A Spatio-Temporal Probabilistic Model for Multi-Sensor Multi-Class Object Recognition

Bertrand Douillard[1], Dieter Fox[2], and Fabio Ramos[1]

[1] ARC Centre of Excellence for Autonomous Systems, Australian Centre for Field Robotics, University of Sydney, Sydney, NSW, Australia
{b.douillard,f.ramos}@cas.edu.au

[2] Dept. of Computer Science & Engineering, University of Washington, Seattle, WA, USA fox@cs.washington.edu

Abstract. This paper presents a general probabilistic framework for multi-sensor multi-class object recognition based on Conditional Random Fields (CRFs) trained with virtual evidence boosting. The learnt representation models spatial and temporal relationships and is able to integrate arbitrary sensor information by automatically extracting features from data. We demonstrate the benefits of modelling spatial and temporal relationships for the problem of detecting seven classes of objects using laser and vision data in outdoor environments. Additionally, we show how this framework can be used with partially labeled data, thereby significantly reducing the burden of manual data annotation.

1 INTRODUCTION

Reliable object recognition is an important step for enabling robots to reason and act in the real world. A high-level perception model able to integrate multiple sensors can significantly increase the capabilities of robots in tasks such as obstacle avoidance, mapping, and tracking.

We present a multi-modal object detector based on Conditional Random Fields (CRFs). CRFs are discriminative models for classification of sequential (dependent) data, directly modelling the conditional probability $p(\mathbf{x}|\mathbf{z})$ of hidden states $\mathbf{x}$ given observations $\mathbf{z}$ [4]. The proposed framework uses the general applicability of CRFs as a unifying methodology to learn spatial and temporal relationships between observations obtained with a laser range-finder and a camera. The model is trained on data collected by a moving vehicle to detect seven classes of objects in an urban environment.

By building on the recently developed Virtual Evidence Boosting (VEB) algorithm [5], the algorithm described here is able to automatically select features during the learning phase. The expert knowledge about the problem is encoded as a selection of features capturing particular properties of the data such as geometry, colour and texture. Given a labeled training set, VEB computes weights for each of these features according to their importance in discriminating the data. Additionally, an extension of VEB for semi-supervised learning is presented to address datasets with partially labeled samples.

2 RELATED WORK

Within the robotics community, researchers have recently developed representations of environments using more than one modality. In [9], a 3D laser scanner and loop closure detection based on photometrical information are brought together in the Simultaneous Localisation and Mapping (SLAM) framework. This approach does not generate a semantic representation of the environment which can be obtained from the same multi-modal data using the approach proposed here. In [11], a robust landmark representation is created by probabilistic compression of high dimensional vectors containing laser and camera information. This representation is used in a SLAM system and updated on-line when a landmark is re-observed. However, it does not readily allow the inference of a landmarks' class which could contribute to higher level reasoning.

Object recognition based on laser and video data has been demonstrated in [7]. Using a sum rule, this approach combines the outputs of two classifiers, each of them being assigned to the processing of one type of data. In contrast, we learn a CRF classifier with the VEB algorithm which performs joint feature selection in both datasets in order to minimise the classification error on training data. The VEB algorithm can, as it is, learn a classifier given as many data types as available and is not restricted to laser and vision inputs [5].

In robotics, CRFs have been applied successfully in the context of semantic place labeling [3] and object recognition [6]. However, neither of these approaches incorporated multiple sensor modalities and performed feature selection via VEB training. In previous work, we showed how cars can be detected in laser and vision data using CRFs [1]. Here, we show how such a system can additionally perform semi-supervised learning and demonstrate its applicability to multi-class scenarios. The key contribution of this work is to present a probabilistic model for multi-class object recognition which integrates spatial and temporal correlations and can be learnt given any types of partially labeled data.

3 Conditional Random Fields

This section provides a brief description of conditional random fields (CRF) and virtual evidence boosting (VEB), an extremely efficient way of learning CRF parameters for arbitrary feature functions (see [5] and [14] for more information).

3.1 Model Description and learning

Conditional random fields (CRF) are undirected graphical models developed for labeling sequence data [4]. CRFs directly model $p(\mathbf{x}|\mathbf{z})$, the *conditional* distribution over the hidden variables $\mathbf{x}$ given observations $\mathbf{z}$. CRFs factorize $p(\mathbf{x}|\mathbf{z})$ as

$$p(\mathbf{x} \mid \mathbf{z}) = \frac{1}{Z(\mathbf{z})} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{z}, \mathbf{x}_c), \tag{1}$$

where $Z(\mathbf{z}) = \sum_{\mathbf{x}} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{z}, \mathbf{x}_c)$ is the normalizing partition function. $\mathcal{C}$ is the set of cliques in the CRF graph. The ϕ_c are *clique potentials*, which are functions that map variable configurations to non-negative numbers. Intuitively, these potentials capture the "compatibility" among the variables in the clique: the larger the potential value, the more likely the configuration. Potentials are constrained to log-linear functions, and learning a CRF requires learning the weights of these functions.

Even though CRFs can handle extremely high-dimensional and dependent feature vectors, the adequate modeling of continuous observations is not straightforward. Recently, Liao and colleagues introduced virtual evidence boosting (VEB), which learns an appropriate discretisation of continuous observations along with the weight parameters of the model [5]. It does this applying the logitboost learning algorithm to both the observations and connections of the CRF. VEB has demonstrated superior performance on both synthetic and real data. The automatic observation discretisation makes VEB extremely flexible and allows the incorporation of arbitrary, continuous and discrete observations.

Through VEB, a CRF model can not only be learnt with fully labeled data but also with partially labeled data. In this context, unlabeled data is ignored when learning the logitboost classifiers for local observations. However, the unlabeled data is used to estimate the *distributions* over all hidden states in the CRF, which can have significant impact on the learning result, as we will show in the experimental results.

3.2 Inference

Inference in CRFs consists either in estimating the marginal distribution of each hidden variable $\mathbf{x}_i$ or in defining the most likely configuration of all hidden variables $\mathbf{x}$ (*i.e.*, MAP estimation), based on their joint conditional probability 1. Both tasks can be solved using BP.

BP provides exact results in graphs with no loops, such as trees or polytrees. However, as the models used in our approach contain various loops, we apply loopy belief propagation, an approximate inference algorithm that is not guaranteed to converge to the correct probability distribution [8]. Fortunately, in our experiments, this approximation is reasonably accurate even when loopy BP failed to converge (the maximum number of iterations is reached).

4 CRFs FOR OBJECT RECOGNITION

This section describes the deployment of the CRF framework to perform object recognition. This paper focuses on performing object detection in an outdoor urban environment given laser data and monocular colour images. Fig. 1 shows two examples of laser scans projected into their corresponding image according to the procedure described in [16]. The CRF framework is applied by converting each scan into a linear chain CRF, as displayed in Fig. 2. Each node of this CRF represents a laser return. The hidden variable to be estimated is the class of the return. The features **z** used in our object recognition CRF model are now described. We then explain how the CRF model of a scan is further incorporated into a more elaborated representation which takes temporal relationships into account.

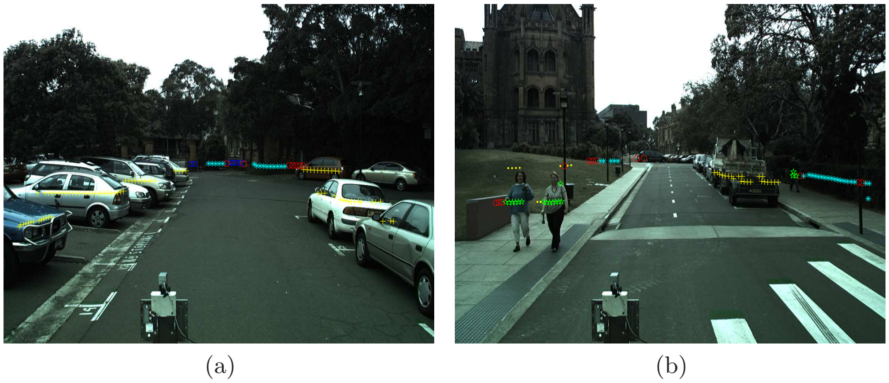

(a) (b)

Fig. 1. Two examples of laser scans and associated images. Laser returns are projected into the image and displayed with different markers indicating their label: a yellow + means class “car”, a magenta ˆ “trunk”, a cyan ∗ “foliage”, a green star “people”, a blue □ “wall”, yellow dot “grass” and a red ○ “other”.

4.1 One time slice model

To jointly estimate all the labels of a laser scan, features **z** are extracted from laser and camera data. Two feature functions are used in the experiments: geometric feature functions and visual feature functions. We now detail each of them.

Geometric laser features

These features capture geometric properties of the objects detected by the laser scan. While local shape can be captured by various types of features, we chose to implement simple shape features measuring distance, angle, and number of out of range returns between two beams. The resulting feature function has the form

$$\mathbf{f}_{\text{geo}}(i, z_A) = \text{concat}\left(\mathbf{f}_{\text{dist}}(\mathrm{i}, \mathrm{z_A}), \mathbf{f}_{\text{angle}}(\mathrm{i}, \mathrm{z_A}), \mathbf{f}_{\text{oor}}(\mathrm{i}, \mathrm{z_A})\right), \tag{2}$$

where i indexes one of the returns in scan z_A. The concat function performs a concatenation operation, and $\mathbf{f}_{\text{oor}}$ refers to the out-of-range feature function. The resulting function $\mathbf{f}_{\text{geo}}(i, z_A)$ returns a vector of dimensionality 213, as specified next.

To generate distance features, we compute for each point $z_{A,i}$ in scan A its distance to other points in scan A. These other points are chosen based on their relative indices in the scan. With k being an index offset, the distance feature corresponding to points $z_{A,i}$ is computed as follows

$$\mathbf{f}_{\text{dist}}\left(i, k, z_A\right) = \left\| z_{A,i} - z_{A,i+k} \right\|. \tag{3}$$

In our implementation this feature is computed for index offsets k varying from -10 to $+10$.

Another way to consider local shape is by computing the angles of points w.r.t their neighbours. The angle of a point $z_{A,i}$ is defined as the angle between the segments connecting a point i to its neighbours. With k and l indicating and index offset, this feature is defined as:

$$\mathbf{f}_{\text{angle}}\left(i, k, l, z_A\right) = \left\| \angle\left(\overline{z_{A,i-k} z_{A,i}}, \overline{z_{A,i} z_{A,i+l}}\right) \right\|. \tag{4}$$

As with the distance feature, we compute a set of angles by varying k and l from -10 to $+10$.

The out of range feature counts the number of “out of range” beams between pairs of successive returns. The idea is to take into account open areas implicitly detected by the laser scan.

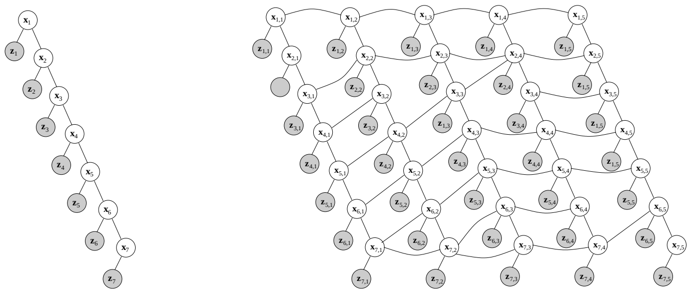
Visual features

In addition to geometrical information, a CRF model learnt with the VEB algorithm can seamlessly integrate vision data provided by a monocular colour camera. The first step consists of registering the vision sensor and the laser range-finder with respect to each other using the calibration procedure described in [16]. Laser returns can then be projected in the corresponding image. The visual features extracted from this image capture colour and texture information in the window (or ROI) centred around the laser return. The edge length of the window is set to be 1 metre for a range of 4 metres. This size is converted into number of pixels using the camera’s intrinsic parameters and adjusted depending on the range measurement. Changing the size of the extracted patch as a function of range is a procedure to cope with the variation in scales as an object moves from the background to the foreground of the image. It was verified that the use of a size varying window improves the experimental results by 4%.

The visual feature function has the form

$$\mathbf{f}_{\text{visu}}(p_i, p_{i-1}) = \text{concat}\left(\mathbf{f}_{\text{texture}}(\mathrm{p_i}, \mathrm{p_{i-1}}), \mathbf{f}_{\text{colour}}(\mathrm{p_i}, \mathrm{p_{i-1}})\right), \tag{5}$$

where p_i is the image patch corresponding to return i. $\mathbf{f}_{\text{texture}}(p_i, p_{i-1})$ returns a vector containing the steerable pyramid [13] coefficients of image patch i, and the difference between the steerable pyramids computed at patch i and at patch $i-1$. $\mathbf{f}_{\text{colour}}(p_i, p_{i-1})$ returns a vector containing the 3D RGB colour histogram of patch i and of its difference with patch $i-1$. Only neighbour $i-1$ is used to limit the dimensionality of $\mathbf{f}_{\text{visu}}$ which is already over 7000.



(a) (b)

Fig. 2. (a) Graphical model of a linear chain CRF for one time slice object recognition. Each hidden node $\mathbf{x}_i$ represents one (non out of range) return in a laser scan. The nodes $\mathbf{z}_i$ represent the features extracted from the laser scan and the corresponding camera image. (b) Graphical model of the spatio-temporal classifier. Nodes $\mathbf{x}_{i,j}$ represent the i-th laser return observed at time j. Temporal links are generated between time slices based on the ICP matching algorithm.

4.2 Recognition over time

Due to the sequential nature of robotics applications, a substantial amount of information can be gained by taking into account prior and posterior data when available. We now present a model that achieves temporal smoothing in addition to exploiting the geometric structure of laser scans. This model is displayed in Fig. 2.

In this work, the temporal connections are instantiated such that they represent the associations found by the Iterative Closest Point (ICP) matching algorithm [17]. The pairwise potentials assigned to these connections are set to identity. Mathematically, $\phi_{\text{temporal}}(\mathbf{x}_i, \mathbf{x}_j) = \delta(\mathbf{x}_i, \mathbf{x}_j)$, where δ is the indicator function. This set-up is justified by the fact that ICP associates returns that were generated by the same physical point. It follows that the integration of temporal information does not require additional learning.

Corresponding to different variants of temporal state estimation, our spatio-temporal model can be deployed to perform three different types of estimation.

- Off-line smoothing: All scans in a temporal sequence are connected using ICP. BP is then run in the whole network to estimate the class of each laser return in the sequence. During BP, each node sends to its neighbours the messages through structural and temporal links (vertical and horizontal links, respectively, in Fig. 2). In our experiments, BP is run for 100 iterations.

- On-line fixed-lag smoothing: Here, scans are added to the model in an on-line fashion. To label a specific scan, the system waits until a certain number of future scans becomes available, and then runs BP taking past and future scans into account.
- On-line filtering: In this case the spatio-temporal model only includes scans up to the current time point.

5 EXPERIMENTS

The experiments were performed using outdoor data collected with a modified car travelling at 0 to 40 km/h. The car drove along several loops in a university campus which has structured areas with buildings, walls and cars, and unstructured areas with bush, trees and lawn fields. The overall dataset contains 4500 images which represents 20 mins of logging. Laser data was acquired at a frequency of 4Hz using a SICK laser. The models presented in Sec. 4 are used to estimate the class of each return in the laser scans.

Results on a binary classification problem are first presented to facilitate the description of the spatial and temporal dependencies' role in the model (some of the binary classification results were already presented in [1], and are shown here for completeness). We then show how the possibility of performing semi-supervised learning can reduce the hand labelling effort by more than half. We finally present results on a classification problem involving seven classes.

5.1 Binary classification

In this first set of experiments we consider two classes: "car" and "other". Table 1 summarises the experimental results in terms of classification accuracy. The accuracies are given in percentages and computed using 10-fold cross validation on a set of 100 manually labeled scans. For each cross validation, different models were trained with 200 iterations of VEB. VEB was computed allowing learning of pairwise relationships only after iteration 100. We found that this procedure increases the weights of local features and improves classification results.

Training set	geo only	visu only	geo+visu	geo+visu
Number of time slices in the model	1	1	1	∓10
CRF	68.93	81.79	83.26	88.08
logitboost	67.64	81.52	83.22	×

Table 1. Binary classification accuracy (in %)

The first line of Table 1 indicates the types of features used to learn the classifier. Four different configurations were tested: first using geometric

Classifier (training set = geo+visu)	logitboost	CRF	CRF
Number of time slices in the model	1	1	∓ 10
String Edit Distance	9.5	5.6	2.4

Table 2. String Edit Distances in the binary classification case

features only, second containing visual features only, third containing both geometric and visual features, and fourth with geometric and visual features integrated over a period of 10 times slices. The second line of table 1 indicates the number of time slices in the network used to perform classification. “1” means that a network as presented in Fig. 3(a) was used. “∓ 10 ” refers to the classifier shown in Fig. 3(b) instantiated with 10 unlabeled scans prior and posterior to the labeled scan.

Two types of classifiers were used: CRFs and logitboost classifiers. CRFs take into account the neighbourhood information to perform classification (Fig. 3(a)). Logitboost learns a classifier that only supports independent classification of each laser return without neighbourhood information [2]. Logitboost is used here for comparison purposes to investigate the gain in accuracy obtained with a classifier that takes into account the structure of the scan.

The first three columns of Table 1 show that classification results are improving as richer features are used for learning. The first three columns also show that the CRF models lead to slightly more accurate classification.

In addition, as presented in Sec. 4.2, a CRF model can readily be extended into a spatio-temporal model. The latter leads to an improvement of almost 5% in classification accuracy (right column of table 1). This shows that the proposed spatio-temporal model, through the use of past and posterior information, performs better for object recognition. The cross in the bottom right of the table refers to the fact that logitboost does not allow the incorporation of temporal information in a straightforward manner.

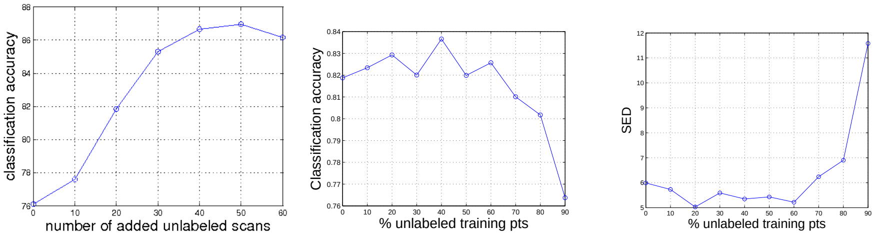
CRF models also generate better segmentation of cars in laser scans. This can be quantified using the metric called String Edit Distance (SED)[12]. By definition the SED is the smallest number of insertions, deletions, and substitutions required to change one string into another. Intuitively, this metric tells whether classification results capture the true arrangement of objects in a scene. It penalises series of estimates that do not respect the true sequence of blocks with the same label. For example, given the ground truth “ccooccoo” (where ’c’ and ’o’ stand for ’car’ and ’other’ respectively), the estimated sequence “cocococo” is more penalised (larger SED) than “ooccoocc”. This is because the latter estimate is more similar to the true sequence in terms of blocks of returns with the same label.

Table 2 presents classification results in terms of SEDs. The numbers show that the spatio-temporal model gives the best results in terms of classification accuracy as well as in terms of SEDs. The CRF classifiers, through their ability to represent spatial and temporal dependencies, are better at capturing the true arrangement of the observed objects. This property will be beneficial when the CRF models are applied to image segmentation problems (which is beyond the scope of this paper). These results resemble the results presented in [3] where it is shown that a CRF based approach is better in capturing the structure of indoor environments.

In order to gauge the difficulty of the task, we also performed logitboost classification using visual Haar features, which results in the well-known approach proposed by Viola-Jones [15]. The accuracy of this approach is 77.09%, which shows that even our single time slice approach (83.26%) outperforms the reference work of Viola & Jones. The improvement in accuracy obtained by the CRF model comes from its aptitude to capture neighborhood relationships and the use of richer features.

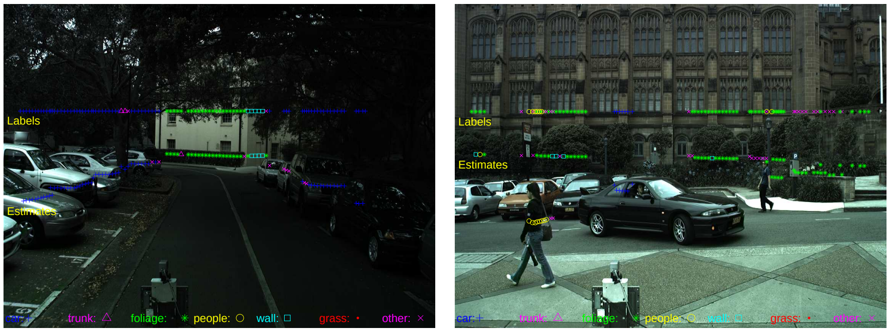
5.2 Semi-supervised learning

Fig. 3 presents binary classification results obtained with models learnt on datasets containing a progressively increasing amount of unlabeled data. Fig. 3(a) shows that adding unlabeled data while maintaining the number of labeled returns constant improves classification accuracy. This shows that the proposed framework allows to perform semi-supervised learning. In Fig. 3(b) and 3(c) the total number of scans used is maintained constant and the number of unlabeled returns is increased. These two plots show that the original accuracy is maintained with only 40% of labeled data. As a consequence, the proposed model enables non negligible economy in the manual labelling process.



(a) (b) (c)

Fig. 3. Semi-supervised learning behaviour for the binary classification case. Each point in the three plots corresponds to the average of 10 one-time-slice models learnt by cross validation. (a) The number of labeled scans is 30. As more unlabeled scans are added to the training set, labeled returns are spread evenly across the training set while their total number is maintained constant. (b) and (c) The training sets contain 90 scans and the testing sets contain 10 scans. The x coordinate means that x% of randomly chosen returns in each of the 90 scans are unlabeled.



(a) (b)

Fig. 4. Example of results obtained with the seven classes classifier. The legend is indicated in the bottom of each image. The top line of markers indicate the manually provided ground truth labels of each scan return directly below. Each scan return projected into the image is displayed with a marker indicating its estimated label.

5.3 Multi-class classification

In this set of experiments, seven classes of objects are involved: car, trunk, foliage, people, wall, grass, and other. The class "trunk" is used since the laser range finder was mounted on the front of the car and was mostly sensing trees at the level of their trunk. The class "foliage" refers to bushes, high grass and the top part of trees.

The classification results obtained with fully labeled data are summarised in table 3. The values show that the spatio-temporal model (right column) gives the best results in terms of classification accuracy as well as in terms of SED.

Classifier	logitboost	CRF	CRF
Number of time slices in the model	1	1	∓10
Accuracy [%]	57.27	59.02	60.73
String Edit Distance	5.9	2.6	0.9

Table 3. Multi-class classification results

Fig. 4 shows a few examples of classification results obtained with the multi-class classifier. It can be seen that most of the returns are correctly identified as belonging to one of the classes car, foliage, wall or people.

Inference in a one time slice CRF takes about 7 seconds on a Intel Xeon 2.33GHz desktop computer. Inference in the spatio-temporal model takes on average 27 seconds per scan. Our implementation of the inference is not very efficient and we believe that an optimized implementation will make it possible to run our system in real time applications. Learning the model requires about three hours and can be performed off-line.

6 CONCLUSIONS AND FUTURE WORK

We have presented a general probabilistic model for object recognition that incorporates spatial and temporal dependencies. Through the use of the VEB learning algorithm, we generate a CRF model that can combine any type of data by selecting features according to their relevance with respect to the classification task.

In future work we will improve the VEB algorithm by incorporating pairwise potentials that take into account local observations in addition to estimated labels, thereby appropriately adapting their smoothing effect. Furthermore, we intend to replace ICP matching by CRF-matching [10], an approach that performs probabilistic scan matching using conditional random fields. We believe that CRF-matching, when incorporated into our spatio-temporal model, will provide more robust estimation of data associations, especially when closing loops. Finally, we will investigate the incorporation of improved geometric features, such as lines and corners.

7 ACKNOWLEDGMENTS

The authors would like to thank Lin Liao for providing the VEB code and Juan Nieto, Jose Guivant, Roman Katz and Oliver Frank for helping with the dataset acquisition. This work is supported by the ARC Center of Excellence programme, the Australian Research Council (ARC), the New South Wales (NSW) State Government, the University of Sydney Visiting Collaborative Research Fellowship Scheme, and DARPA's ASSIST and CALO Programmes (contract numbers: NBCH-C-05-0137, SRI subcontract 27-000968).

References

1. B. Douillard, D. Fox, and F. Ramos. A spatio-temporal probabilistic model for multi-sensor object recognition. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007.
2. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.
3. S. Friedman, D. Fox, and H. Pasula. Voronoi random fields: Extracting the topological structure of indoor environments via place labeling. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.
4. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning (ICML)*, 2001.
5. L. Liao, T. Choudhury, D. Fox, and H. Kautz. Training conditional random fields using virtual evidence boosting. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

6. B. Limketkai, L. Liao, and D. Fox. Relational object maps for mobile robots. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.
7. G. Monteiro, C. Premebida, P. Peixoto, and U. Nunes. Tracking and classification of dynamic obstacles using laser range finder and vision. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006.
8. K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999.
9. P. Newman, D. Cole, and K. Ho. Outdoor SLAM using visual appearance and laser ranging. In *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, Orlando, USA, 2006.
10. F. Ramos, D. Fox, and H. Durrant-Whyte. CRF-matching: Conditional random fields for feature-based scan matching. In *Proc. of Robotics: Science and Systems*, 2007.
11. F. Ramos, J. Nieto, and H.F. Durrant-Whyte. Recognising and modelling landmarks to close loops in outdoor slam. In *Proc. of the IEEE International Conference on Robotics & Automation (ICRA)*, 2007.
12. D. Sankoff and J. Kruskal, editors. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, 1983.
13. E. Simoncelli and W. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. of 2nd International Conference on Image Processing*, 1995.
14. C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*, chapter hi. MIT Press, 2006.
15. P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
16. Q. Zhang and R. Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sendai, Japan, 2004.
17. Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.