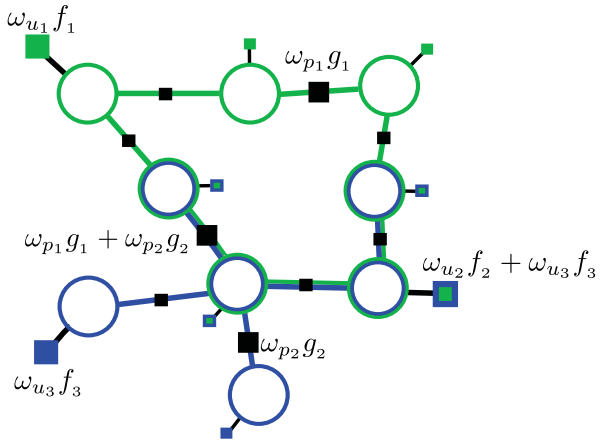


Fig. 2: CRF over a graphical model with two different sources of information.

where u and p stand for unary and pairwise, respectively; and $\mathbf{w}^T = [\mathbf{w}_u^T, \mathbf{w}_p^T]$ is a weight vector, which represents the importance of each feature function. CRFs learn these weights discriminatively by maximizing the conditional likelihood of labeled training data.

Fig. 2 shows an example where the random variables capture two different sources of information A and B, thus we can define different unary and pairwise potentials depending of the type of available information. The CRFs formulation for the graphical model shown in Fig. 2 in terms of Eq. 2 is:

$$p(\mathbf{x}|\mathbf{z}) = \frac{1}{Z(\mathbf{z})} \exp\Bigg(\sum_{i \in \mathcal{N}_{A \setminus B}} \omega_{u_1} f_1(\mathbf{x}_i, \mathbf{z}_A) + \sum_{i \in \mathcal{N}_{A \cap B}} \omega_{u_2} f_2(\mathbf{x}_i, \mathbf{z}_{A,B}) + \sum_{i \in \mathcal{N}_B} \omega_{u_3} f_3(\mathbf{x}_i, \mathbf{z}_B) + \sum_{i,j \in \mathcal{E}_A} \omega_{p_1} g_1(\mathbf{x}_{i,j}, \mathbf{z}_A) + \sum_{i,j \in \mathcal{E}_B} \omega_{p_2} g_2(\mathbf{x}_{i,j}, \mathbf{z}_B) \Bigg)$$

where A and B stand for "green" and "blue" domain. We observe different feature function depending on the available domain: f_1 is computed for nodes with access only to domain A; f_3 is computed for nodes with access to domain B; and f_2 is added for nodes with access to both, A and B. Similarly for edges, two pairwise functions are computed, g_1 if the edge connects two nodes with information from A available and g_2 with information from B.

With this formulation we can obtain either the marginal distribution over the class of each variable $\mathbf{x}_i$ by solving Eq. 2, or the most likely classification of all the hidden variables $\mathbf{x}$. The latter can be formulated as the *maximum a posteriori* (MAP) problem, seeking the assignment of $\mathbf{x}$ for which $p(\mathbf{x}|\mathbf{z})$ is maximal.

In summary the conditional probability distribution can be modelled by defining a graph structure relating the random variables, and the feature functions with the proper domain depending on the information available.

III. Semantic Segmentation with CRFs

Given the general formulation for CRFs, we want to apply it to our problem of semantic segmentation. First we show how to solve the problem using only visual information acquired from a monocular camera. Then, we describe the proposal of [3] when we have two sensor modalities with common spatial coverage, in this case visual and 3D information. And finally, we show our strategy for the same two sensor modalities but with different spatial coverage.

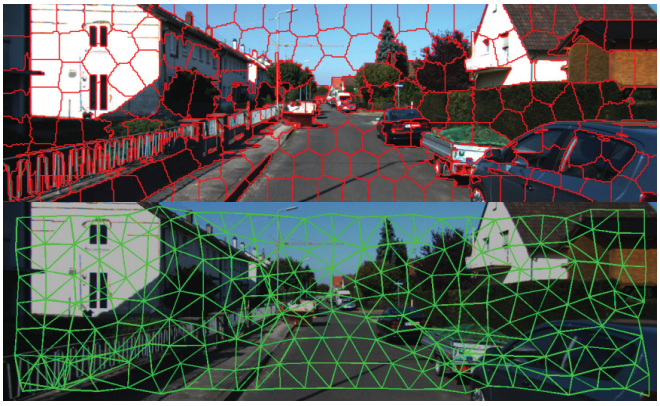


Fig. 3: On top, the original image with the SLIC superpixel over-segmentation [1]. Below, the image graph computed from the superpixels' neighbourhood.

A. Image only

This is a classic problem in computer vision. The strategy explained here is a basic approach which allows us explain the recipe for semantic segmentation with CRFs.

1st: The first step is to define the graph. We adopt commonly used strategy of superpixel over-segmentation to obtain the nodes in the graph, and the edges are defined by the superpixel neighbourhood in the image. To illustrate this step the image over-segmentation and the image graph are shown in Fig. 3.[1]

2nd: In the second step we define the unary and pairwise feature functions. For the unary feature function (Eq. 3) we use a k-NN classifier (Eq. 4) as defined in [24].

$$f_1(\mathbf{x}_s, \mathbf{z}_{im}) = -\log P_s(\mathbf{x}_s|\mathbf{z}_{im}) \tag{3}$$

$$P_s(\mathbf{x}_s = l_j|\mathbf{z}_{im}) = \frac{1}{\sum_{j=1}^{m} \left(\frac{\mathrm{f}(l_j)}{\bar{\mathrm{F}}(l_j)} \frac{\bar{\mathrm{f}}(l_j)}{\mathrm{F}(l_j)} \right)} \frac{\mathrm{f}(l_j)}{\bar{\mathrm{F}}(l_j)} \frac{\bar{\mathrm{f}}(l_j)}{\mathrm{F}(l_j)} \tag{4}$$

where $\mathrm{f}(l_j)$ (resp. $\bar{\mathrm{f}}(l_j)$) is the number of neighbours to superpixel s with label l_j (resp. not l_j) in the kd-tree. And $\mathrm{F}(l_j)$ (resp. $\bar{\mathrm{F}}(l_j)$) is the counting of all the observations in the training data with label l_j (resp. not l_j).

In the outdoors case [3], the 13-D feature vector is composed from the mean and standard deviation of LAB and RGB color spaces of the superpixel, and by the vertical superpixel centroid in the image. In the indoors case [4], the 8-D feature vector is composed from the mean and standard deviation of LAB color space, the vertical superpixel centroid in the image, and the entropy of the probability distribution

[1] For the sake of visual clarity the example in Figs. 3, 4 and 5 use a coarser over-segmentation than the actually used in the experimental section.

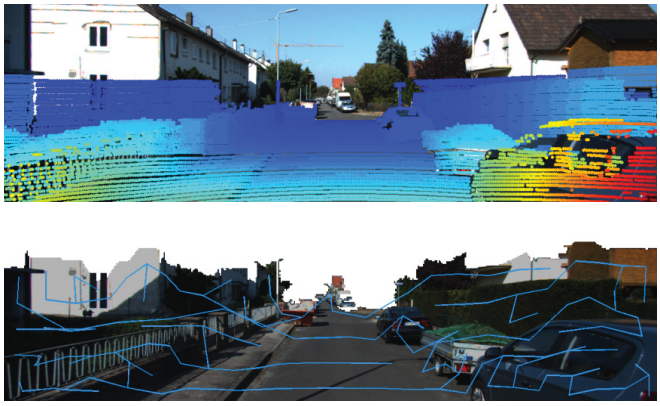


Fig. 4: On top, the 3D point cloud re-projected on the image. Below, the 3D graph (blue) computed as the minimum spanning tree over the Euclidean 3D distances between superpixels and re-projected on the image.

for the superpixel's boundaries belonging to the dominant vanishing points.

The pairwise feature function is defined in Eq. 5.

$$g_1(\mathbf{x}_{i,j}, \mathbf{z}_{im}) = \begin{cases} 1 - \exp\left(-\|c_i - c_j\|_2\right) & \rightarrow \quad l_i = l_j \\ \exp\left(-\|c_i - c_j\|_2\right) & \rightarrow \quad l_i \neq l_j \end{cases} \tag{5}$$

where $\|c_i - c_j\|_2$ is the L2-Norm of the difference between the mean colors of two superpixels in the LAB-color space and l is the class label.

3rd: The next step is the learning stage and finally the inference. Given that we have defined a cyclic graph structure the inference is carried out by loopy belief propagation and the learning of the weights $[w_{u_1}, w_{p_1}]$ by minimizing the negative log pseudo-likelihood [11].

B. Image and 3D: Intersection

When information from another sensor becomes available we want to infer the class in the spatial regions that are covered by all the sensors. In this case, we want to use the visual information jointly with the 3D information from a laser range finder or from a depth sensor. The steps described here are a summary of the proposals in [3] for outdoors and [4] for indoors.

1st: We use the superpixel over-segmentation to obtain the nodes in the graph. Hence, the 3D point cloud is clustered through this over-segmentation by re-projecting them on the image. The graph edges are defined by the minimum spanning tree (MST) over the Euclidean distances between 3D superpixel's centroids in the scene. The area of coverage of the laser and the 3D graph are shown in Fig. 4.

2nd: The unary and pairwise feature functions take the same form of Eqs. 3 and 5. But with the new information available we augment the feature vector for the kd-tree in the k-NN classifier. The previous feature vector is augmented with a 8-D vector of 3D features, containing the 3D centroid, the local and neighbourhood planarity, the vertical orientation, and the mean and std for the differences in depth for the superpixel wrt. its image neighbours. In the indoor case the horizontal coordinate for the centroid is not used, ending in a 7-D vector of 3D features. Eq. 6 shows the unary feature function using this new k-NN classifier and Eq. 7 is used jointly with Eq. 5 as pairwise feature functions.

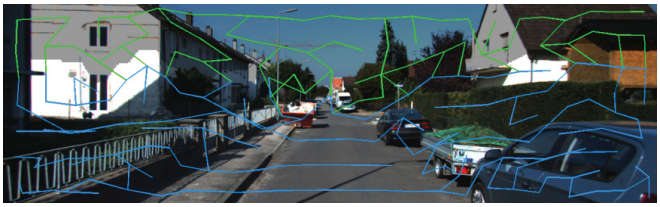


Fig. 5: Full graph. The edges determined by the 3D distances are in blue, and those by color distances are in green. We have shaded the image area with no 3D data related.

$$f_2(\mathbf{x}_s, \mathbf{z}_{im,3D}) = -\log P_s(\mathbf{x}_s | \mathbf{z}_{im,3D}) \tag{6}$$

$$g_2(\mathbf{x}_{i,j}, \mathbf{z}_{3D}) = \begin{cases} 1 - \exp\left(-\|\vec{p}_i - \vec{p}_j\|_2\right) & \rightarrow \quad l_i = l_j \\ \exp\left(-\|\vec{p}_i - \vec{p}_j\|_2\right) & \rightarrow \quad l_i \neq l_j \end{cases} \tag{7}$$

where $\|\vec{p}_i - \vec{p}_j\|_2$ is the L2-Norm of the difference between centroid's 3D positions.

3rd: In this case the MST over 3D has induced the *tree* graph structure, as such the inference process can be carried out in an exact way efficiently by the belief propagation algorithm [11]. The learning of the weights $[w_{u_2}, w_{p_1}, w_{p_2}]$ is done as before.

C. Image and 3D: Union

In the previous approach the graphical model is induced by 3D point cloud, and is able to explain only the intersection of the field of view of the camera and the 3D sensor. In the proposal presented below, we show how to assign semantic labels to the union of spatial coverage of both sensors.[2] We do so by augmenting the three ingredients of the CRFs approach (graph structure, potentials and learning and inference) to incorporate non-overlapping sensor coverage. In the experimental section we demonstrate improvements in the semantic segmentation accuracy, while maintaining the efficiency of the system.

1st: To build the full graph we need edges relating the 3D point cloud, the image and some connection between them. We first rely on the approach presented in the previous section and construct the sub-graph over the intersection of sensors coverage. Namely, we use the superpixels to cluster the point cloud but note that any other 3D clustering method is suitable for this purpose if the image is not available. Then, the 3D sub-graph is identical to that shown in Fig. 4. For the image sub-graph we use the image superpixels' neighbourhood without 3D information, a subset of Fig. 3 bottom. Note that some of the neighborhood edges end in the nodes of the 3D sub-graph giving us the connection between sub-graphs. Within this set of connections we find a MST over the distances between the superpixels' LAB-color space,

[2]The 3D laser has actually 360° of horizontal field of view but the ground truth labels are not available outside of image field of view. Note that the approach proposed here is equally applicable to parts with only 3D data extending the 3D graph to some point cloud over-segmentation, in which case the graphical model would take the form of Fig. 2.

TABLE I: KITTI dataset, semantic segmentation recall accuracy in pixel-wise percentage.

	ground		objects					building	vegetation	sky	Average	Global	Coverage
	road	pavem.	car	fence	post	people	sign.						
Image only													
Sengupta et al. [21]	98.3	91.3	93.9	48.5	49.3	—	—	97.0	93.4	—	81.7	88.4	—
CRF–Im	97.8		61.1					87.4	94.6	97.6	87.7	85.5	100
CRF–Im ∩ 3D[a] [3]	97.3		82.9					82.8	86.9	—	87.5	88.4	60.1
CRF–Im ∪ 3D	96.6		83.6					86.1	94.3	93.7	91.6	90.1	100

[a] Accuracy computed over the effective coverage region.
— stands for values that are no provided in [21] or are not possible to obtain with the approach of [3].

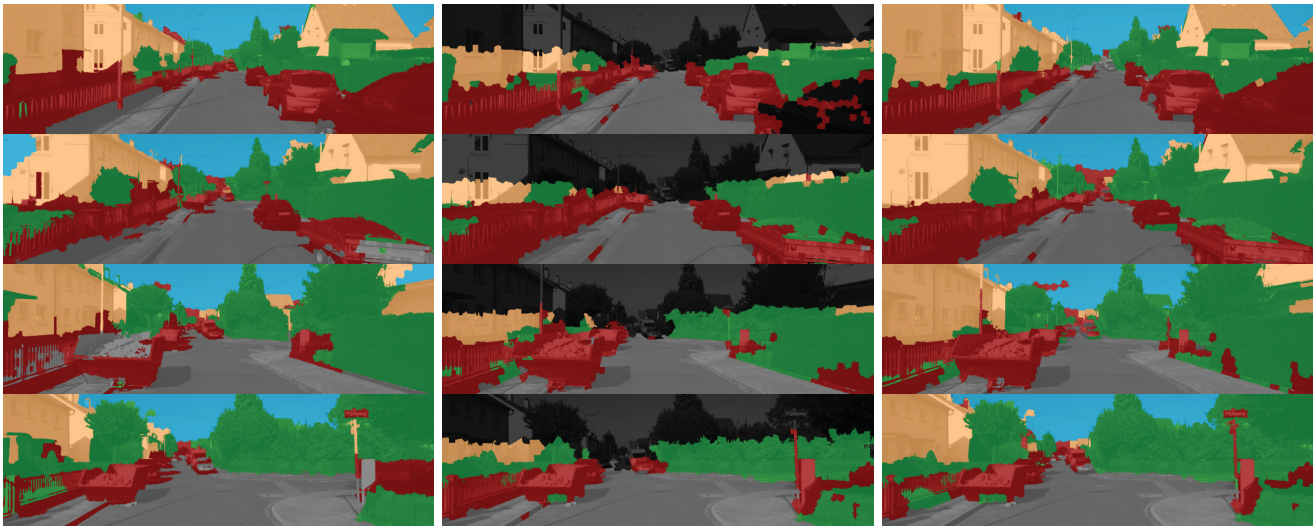


Fig. 6: Results of different settings on KITTI dataset. From left to right: CRF–Im, CRF–Im ∩ 3D and CRF–Im ∪ 3D. Color code: ■*ground*, ■*objects*, ■*building*, ■*vegetation* and ■*sky*.

inducing the image sub-graph. In Fig. 5 we show the graph for the full coverage, the 3D sub-graph in blue and the image sub-graph in green.

2nd: We have already shown very simple and effective features for each sensing modality. We will use the associated feature functions with their proper domain, as follows:

$$\begin{aligned} f_1(\mathbf{x}_s, \mathbf{z}_{im}) \quad & if \quad s \in im \setminus 3D \\ f_2(\mathbf{x}_s, \mathbf{z}_{im,3D}) \quad & if \quad s \in im \cap 3D \\ g_1(\mathbf{x}_{i,j}, \mathbf{z}_{im}) \quad & if \quad i \wedge j \in im \\ g_2(\mathbf{x}_{i,j}, \mathbf{z}_{3D}) \quad & if \quad i \wedge j \in 3D \end{aligned}$$

This is the same case already shown in Fig. 2 after removing f_3.

3rd: The weight vector to learn in this setting is composed by $[w_{u_1}, w_{u_2}, w_{p_1}, w_{p_2}]$, which is again obtained by minimizing the negative log pseudo-likelihood. The inference is carried out by loopy belief propagation because even though the two sub-graphs are trees, the full graph contains loops. However, we have found that convergence was achieved in all of our experiments in very few iterations.

IV. Experiments

We report the performance of the proposed method on two different real environments, urban outdoors and indoors. In outdoors we use Velodyne laser range sensor and camera and in indoors Kinect RGB-D sensor.

A. Outdoors: Image + 3D laser scan

We use the KITTI dataset [7], which contains images (1240x380) and 3D laser data taken for a vehicle in different urban scenarios. There are 70 manually labelled images as ground truth made available by [21], 45 for training and 25 for testing. The original classes released by [21] are: road, building, vehicle, people, pavement, vegetation, sky, signal, post/pole and fence. We have mapped those to five more general classes: ground (road and pavement), building, vegetation, and objects (vehicle, people, signal, pole and fence) and sky.

In this experiment we test the three cases of sensor overlap described in Section III; semantic segmentation using only the image information: CRF–Im, using the intersection between image and 3D: CRF–Im ∩ 3D, and using the full coverage: CRF–Im ∪ 3D. The results are shown in Fig. 6 and Table I for a qualitative and quantitative evaluation, respectively. As reference we show the results reported in [21] in Table I. Using only the image information, CRF-Im, we can obtain very good results in classes like *ground*, *vegetation* and *sky*, but a poor performance in *objects*. Adding shape evidence, CRF-Im ∩ 3D, the performance on the *objects* class is boosted with the disadvantage of parts of the scene with no semantic explanation. When we use all the available information, CRF-Im ∪ 3D, we obtain the best average performance over all the classes and solve the deficiencies of using only one modality without sacrificing the coverage of the scene. For instance in Fig. 6, the mistake

TABLE II: NYU dataset, semantic segmentation recall accuracy in pixel-wise percentage.

	ground	furniture	props	structure	Average	Global	In-painting	Coverage
Silberman et al. [22]	68	70	42	59	59.6	58.6	Required	100
Couprie et al. [5]	87.3	45.3	35.5	86.1	63.5	64.5	Required	100
CRF–Im $\cap$ 3D [4]	88.4	64.1	30.5	78.6	65.4	67.2	Required	100
CRF–Im $\cap$ 3D raw-depth[b]	88.5	69.0	23.1	78.6	64.8	67.4	No	74.6
CRF–Im $\cup$ 3D	87.9	63.8	27.1	79.7	64.3	67.0	No	100

[b] Accuracy computed over the effective coverage region.

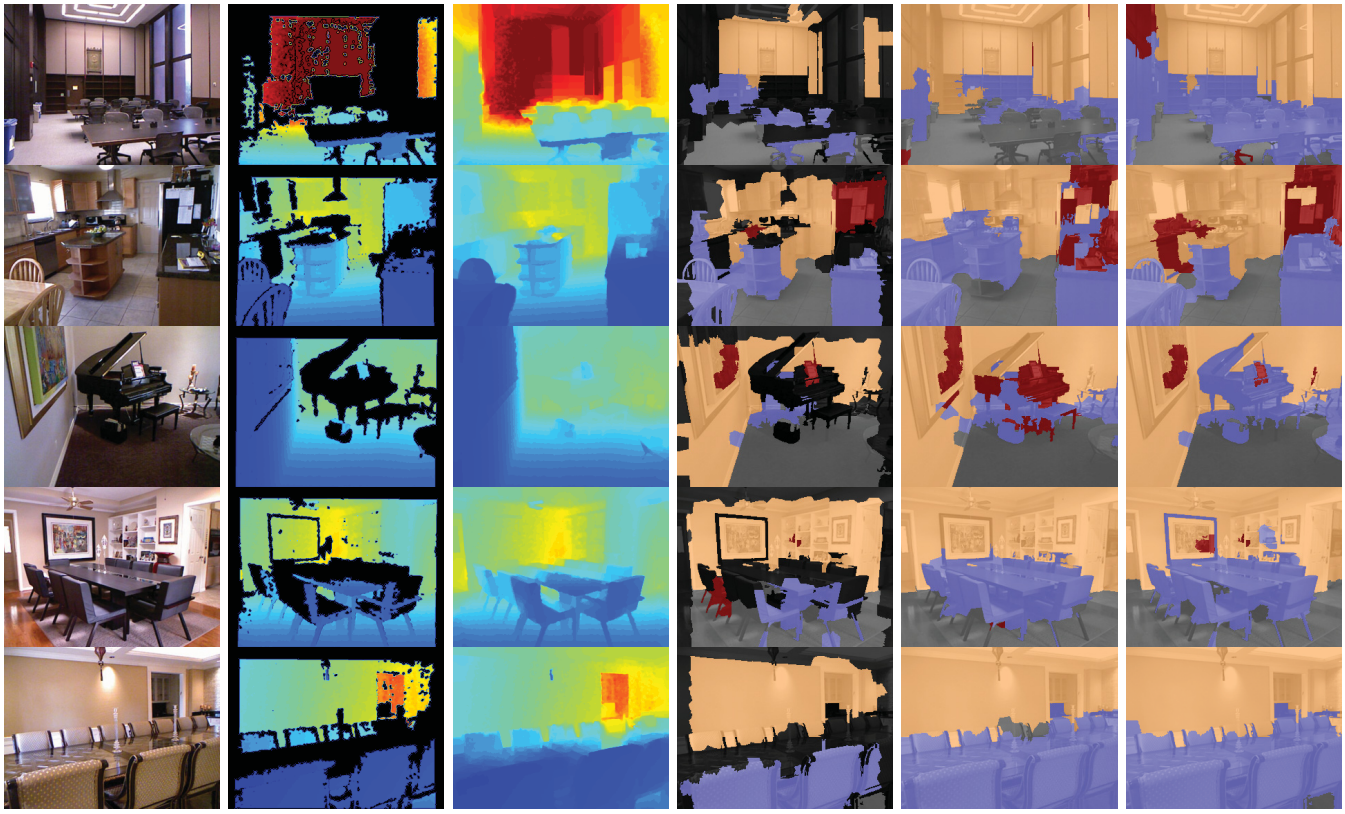


Fig. 7: Results of different settings on NYU dataset. First three columns show the RGB, the raw depth and the in-painted depth images. The next two columns show the results of the system of [4] for CRF–Im $\cap$ 3D using the raw and in-painted depths. The last column is the result from our proposal CRF–Im $\cup$ 3D using RGB and raw depth channels. Labels color code: ■*ground*, ■*structure*, ■*furniture* and ■*props*.

of assigning *ground* to the phone cable box is solved with shape evidence, last row. And, the propagation through the graph of *building* class solves the wrong assignment of *sky* to part of the house, second row.

In terms of efficiency the average cost, after segmentation and feature computation, is 40ms for CRF–Im, 16ms for CRF–Im $\cap$ 3D, and 23ms for CRF–Im $\cup$ 3D. With GPU implementation for SLIC superpixels and features extraction in Matlab any of the settings runs in less than one second.

B. Indoors: Kinect sensor

We use the NYU V2 RGB-D dataset [22], which contains 1449 labeled frames. The labeling spans over 894 different classes produced using Amazon Mechanical Turk. The authors of the dataset also provide a train and test splits and a mapping from 894 categories to 4 classes: *ground*, *structure*, *furniture* and *props*, as was used in [22]. We take 795 frames for learning, and the remaining 654 frames for testing and quantitative comparison.

In this experiment we want to evaluate the effectiveness of in-painting techniques for obtaining dense depth vs using our system CRF–Im $\cup$ 3D over raw depth. In Table II we show three state of the art results [4,5,22] using dense depth, also, as reference the result of [4] using the raw depth (CRF-Im$\cap$3D raw-depth), and in the last row the result from our proposal CRF–Im $\cup$ 3D.

Our solution, CRF–Im $\cup$ 3D, achieves state of the art performance without the expensive extra in-painting stage. A visual comparison in Fig. 7, shows the benefits of using the union of both modalities more than their intersection. For example, in the first and third rows of Fig. 7 the depth data for the tables are missed and the in-painting does not estimate the correct values, leading CRF–Im $\cap$ 3D to assign the class *ground* in those regions, something dangerous for the navigation system in a mobile robot. Also in the third row, the missing depth for the piano cover is filled by the in-painting with the depth of the wall behind, resulting in wrong labeling as structure. Those cases are correctly handled by our CRF–Im $\cup$ 3D strategy, where in the missing depth regions only image information gives the local evidence

and the information from regions with depth is correctly propagated through the graph to them. Note that for scenes without missing depth both systems ($\cup$ and $\cap$) are equivalent (up to no overlap for different field of views).

In the indoors setting experiment the average computational cost, after segmentation and feature computation, is 7ms for CRF−Im$\cap$3D using in-painting (15s), and 10ms for CRF−Im$\cup$3D using raw depth.

V. Conclusions

In this work we have addressed the problem of semantic mapping using the information from different sensors. We exploited the versatility and flexibility of the conditional random fields, to connect and use different sensory modalities in case they have different coverage areas. The presented proposal also handles the cases of missing data commonly encountered in RGB-D sensors and correctly propagates evidence from areas with available 3D information.

We have tested our proposal on real data, from outdoors and indoors environments, demonstrating its advantages over the existing alternatives to the problem of semantic segmentation. In our experiments due to the non-empty intersection between sensor coverages, we obtained a connected graph, but this is no a requirement for our approach. In fact, if the final graph is a forest of tree we can still carry out the inference process. The approach presented in this paper is very standard, and can easily be adapted to the recursive approach described in [3] for on-line semantic segmentation. The outcome from our system can be used and enhanced by specific object detectors in any (combination of) sensor modalities.

In the future work we plan to apply our semantic segmentation proposal to stereo cameras alone [13], and in combination to the 360° 3D laser sensor [7]. The approach proposed and evaluated here is not limited to a fixed number of sensors or sensor modalities. For instance, we can formulate the system for any number, and any kind of cameras: CRF−$\text{Im}_1 \cup \text{Im}_2 \cup \ldots \text{Im}_N$, or combination with alternative sensors (e.g. radar, thermal, or infra-red cameras) [15,17,19]: CRF−Im$\cup$Th$\cup$Radar$\cup$IR$\cup$3D.

The proposed augmented CRF framework which enables single inference of sub-graphs, induced by different sensing modalities, can be also applied to data association in place recognition tasks [2].

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274 –2282, nov. 2012.

[2] C. Cadena, D. Gálvez-López, J.D. Tardós, and J. Neira. Robust place recognition with stereo sequences. *IEEE Transaction on RObotics*, 28(4):871 –885, 2012.

[3] C. Cadena and J. Košecká. Recursive Inference for Prediction of Objects in Urban Environments. In *International Symposium on Robotics Research*, Singapore, December 2013.

[4] C. Cadena and J. Košecká. Semantic Parsing for Priming Object Detection in RGB-D Scenes. In *3rd Workshop on Semantic Perception, Mapping and Exploration*, Karlsruhe - Germany, June 2013.

[5] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *CoRR*, abs/1301.3572, 2013.

[6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1, 2013.

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and PatternRecognition (CVPR)*, 2012.

[8] S. Gupta, P. Arbeláez, and J. Malik. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2013.

[9] H. Hu, D. Munoz, J.A. Bagnell, and M. Hebert. Efficient 3-d scene analysis from streaming data. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, June 2013.

[10] A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *ICCV Workshop on Consumer Depth Cameras for Computer Vision*, 2011.

[11] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[12] L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P.H.S. Torr. What, Where and How Many? Combining Object Detectors and CRFs. In *Computer Vision - ECCV 2010*. Springer Berlin Heidelberg, 2010.

[13] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[14] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Trans. Graph.*, 23(3):689–694, August 2004.

[15] W. Maddern and S. Vidas. Towards robust night and day place recognition using visible and thermal imaging. In *RSS 2012: Beyond laser and vision: Alternative sensing techniques for robotic perception*, 2012.

[16] B. Micusik, J. Košecká, and G. Singh. Semantic parsing of street scenes from video. *The International Journal of Robotics Research*, 31(4):484–497, 2012.

[17] A. Milella, G. Reina, J. Underwood, and B.Douillard. Visual ground segmentation by radar supervision. *Robotics and Autonomous Systems*, (0):–, 2012.

[18] D. Munoz, J.A. Bagnell, and M. Hebert. Co-inference for multi-modal scene analysis. In *Computer Vision - ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 668–681. Springer Berlin Heidelberg, 2012.

[19] T. Peynot, S. Scheding, and S. Terho. The Marulan Data Sets: Multi-sensor Perception in a Natural Environment with Challenging Conditions. *The International Journal of Robotics Research*, 29(13):1602–1607, 2010.

[20] I. Posner, M. Cummins, and P. Newman. A generative framework for fast urban labeling using spatial and temporal context. *Autonomous Robots*, 26:153–170, 2009.

[21] S. Sengupta, E. Greveson, A. Shahrokni, and P.H.S. Torr. Urban 3D Semantic Modelling Using Stereo Vision. In *ICRA*, 2013.

[22] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 2012.

[23] G. Singh and J. Košecká. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2013.

[24] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *Computer Vision - ECCV 2010*. Springer Berlin Heidelberg, 2010.

[25] R.A. Triebel, R. Paul, D. Rus, and P. Newman. Parsing outdoor scenes from streamed 3d laser data using online clustering and incremental belief updates. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[26] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 686 –693, oct. 2009.