

An Approach to Spatio-Temporally Consistent Scene Classification in Urban Workspaces: An Overview

I. Posner, M. Cummins and P. M. Newman

Mobile Robotics Group, Department of Engineering Science, Oxford University, Parks Road, Oxford, OX1 3PJ

Abstract

This paper outlines¹ a two-stage probabilistic approach to the semantic labelling of regions in an urban scene. In the first stage inference is performed locally in order to classify each image region separately. In the second stage scene-wide as well as temporal context is taken into account using a Markov random field (MRF) in order to refine the initial classifications. Although a range of classification frameworks could be employed to perform the first stage classification, we chose to explore a generative probabilistic classifier formulation for this purpose. We will indicate how this approach harbours potential with respect to classification speed as well as versatility of the proposed system. The approach is evaluated on real data gathered by a robot over 17 km of track through an urban environment.

Keywords : Generative model, TAN classifier, Markov random field, 3D laser, vision.

Introduction

This paper outlines a probabilistic method which achieves fast labelling of regions in a scene by performing inference at multiple scales: locally, using scene wide context and, finally, using context provided by evidence across an individual scene. Although the application here leverages a combination of 3D range and image data the proposed framework is by no means limited to these modalities. At a local scale, classification is based on the co-occurrence of appearance descriptors, which capture both visual and surface orientation information. We frame this classification problem in probabilistic terms, which provides an efficient and flexible framework. Secondly, at the scene-wide scale, we use a Markov Random Field (MRF) to model the expected spatial relationships between patch labels, thus capturing some of the strong structural relationships between parts of a typical

urban scene. Our MRFs have a relatively low node-count, just one node for each scene patch, yielding rapid inference.

Related Works

Recently there has been a surge in the literature regarding environment understanding within robotics, particularly as available sensory data become richer and the limitations of un-annotated maps become more apparent. Particularly relevant to the work presented here are papers which consider a combination of vision and laser data in an outdoor setting. [3] considers the task of pedestrian and vehicle detection, using 2D laser data. In [4] a more sophisticated inference framework based on Conditional Random Fields was brought to bear on the vehicle detection problem, with preliminary results also reported for multi-class labelling. 3D laser data were combined with visual information in [5], which used support

¹ A more detailed technical account of a similar system was first published in [1]. A considerably more detailed account of this system is soon to appear in [2]. The reader is referred to either of these publications for more information

vector machines for classification but did not make use of contextual information. The work presented here also leverages a combination of laser data with vision. Our main contribution lies in the definition of an efficient contextual inference framework, based on a graph over plane patches rather than over measurements (e.g. laser range data) directly. This yields substantial speed increases over previous approaches. As an integral part of this framework we further define a generative bag-of-words classifier and describe an efficient inference procedure for it. Finally, the work presented here further distinguishes itself from related work by combining information from two complimentary sensors – full 3D geometry and appearance. Thereby our approach gains the capacity of providing *more detailed* workspace descriptions such as the surface-type of building(s) encountered or the nature of ground traversed.

Urban Workspace Classes and Features

To enable comparison with our own prior work the classes and features used here are the same as those used in [6] and are summarised in Table 1 and Table 2, respectively. The reader is referred to the earlier publication for a detailed motivation and description.

Table 1: Classes

Class	Description
<i>Ground Type</i>	
Pavement/Tarmac	Road, footpath.
Dirt Path	Mud, sand, gravel.
Grass	Grass.
<i>Building Type</i>	
Smooth Wall	Concrete, plaster, glass.
Textured Wall	Brickwork, stone.
<i>Object</i>	
Foliage	Bushes, tree canopy.
Vehicle	Car, van.

Table 2: Features

Feature Descriptions	Dimensions
3D Geometry Orientation of surface normal of the local plane	1
2D Geometry Location in image: mean of normalized x and y	2
Colour HSV: hue & sat. histograms in a local neighbourhood	30
Texture HSV: hue & sat. variance in a local neighbourhood	2

Classification Framework

The inference framework proposed in this paper is a multi-level approach based on successive combinations of lower-level features. The lowest level input to our system is the collection of laser points in the scene. Each laser point is described by a feature vector using the feature set described above. Rather than deal with raw data directly, we adopt the bag-of-words representation [7], where the feature vectors are quantised with respect to a ‘vocabulary’. The vocabulary is constructed by clustering all the feature vectors from a set of training data, using a fixed-radius incremental clustering algorithm.

This yields a vocabulary of size $|v|$, defined by the cluster centres. The vocabulary size is determined by a user-specified threshold. For this work we use a vocabulary of approximately 6,500 words. When the system has been trained, each incoming laser point yields a feature vector, which is quantised to the approximate nearest cluster centre using a kd-tree. The image is segmented into patches using an off-the-shelf algorithm [8]. The laser points contained within each patch then define a bag-of-words. The bag-of-words for each patch in the scene is the input to the next level of the system, which aims to provide a soft classification of any given patch. The remainder of this section provides a description of the generative probabilistic

model we employ for this classification task.

Stage I: Local Classification

Our patch-level classifier is inspired by the probabilistic appearance model introduced in SEN11 (also see [9] for details) and the theory presented below is an extension of that work into a more general classification framework. Building on the output of the lower level vector quantisation step, an observation of a patch $\mathbf{z} = \{z_1, \dots, z_{|v|}\}$ is a collection of binary variables where each z_i indicates the presence (or absence) of the i^{th} word of the vocabulary within the patch. We would like to compute $p(\mathcal{C}|\mathbf{z})$, the distribution over the class labels given the observation, which can be computed according to Bayes rule:

$$p(\mathcal{C}^k|\mathbf{z}) = \frac{p(\mathbf{z}|\mathcal{C}^k)p(\mathcal{C}^k)}{p(\mathbf{z})} \quad (1)$$

where $p(\mathbf{z}|\mathcal{C}^k)$ is the class-conditional observation likelihood, $p(\mathcal{C}^k)$ is the class prior and $p(\mathbf{z})$ normalises the distribution.

Given a vocabulary, individual classes are represented within the classification framework by a set of class-specific examples, which we call *exemplars*. Concretely, for each class k the model consists of n_k exemplars $\mathcal{C}^k = \{\mathcal{C}_1^k, \dots, \mathcal{C}_{n_k}^k\}$ where \mathcal{C}_i^k is the i^{th} exemplar of class k . Exemplars themselves are defined in terms of a hidden ‘existence’ variable e , each exemplar \mathcal{C}_i^k being described by the set $\{p(e_1 | \mathcal{C}_i^k), \dots, p(e_{|v|} | \mathcal{C}_i^k)\}$. The term e_j is the event that a patch contains a property or artefact which, given a perfect sensor, would cause an observation of word z_j . However, we do not assume a perfect sensor — observations \mathbf{z} are related to existence e via a sensor model which is specified by

$$\mathcal{D}: \begin{cases} p(z_j = 1|e_j = 0), \text{ false positive probability.} \\ p(z_j = 0|e_j = 1), \text{ false negative probability.} \end{cases}$$

with these values being a user-specified input. The reasons for introducing this extra layer of hidden variables, rather than modelling the exemplars as a density over observations directly, are twofold. Firstly, as described in the previous section, it provides a natural framework for incorporating data from multiple sensors, where each sensor has different (and possibly time-varying) error characteristics. Secondly, as we will discuss later, it allows the calculation of $p(\mathbf{z}|\mathcal{C}^k)$ to blend local patch-level evidence with a global model of word co-occurrence.

The key step in computing the pdf over class labels as per Equation 1 is the evaluation of the conditional likelihood $p(\mathbf{z}|\mathcal{C}^k)$. This can be expanded as an integration across all the exemplars that are members of class k :

$$p(\mathbf{z}|\mathcal{C}^k) = \sum_{i=1}^{n_k} p(\mathbf{z}|\mathcal{C}_i^k, \mathcal{C}^k) p(\mathcal{C}_i^k|\mathcal{C}^k) \quad (2)$$

where \mathcal{C}^k is the class k , and \mathcal{C}_i^k is an exemplar of the class. Given $p(\mathcal{C}^k | \mathcal{C}_i^k) = 1$ (an assumption that none of the training data is mis-labelled) and $p(\mathcal{C}_i^k | \mathcal{C}^k) = 1/n_k$ (all exemplars within a class are equally likely), this becomes

$$p(\mathbf{z}|\mathcal{C}^k) = \frac{1}{n_k} \sum_{i=1}^{n_k} p(\mathbf{z}|\mathcal{C}_i^k) \quad (3)$$

The likelihood with respect to the exemplar can now be expanded as

$$\begin{aligned} p(\mathbf{z}|\mathcal{C}_i^k) &= p(z_1|z_2, \dots, z_n, \mathcal{C}_i^k) \\ &\times p(z_2|z_3, \dots, z_n, \mathcal{C}_i^k) \\ &\times \dots \times p(z_n|\mathcal{C}_i^k) \end{aligned} \quad (4)$$

This expression cannot be tractably computed — it is infeasible to learn the high-order conditional dependencies between appearance words. We thus seek to approximate this expression by a simplified form which can be tractably

computed and learned for available data. A popular choice in this situation is to make a naive Bayes assumption - treating all variables z as independent. However, visual words tend to be far from independent, and it has been shown in similar contexts that learning a better approximation to their true distribution substantially improves performance [9]. The learning scheme we employ is the Chow Liu tree, which locates a tree-structured Bayesian network that approximates the true distribution [10]. Chow Liu trees are optimal within the class of tree-structured approximations, in the sense that they minimise the KL divergence between the approximate and true distributions. Because the approximation is tree-structured, its evaluation involves only first-order conditionals, which can be reliably estimated from practical quantities of training data. Additionally, Chow Liu trees have a simple learning algorithm that consists of computing a maximum spanning tree over the graph of pairwise mutual information between variables — this readily scales to very large numbers of variables. The Chow Liu tree can be learnt from unlabeled training data across all classes, and approximates the distribution $p(z)$. To compute $p(z | C^k)$, the class-specific density, we find an expression that combines this global occurrence information with the class model outlined above. Returning to Equation 4 and employing the Chow Liu approximation yields

$$p(\mathbf{z}|C_i^k) \approx p(z_r|C_i^k) \prod_{q=1}^{|\mathbf{v}|} p(z_q|z_{p_q}, C_i^k) \quad (5)$$

where z_r is the root of the Chow Liu tree and z_{p_q} is the parent of z_q in the tree. Each term in Equation 5 can be further expanded as an integration over the state of the hidden variables in the exemplar appearance model. Assuming that sensor errors are independent of class and making the approximation $p(e_j|z_j)=p(e_j) \forall i \neq j$, each term in Equation 5 can be expressed entirely in terms of the known detector

model and marginal and conditional observation probabilities. These can be estimated from training data. Thus we have a procedure for computing $p(\mathbf{z}|C^k)$. Returning to Equation 1, the prior $p(C^k)$ can be learned simply from labelled training data, $p(\mathbf{z}|C^k)$ we have discussed above, and to normalise the distribution we make the naive assumption that our set of classes fully partitions the world. The posterior distribution across classes, $p(C^k|\mathbf{z})$, can now be computed for each patch.

Stage II: Context-Based Classification

The estimation of the set of most likely values of a set of interdependent random variables from available data is a standard machine learning problem. Such context-dependent inference can be achieved using a family of graphical models known as Markov Random Fields (MRFs). An MRF models the joint probability distribution, $p(\mathbf{x}, Z)$, over the (hidden) states of the random variables, \mathbf{x} and the available data, Z . For pairwise MRFs, it is well known that this joint probability can be maximised by equivalently minimising an energy function incorporating a unary term modelling the data likelihood for each node and a binary term specifying the interaction potentials between neighbouring nodes over the set of possible values [11]. Under the assumption of every datum being equally likely (i.e. $p(Z)$ being uniform) a minimisation of this energy function is equivalent to finding the most likely configuration of labels given the observed data - i.e. a maximum a priori (MAP) estimate of $p(\mathbf{x}|Z)$. In the following we describe how an MRF can be applied in the context of our scene labelling endeavour. In particular, we outline how the model structure of an MRF is derived for each scene from the available data, how the model parameters are obtained and, finally, how a MAP estimate over $p(\mathbf{x}|Z)$ is achieved.

MRFs are a family of graphical models where the set of interdependent variables is

modelled as a graph $G=(\mathcal{V}, \mathcal{E})$, where \mathcal{V} denotes the set of vertices and \mathcal{E} denotes the set of edges, respectively. In the context of our scene labelling problem, each vertex represents an image patch as produced by the first stage of our system. Neighbourhood relations within each scene are defined between patches sharing a common border, information provided readily by the image segmentation. Of course, adjacency in an image implies, but does not guarantee, adjacency in the 3D scene. Therefore, in estimating adjacency

from 2D information a trade-off is made between the ability of determining neighbourhood relations efficiently and the introduction of incorrect adjacencies due to the loss of depth information. In practice, we found the number of false adjacencies introduced by this approach to be negligible.

Typical examples of graph structure extracted from scenes recorded by our mobile platform are shown in Figure 1.



Figure 1: Typical graphs extracted from urban scenes as recorded by our mobile robot. Top: the original scenes. Bottom: the corresponding segmented images with the extracted graph overlaid. Circles indicate nodes, lines indicate edges. For images patches which are not marked as nodes no reliable geometry estimates could be extracted from the laser data. Reproduced from [1]

It should be noted that the one-to-one correspondence between vertices and image patches implies that the number of nodes in the MRF for a particular frame is independent of the number of measurements taken of the scene. Thus, the abstraction away from individual measurements (e.g. laser range data) to the patch level decouples the complexity of our inference stage from the density of the underlying data. This provides a substantial advantage in terms of speed over related works [4, [12] where the complexity of the

graphical models is directly proportional to the density of the underlying data.

The specification of an energy function to be optimised provides a convenient and intuitive way of incorporating scene properties. Consider the set of labels, $\mathbf{x} \in \mathbb{Z}^{N_n}$, for a particular configuration of a graph with N_n nodes. Each node s has an observation vector, \mathbf{z}^s , associated with it and can be assigned one of N_c labels such that $\mathbf{x}^s \in \{1, \dots, N_c\}$. We specify the energy of any such configuration to be given by

$$E(\mathbf{x}|\theta) = \sum_{s \in \mathcal{V}} \theta_s(x_s) + \sum_{(s,t) \in \mathcal{E}} \theta_{st}(x_s, x_t) \quad (6)$$

where we adopt the notation of [13] in that θ defines the parameters of the energy: $\theta_s(\bullet)$ is a unary data penalty function; and $\theta_{st}(\bullet)$ is a pairwise interaction potential. θ_s specifies the cost of assigning a given vertex any of the available labels. Intuitively, for a given node s , θ_s can be specified as a function of the posterior distribution over all classes for that node given the associated data, $p(\mathcal{C}|z^s)$, as provided by the patch classifier introduced in Stage I. In particular, the penalty of assigning label k to node s can be expressed as

$$\theta_s(x_{sk}) = 1 - p(\mathcal{C}^k | z^s) \quad (7)$$

The complement of $p(\mathcal{C}^k | z^s)$, is used since θ_s refers to a penalty function which is to be minimised. The pairwise potential θ_{st} encodes prior domain information in the form of penalties incurred by assigning specific labels to adjacent (i.e. connected) nodes. This is an intuitive formulation of the preference that nodes of certain labels are more likely to be connected to nodes of certain other labels. It follows that θ_s can be specified in terms of a square-symmetric matrix Φ of size $Nc \times Nc$ such that

$$\theta_{st}(x_i, x_j) = 1 - \phi_{i,j} \quad (8)$$

where again the complement is used since a penalty function is specified. However, rather than specifying a single matrix we specify two such matrices Φ_t and Φ_s , for the temporal and spatial edges respectively. For spatial edges we specify Φ_s such that, for two classes i and j ,

$$\phi_{i,j} = \frac{L_{i,j}}{L_i + L_j - L_{i,j}} \quad (9)$$

Here $L_{i,j}$ denotes the total number of links connecting nodes of labels i and j , and L_i denotes the total number of links

originating from nodes of label i . It follows that $\phi_{i,j} \leq 1 \quad \forall(i, j)$. Appropriate values for both $L_{i,j}$ and L_i are obtained from a hand-labelled training set. The temporal edges Φ_t are specified such that

$$\phi_{i,j} = 1, \quad \forall i \neq j, \quad (10)$$

$$\phi_{i,i} = 0, \quad (11)$$

thus enforcing a uniform penalty on all inconsistent temporal labels. MAP estimation is performed using sequential tree-reweighted message passing (TRW-S) [13] because of its desirable convergence properties and speed.

Results

The algorithm presented above was tested using two extensive outdoor data sets spanning nearly 17 km of track gathered with an ATRV mobile platform. The system was equipped with a colour camera mounted on a pan-tilt unit and a custom-made 3D laser scanner consisting of a standard 2D SICK laser range finder (75Hz, 180 range measurements per scan) mounted in a reciprocating cradle driven by a constant velocity motor.

The camera records images to the left, the right and the front of the robot in a pre-defined pan-cycle triggered by vehicle odometry at 1.5 m intervals. The Jericho data set was recorded in a built-up area in Oxford over 13.2 km of track (16,000 images in total). The Oxford Science Park data set was recorded in the science park area in Oxford over 3.3 km of track (8,536 images in total). The two datasets were collected in different areas of the city, with only a very small overlap between the two regions.

Due to space constraints only absolute performance across all classes is considered here. For a more complete analysis including common misclassifications (confusion) the reader is referred to [2].

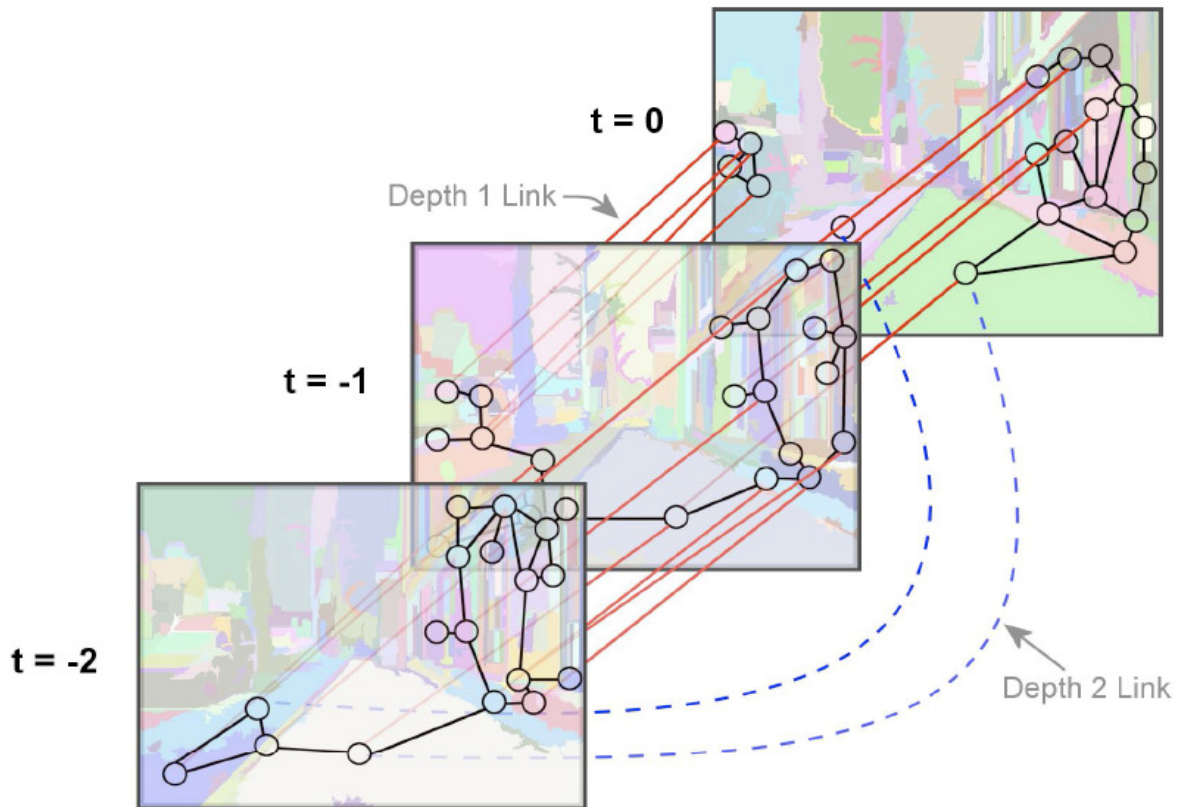


Figure 2: Conceptual illustration of the temporal MRF for three successive images. Spatial links are shown in black, depth one temporal links in red. Some depth two temporal links are also shown as blue dashed lines. Inference is carried out jointly over this spatio-temporal graph. Reproduced from [2]

Table 3: Detailed classification results for the Oxford Science Park data set. Results for the spatio-temporal column were obtained over a three-frame window with a history depth of one frame

Class Details		Pre MRF			Spatial Context			Spatio-Temporal Context		
Name	# Patches	Precision [%]	Recall [%]	$F_{0.5}$	Precision [%]	Recall [%]	$F_{0.5}$	Precision [%]	Recall [%]	$F_{0.5}$
Gr	82	80.3	69.5	77.9	89.2	40.2	71.7	95.5	25.6	61.8
Ta	1286	79.5	86.1	80.8	89.2	94.9	90.3	89.9	95.7	91.0
Di	127	21.4	47.2	24.0	60.2	46.5	56.8	75.6	46.5	67.2
Te	2199	73.3	75.5	73.8	74.0	93.8	77.3	74.3	97.5	78.0
Sm	898	54.1	36.2	49.2	76.4	39.0	64.1	86.3	40.7	70.5
Bu	175	41.7	38.9	41.1	59.6	35.4	52.5	61.5	32.0	52.0
Ve	165	35.9	34.6	35.6	69.1	33.9	57.3	79.7	30.9	60.6

Legend for class shortcuts: Grass, Tarmac/Paved, Dirt Path, Textured Wall, Smooth Wall, Bush/Foliage, Vehicle

Table 3 indicates that the patch classifier (pre-MRF) provides a baseline classification of mixed quality. Good results are achieved mainly for common classes (e.g. pavement / tarmac, textured wall and smooth wall, as well as for some less common ones (e.g. grass). The effect of both spatial and temporal context is pronounced. For common classes we note a boost to both precision and recall. For rarer classes such as vehicle and particularly grass the MRF has the effect of boosting precision at the cost of some drop in recall.

This tends to happen as weaker partial detections of objects are reassigned based on surrounding labels, typically eliminating many false positives but suppressing weaker true positives.

The benefits of MRF smoothing in terms of recall are more varied. For less common classes, a significant amount of over-smoothing occurs. However, the final results demonstrate generally good precision and reasonable recall

performance, particularly for the common classes.

The mean total processing time per frame was measured to be 4.0 seconds. For a more detailed breakdown of the processing time as well as a more detailed account of system performance the reader is referred to [2].

Conclusions

This paper has described and provided an overview of a two-stage approach to fast region labelling in maps of urban environments. Although the approach described here made specific use of both 3D laser and image data, the algorithms described are not limited to these modalities. The principal contribution of this work is the introduction of a layered classification framework which considers local scene properties in the first stage and then applies spatial as well as temporal context to refine these initial classifications. The results demonstrate the improvements in classification performance obtained by accounting for spatial and temporal context. This is despite the fact that the neighbourhood relations encoded in the MRF are a relatively weak cue; stronger information such as relative location and containment relations would be expected to improve the results. Furthermore, the inclusion and learning of a relative weighting between unary and binary potentials is expected to improve results since it provides a mechanism to minimise the over-smoothing effected by the MRF in the current system. Conceivably the most direct route to better performance is the addition of more informative features.

Further contributions are the development of efficient and principled methods to accomplish each classification stage. While any classifier capable of providing soft class assignments can conceivably be employed in the first stage of our framework, we opt to describe a probabilistic bag-of-words approach, which

has the advantage of providing a sensor model as a mechanism to incorporate the notion that some of the robot's observations are more trustworthy than others. In addition, the class models can readily be updated online. Furthermore, the formulation of the MRF model allows the efficient integration of contextual information. As a result, the overall per-scene compute time of this method is compelling: at 4.0 seconds it is suitable for online deployment.

References

- [1] I. Posner, M. Cummins, and P. Newman, *Fast probabilistic labeling of city maps*, in Proc. of Robotics: Science and Systems (RSS), June 2008.
- [2] I. Posner, M. Cummins, and P. Newman, *A generative framework for fast urban labeling using spatial and temporal context*, accepted for publication in Autonomous Robots.
- [3] G. Monteiro, C. Premebida, P. Peixoto, and U. Nunes, *Tracking and Classification of Dynamic Obstacles Using Laser Range Finder and Vision*, in Workshop on 'Safe Navigation in Open and Dynamic Environments - Autonomous Systems versus Driving Assistance Systems', in Proc. Intl. Conf. on Intelligent Robots and Systems (IROS), 2006.
- [4] B. Douillard, D. Fox, and F. Ramos, *A Spatio-Temporal Probabilistic Model for Multi-Sensor Multi-Class Object Recognition*, in Proc. 13th Intl. Symp. of Robotics Research (ISRR), 2007.
- [5] I. Posner, D. Schroeter, and P. Newman, *Describing Composite Urban Workspaces*, in Proc. of the Int. Conference on Robotics and Automation (ICRA), 2007.
- [6] I. Posner, D. Schroeter, and P. Newman, *Online generation of scene descriptions in urban environments*, Robot. Auton. Syst., 56(11):901–914, 2008.
- [7] J. Sivic and A. Zisserman, *Video Google: A text retrieval approach to object matching in videos*, in Proc. Intl. Conference on Computer Vision (ICCV), Nice, France, October 2003.
- [8] P. F. Felzenszwalb and D. P. Huttenlocher, *Efficient graph-based image segmentation*, Int. J. Comput. Vision, vol. 59, no. 2, pp. 167–181, 2004.

- [9] M. Cummins and P. Newman, *Probabilistic appearance based navigation and loop closing*, in Proc. Intl. Conference on Robotics and Automation (ICRA), Rome, April 2007.
- [10] C. Chow and C. Liu, *Approximating Discrete Probability Distributions with Dependence Trees*. IEEE Transactions on Information Theory, vol. IT-14, no. 3, May 1968.
- [11] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 6, no. 6, November 1984.
- [12] D. Angelo, B. Taskbar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Y. Ng, *Discriminative learning of Markov random fields for segmentation of 3D scan data*, in CVPR (2). IEEE Computer Society, 2005, pp. 169–176.
- [13] V. Kolmogorov, *Convergent tree-reweighted message passing for energy minimization*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, pp. 1568–1583, 2006.

Acknowledgements

The work reported in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence.