# Towards Simultaneous Place Classification and Object Detection based on Conditional Random Field with Multiple Cues

L. Shi, S. Kodagoda, *Member, IEEE*, and M. Piccardi, *Member, IEEE*

***Abstract*— Simultaneous place classification and object detection (SPCOD) is an algorithm which is able to categorize the environment (place) and detect the objects presented in the environment. Although both place classification and object detection problems have been in discussion in literature, as a concept SPCOD is still in its early stage of research. Focusing mainly on the discrimination ability of SPCOD, in this paper we have proposed a pairwise conditional random field (CRF) framework to integrate mature techniques on laser data based place classification and vision based off-the-shelf object descriptor. Extensive experimental results on a public data set demonstrate the effectiveness of the proposed method.**

## I. Introduction

The concept of simultaneous place classification and object detection (SPCOD) refers to the task of simultaneously distinguishing differences between various environmental locations and classifying part of the environment as representing certain objects. More importantly, the targets being processed should be assigned with human-defined labels (kitchen, office, chair, fridge, etc). The SPCOD has a variety of applications in robotics, such as human-robot interaction [1] and mapping [2].

In our perspective, SPCOD is a natural evolution of the current research on place classification which has been carried out based on various sensory modalities to a reasonable success. Vision based solutions for place classification have been in the forefront for many years [3][4][5], with a basic assumption that a realistic scene can be represented by a visual descriptor without any loss of discriminative information [6]. Pronobis et al. designed a vision-only recognition algorithm for place classification using rich global descriptors from image and support vector machine (SVM) as a discriminative classifier. Ranganathan and Jongwoo labeled the robot trajectories using CRF [4]. Vasudevan and Siegwart suggested functional concepts of places based on the objects and inter-object relationships [3]. In terms of 2D laser range sensor based solutions, Mozos et al. extracted hundreds of simple features and employed the AdaBoost classifier to label indoor environments consisting of rooms, corridors, doorways and halls [7]; Sousa et al. used a subset of the aforementioned features with SVM classifier [8]. In previous work, we were able to achieve an accuracy of 96% using i.i.d. classifiers on similar data sets for multiclass classification [2][9].

Research results show that graphical-models-based algorithms incorporating the contextual information usually exhibit superior performance than those relying only on the local observations [1]. Typical forms of this information are spatial/temporal consistency and place-object context. Therefore, combining well-developed object detection technique with current place classification methods are becoming a reality. In this respect, Murphy et al. used a tree-structured graphical model to facilitate object-presence detection, and vice versa [10]. Torralba et al. employed the semantic knowledge to provide contextual priors for object recognition [11]. More recently, the proposal of the concept simultaneous place and object recognition further reflected the need for modeling bidirectional interaction between places and objects for simultaneous reinforcement (which is somewhat analogous to simultaneous localization and mapping (SLAM)). Kim et al. extended the hidden Markov model (HMM) by incorporating the bidirectional context of objects, and Luo et al. proposed a hierarchical random field [12][13]. They all used low level features from images for object recognition, and modeling the coexistence of objects is reported to be computationally expensive.

As an effort towards developing fully functional SPCOD, we propose a supervised learning framework based on pairwise CRF to classify places and determine the presence/absence of target objects simultaneously. Experiments show that the proposed method is capable of producing appealing results.

The rest of this paper is arranged as follows. Section II describes the details of pairwise CRF and the methods proposed for SPCOD, as well as the standard SVM classification scheme which provides an alternative approach for comparison. Section III introduces the features extracted from multiple cues in two sensory modalities. Section IV describes the experimental setup. In Section V, experimental results are presented; and Section VI concludes the paper with pointers to future directions of the research.

## II. Background Algorithms and SPCOD methods

### A. Pairwise CRF and Parameter Estimation

Probabilistic graphical model like CRF captures both uncertainty and logical structure to compactly represent complex phenomena [14]. To be specific, CRF is a discriminative model which is used to directly estimate the conditional probability distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$, where $\boldsymbol{y}$ and $\boldsymbol{x}$ are labels (to be predicted) and observations respectively. In

Lei Shi and Sarath Kodagoda are with the Centre for Autonomous Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia.

Massimo Piccardi is with the School of Computing and Communications, Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia.

E-mails: {Lei.Shi-1 | Sarath.Kodagoda | Massimo.Piccardi}@uts.edu.au}

the work presented in this paper, an implementation of a CRF with pairwise potentials [15] is employed. The conditional distribution of pairwise CRF is defined in equation (1)

$$p(\boldsymbol{y}\mid\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x})}\prod_{<ij>}\psi_{ij}\left(y_i,y_j,\boldsymbol{x}\right)\prod_i\psi_i\left(y_i,\boldsymbol{x}\right) \tag{1}$$

where $\psi_i$ and $\psi_{ij}$ represent node and edge potentials respectively, and $Z(\boldsymbol{x})$ is the normalizing partition function. The node potential is a function of node features $\boldsymbol{x}_i$ and parameter matrix $\boldsymbol{w}$, and the edge potential is a function of edge features $\boldsymbol{x}_{ij}$ and parameter matrix $\boldsymbol{v}$. We choose $\boldsymbol{x}_i=[1,\boldsymbol{f}_i]$ and $\boldsymbol{x}_{ij}=[1,\boldsymbol{f}_i,\boldsymbol{f}_j]$ as the forms of the node and the edge features respectively, where $\boldsymbol{f}_i$ contains the local features associated with node $i$. As a common practice, the node features are derived from the local observation and the edge features share the node features from two end nodes of the edge [16]. In order to reduce the risk of over parameterization, the same set of parameter matrices are applied on all nodes and edges, and the node and edge potentials are set in the following forms.

$$\psi_i\left(\cdot,\boldsymbol{x}\right)=\left(e^{\boldsymbol{w}_1^T\boldsymbol{x}_i},\cdots,e^{\boldsymbol{w}_{n-1}^T\boldsymbol{x}_i},1\right) \tag{2}$$

$$\psi_{ij}\left(\cdot,\cdot,\boldsymbol{x}\right)=\begin{pmatrix} e^{\boldsymbol{v}_{11}^T\boldsymbol{x}_{ij}} & \cdots & e^{\boldsymbol{v}_{1,n-1}^T\boldsymbol{x}_{ij}} & e^{\boldsymbol{v}_{1,n}^T\boldsymbol{x}_{ij}} \\ \vdots & \ddots & \vdots & \vdots \\ e^{\boldsymbol{v}_{n-1,1}^T\boldsymbol{x}_{ij}} & \cdots & e^{\boldsymbol{v}_{n-1,n-1}^T\boldsymbol{x}_{ij}} & e^{\boldsymbol{v}_{n-1,n}^T\boldsymbol{x}_{ij}} \\ e^{\boldsymbol{v}_{n,1}^T\boldsymbol{x}_{ij}} & \cdots & e^{\boldsymbol{v}_{n,n-1}^T\boldsymbol{x}_{ij}} & 1 \end{pmatrix} \tag{3}$$

where $\boldsymbol{w}=[\boldsymbol{w}_1,...,\boldsymbol{w}_{n-1}]$, $\boldsymbol{v}=[\boldsymbol{v}_{11},...,\boldsymbol{v}_{n-1,n-1}]$, and $n$ is the number of target classes, i.e. $y_i\in\{1,\cdots,n\}$.

For the convenience of further analysis, equation (1) can be written in another way:

$$p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})=\exp\left(\left\langle\phi(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{\theta}\right\rangle-z(\boldsymbol{\theta}\mid\boldsymbol{x})\right) \tag{4}$$

where $z(\boldsymbol{\theta}\mid\boldsymbol{x})=\ln\sum_{\boldsymbol{y}}\exp\left(\left\langle\phi(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{\theta}\right\rangle\right)$, $\boldsymbol{\theta}=[\boldsymbol{w},\boldsymbol{v}]$,

and $\phi(\boldsymbol{x},\boldsymbol{y})$ is called sufficient statistics.

By applying clique decomposition, $\phi(\boldsymbol{x},\boldsymbol{y})$ can be calculated by summing the clique potentials over all nodes and edges [16].

$$\phi(\boldsymbol{x},\boldsymbol{y})=\left(\sum_{<ij>}\psi_{ij}\left(y_i,y_j,\boldsymbol{x}\right),\sum_i\psi_i\left(y_i,\boldsymbol{x}\right)\right) \tag{5}$$

In the parameter estimation stage, given a training set $\{X,Y\}=\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^m$ with $m$ instances, the distribution can be written as

$$P(Y\mid X;\boldsymbol{\theta})=\prod_{i=1}^m p(\boldsymbol{y}_i\mid\boldsymbol{x}_i;\boldsymbol{\theta}) \tag{6}$$

The maximum conditional likelihood estimation is adopted to estimate the parameter $\boldsymbol{\theta}$, i.e.

$$\boldsymbol{\theta}^*=\arg\max_{\boldsymbol{\theta}}P(Y\mid X;\boldsymbol{\theta}) \tag{7}$$

Equation (7) is equivalent to minimizing the negative log-likelihood $\boldsymbol{\theta}^*=\arg\min_{\boldsymbol{\theta}}nll(\boldsymbol{\theta})$. As per the common practice, an L2-regularization term $\lambda\|\boldsymbol{\theta}\|^2$ can be added to equation (6) to improve cross-validation results.

The negative log-likelihood function can be written as

$$nll(\boldsymbol{\theta})=-\sum_{i=1}^m\left[\left\langle\phi(\boldsymbol{x}_i,\boldsymbol{y}_j),\boldsymbol{\theta}\right\rangle-z(\boldsymbol{\theta}\mid\boldsymbol{x}_i)\right] \tag{8}$$

and the gradient of equation (8) is also required.

$$\nabla nll(\boldsymbol{\theta})=-\sum_{i=1}^m\left[\phi(\boldsymbol{x}_i,\boldsymbol{y}_i)-E_{p(\boldsymbol{y}\mid\boldsymbol{x}_i;\boldsymbol{\theta})}\left[\phi(\boldsymbol{x}_i,\boldsymbol{y})\right]\right] \tag{9}$$

where $E_{p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})}\left[\phi(\boldsymbol{x},\boldsymbol{y})\right]=\sum_{\boldsymbol{y}}p(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})\,\phi(\boldsymbol{x},\boldsymbol{y})$

The prediction process consists of calculating $\boldsymbol{y}^*=\arg\max_{\boldsymbol{y}}p\left(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta}^*\right)$ for any new observation $\boldsymbol{x}$.

In this paper the loopy belief propagation (LBP), which is a generalization of the forwards-backwards message passing algorithm to loopy graphs, has been adopted for parameter estimation and inference.

### *B. Support Vector Machine*

SVM is a well-established classifier which has been proved to offer excellent generalization ability. In this paper, the results from SVM are used for comparison purposes.

The basic idea of SVM is to map the data into a high dimensional feature space and find an optimal separating hyper-plane with the maximal margin.

Consider a training set of instance-label pairs $D=\{(\boldsymbol{x}_1,y_1),...,(\boldsymbol{x}_m,y_m)\}$, $\boldsymbol{x}_i\in R^n$ and $y_i\in\{1,-1\}$, $i=1,...,m$ where $m$ is the number of samples, instances $\boldsymbol{x}_i$ are firstly mapped into a higher dimension feature space $F$ via a nonlinear mapping $\phi: R^n\to F$.

Theoretically, a soft-margin SVM constructs an optimal hyper-plane $\boldsymbol{w}^T\boldsymbol{x}+b=0$ with maximum-margin and bounded error by solving,

$$\min_{\boldsymbol{w},b}\ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+C\sum_{i=1}^m\xi_i \tag{10}$$

$$s.t.\ \ y_i\left(\boldsymbol{w}^T\phi(\boldsymbol{x}_i)+b\right)\ge 1-\xi_i,\ \ \xi_i\ge 0,\ \ i=1,...,m \tag{11}$$

where $\boldsymbol{w}$ and $b$ denote the weight vector and bias respectively in the equation of the optimal hyper-plane $\boldsymbol{w}^T\boldsymbol{x}+b=0$. The positive constant $C$ is a penalty parameter used to control the amount of regularization, and the non-negative slack variable $\xi_i$ accounts for the amount of misclassification.

In this paper, the linear kernel C-Support Vector Classification scheme included in the LIBSVM package [17] is utilized as a multi-class classifier.

### C. *Candidate Methods for SPCOD*

In this research, three methods based on pairwise CRF are proposed and two of them are designed for SPCOD.

As shown in Fig. 1, single sensor modality place classification (S-PC) is similar to the VRF proposed by Friedman et al. for place classification [18]. It is a CRF built on top of spatially connected topological map. This method does not have object detection capability but it serves as the root of the other two methods. In addition, it provides useful reference results in experiments.

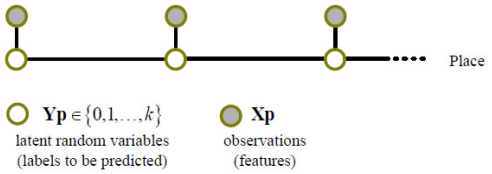


Fig. 1. Single sensory modality place classification model

The multiple cues SPCOD with individual object feature (M-SPCOD-IOF) shown in Fig. 2 is a derivation of S-PC by adding object nodes. The edges between place nodes and object nodes do not represent spatial connectivity any more but spatial co-existence. The co-existence of objects are not modeled otherwise the complexity of the model will increase dramatically with the categories of object nodes.

As an ideal setup, each object is described by only one feature which is the highest value in the Object Bank response map according to the specific object.

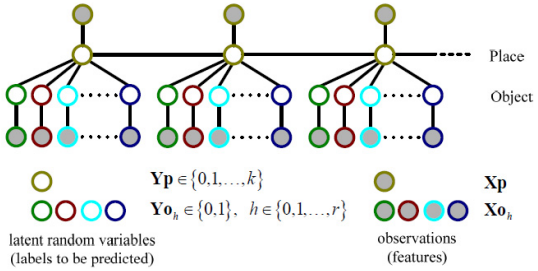


Fig. 2. Multiple cues SPCOD with individual object feature (the figure is better viewed in color)

The multiple cues SPCOD with object feature pool (M-SPCOD-OFP) shown in Fig. 3 further expands M-SPCOD-IOF by introducing an object feature pool (OFP) due to the consideration that an object may not be described well using corresponding object descriptor in the Object Bank, and it could also be represented better by other object descriptors [21]. A case in point is that a computer screen in an image may also respond well to "television filter" and "laptop filter", or even "window filter" due to the reflection. Since that the pairwise CRF deals with relatively high dimensional feature space well, introducing more features for each object is expected to provide better system performance. Therefore the OFP for a scene contains the peak values of responses from all available object filters (currently there are 208 object filters in Object Bank) which presumably represent the presence/absence of corresponding objects. It is the classifier's task to assign proper parameters to combine these features for each object node.

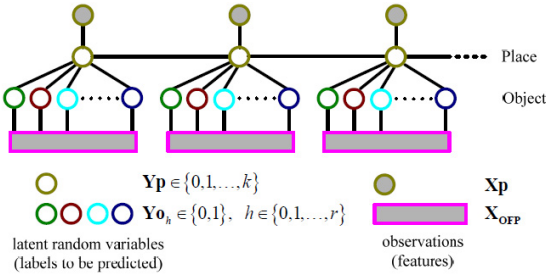


Fig. 3. Multiple cues SPCOD with object feature pool (the figure is better viewed in color)

It is to be noted that all the solutions proposed in this section are based on cyclic graphs, so that they are able to work on general graphs.

## III. Feature Construction

### A. *Features from 2D Laser Range Data*

The typical output of a 2D scanning laser range finder (LRF) is a beam sequence corresponding to a constant angle interval, which represents a point set in Euclidean Plane.

For the purpose of place classification, various features extracted from 2D laser range data have been reported in the literature including spectral features [19] and single-valued features capturing statistical and geometric information [7]. Mozos et al. suggested 22 categories of single-valued features and extracted 150 features considering different thresholds [7]. Sousa et al. constructed a feature set using 14 single-valued features [8].

For this research, the following 24 features including a subset of abovementioned simple features and those originally derived by Hjorth to describe time domain signal [20] have been adopted. Precise mathematical definitions of these features can be found in [7][20].

- The area $A$ of the polygon $Z$ specified by the observed points; the perimeter $P$ of $Z$; the normalized circularity of $Z$; the quotient of $A/P$
- The average and the standard deviation of both the beam length and the normalized beam length
- The average, normalized average and the standard deviation of the difference between the length of consecutive beams
- The average and the standard deviation of the relationship between the length of consecutive beams
- The average, the normalized average, the standard deviation and the normalized standard deviation of the distances between the centroid and the shape boundary
- The major axis *Ma* and minor axis *Mi* of the ellipse that approximates $Z$; the quotient of *Ma/Mi*
- The *Kurtosis*, *activity*, *mobility* and *complexity* of the beam length sequence [20]

### B. *Features from Image*

Both low-level and high-level features have been developed for the purpose of image understanding. Low-level features like GIST or SIFT-SPM are derived at the pixel level

to characterize images in a statistical way and high-level features like Object Bank usually analyses images in a semantically meaningful way [21][22]. As it is suggested that low-level image features are inadequate to capture complex semantic meaning required to solve high-level tasks [23], in this research the Object Bank is adopted as image-based object features because it serves our objectives well with extra convenience as off-the-shelf object filters.

Object Bank constructs feature vector through collecting responses to object filters which are trained classifiers in HOG feature space [24]. The responses are basically heat maps indicating the strength of the response when the object filter is placed at each position [24]. The benefits of using Object Bank include: a) it provides pre-trained off-the-shelf object filters; b) it holds useful information on the precise object position. However, Object Bank also has some problems: a) it does not hold any information on whether or not the object is on the scene; b) it does not guarantee adequate semantic coherence in terms of object detection [21][24]; c) its high performance on scene classification tasks is reported to be due to the high-dimensional vectors of local scale-space features [21].

In our methods, these problems are avoidable since 1) the decisions on the present/absence of objects are made by classifiers; 2) an object feature pool is introduced to use Object Bank features as latent information; and 3) only one feature corresponding to each object filter is used to construct a low-dimensional feature vector. An example of constructing the object feature pool for the M-SPCOD-OFP method mentioned in Section II.C is shown in Fig. 4. For each object filter, $8\times8$ response maps in 12 scales are generated and only the maximum strength is employed as the feature indicating the presence/absence of this object.

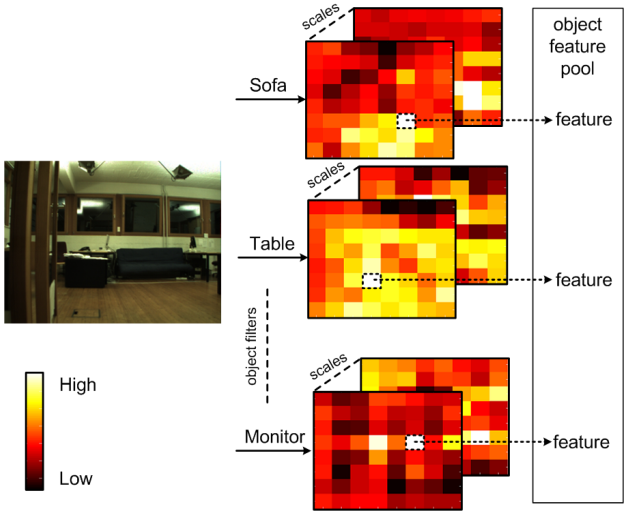


Fig. 4. Feature extraction process for M-SPCOD-OFP (the figure is better viewed in color)

## IV. Environment And Data Set

For the research reported in this paper, a subset of the freely available COsy Localization Database (COLD) [25] is adopted to validate the proposed methods. COLD contains 76 image sequences collected in three different indoor environments, using the same sensor setup in rooms of different functionality and under various environmental conditions [25]. The three data sets available are called COLD-Saarbruecken, COLD-Freiburg and COLD-Ljubljana respectively. However, the COLD-Saarbruecken data set does not contain sufficient images with high occurrence of objects, and the COLD-Ljubljana data set covers only one area for each environment type without providing any range data. Therefore, the following experiments are based on 16 sequences in the COLD-Freiburg data set as the environment consists of rich object density with the bonus of the available laser range finder data.

Based on the properties of the COLD-Freiburg data set, data collected from 3 trajectories containing all target places and objects were selected for training, and another 13 trajectories under different lighting conditions were divided into three groups for test. To clarify the terms, a *sample* is the data collected on a whole trajectory, and an *instance* is an observation (laser range finder data and image) made on a node (a pose) of a trajectory. To reduce the overhead of the system, we spatially down sample the observations to keep them about 0.5 meters apart on the trajectory.

In the test data, *Group 1* consists of 6 samples (shorter trajectories in the same area where the training data are gathered) under 3 lighting conditions (cloudy, night, and sunny). *Group 2* covers 3 samples (approximate revisits of the training trajectories) under 3 lighting conditions (cloudy, night, and sunny). *Group 3* contains 4 samples (trajectories in a new area which slightly overlaps with the training trajectories) under 2 lighting conditions (cloudy, sunny). From the perspective of training, data from *Group 1*, *Group 2* and *Group 3* can be thought as familiar data, approximately replicated data, and unfamiliar data respectively.

In the COLD-Freiburg data set, a single observation consists of a sequence of 180° LRF data collected by a SICK laser scanner and a $640\times480$ pixels color image came from a perspective camera. The corresponding pose of the robot which was estimated during the acquisition process using a laser-based localization technique was also provided [25].

As for the ground truth for training, place nodes are labeled by matching the blue-print map with poses where the observations are made, and the presence/absence of target objects in a scene is manually labeled. Please note that in this research, no effort has been made on providing ground truth about the precise object position in images. The target concepts include 6 place types and 10 objects.

## V. Experimental Results

In the following experiments, performances of different methods on place classification and object detection are discussed.

### A. Performances of Place Classification

In this experiment, places are classified into six categories (printer room, corridor, kitchen, office, bathroom and stairs) using different methods as describe in Section II.C. Overall place classification accuracies achieved are compared in TABLE I.

TABLE I
OVERALL PLACE CLASSIFICATION ACCURACIES

| | Group 1 (%) | Group 2 (%) | Group 3 (%) |
|---|---|---|---|
| SVM | 79.75 | 78.73 | 75.27 |
| S-PC | 90.93 | 94.93 | 79.81 |
| M-SPCOD-IOF | 93.31 | 92.51 | 83.73 |
| M-SPCOD-OFP | 96.28 | 94.17 | 81.98 |

Detailed testing results further shows that all three graph-model-based methods (S-PC, M-SPCOD-IOF and M-SPCOD-OFP) outperforms SVM on 13 out of 13 samples, highlighted the importance of modeling the contextual information of instances.

M-SPCOD-IOF and M-SPCOD-OFP outperform S-PC on 9 and 11 out of 13 samples respectively, which suggests the benefits of adding object information on improving the accuracies of place classification.

M-SPCOD-OFP outperforms M-SPCOD-IOF on 8 out of 13 samples, and shows slightly better accuracies than the latter according to TABLE I. Therefore in current setup the choice on the type of object feature does not seem to affect place classification accuracies significantly.

### *B. Performances of Object Detection*

In this experiment, single observation from the camera is categorized into presence or absence of the following 10 objects: chair, sofa, table/writing desk, computer monitor, printer, stairs/step, dish washer, fridge, closet/cupboard, and sink. Instead of using the overall accuracy as the judging criteria, specificity and sensitivity are adopted to describe the presence/absence of a particular object on the scenes. This is done to remove the effects of a biased prior distribution (i.e. there are much more *absence* than *presence* in the nature world). Definitions of specificity and sensitivity are shown in equations (12) and (13).

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}} \tag{12}$$

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{13}$$

As TABLE II shows quite close and high specificities reported by all methods, the attention should be focused more on sensitivities. According to TABLE III, M-SPCOD-IOF outperforms SVM (IOF) as the latter may not be able to work properly using only one feature. On the contrary, the performances of M-SPCOD-OFP and SVM (OFP) are very comparable (as indicated in both TABLE III and TABLE IV) with superior performances than M-SPCOD-IOF. This justifies the introduction of the object feature pool.

TABLE II
OBJECT CLASSIFICATION ACCURACIES EXPRESSED IN SPECIFICITY

| | Group 1 (%) | Group 2 (%) | Group 3 (%) |
|---|---|---|---|
| SVM (IOF) | 99.96 | 100.00 | 99.93 |
| M-SPCOD-IOF | 97.00 | 96.90 | 92.95 |
| | | | |
| SVM (OFP) | 97.12 | 97.24 | 92.53 |
| M-SPCOD-OFP | 97.47 | 98.06 | 94.72 |

TABLE III
OBJECT CLASSIFICATION ACCURACIES EXPRESSED IN SENSITIVITY

| | Group 1 (%) | Group 2 (%) | Group 3 (%) |
|---|---|---|---|
| SVM (IOF) | 0.69 | 1.64 | 1.73 |
| M-SPCOD-IOF | 41.87 | 42.97 | 27.53 |
| | | | |
| SVM (OFP) | 63.22 | 62.10 | 37.68 |
| M-SPCOD-OFP | 62.78 | 65.18 | 32.06 |

TABLE IV
OBJECT-SPECIFIC CLASSIFICATION ACCURACIES EXPRESSED IN SENSITIVITY

| | SVM (OFP) (%) | M-SPCOD-OFP (%) |
|---|---|---|
| Chair | 52.30 | 57.60 |
| Sofa | 58.33 | 44.64 |
| Table, desk | 64.84 | 63.25 |
| Monitor | 36.07 | 43.08 |
| Printer | 53.15 | 53.71 |
| Stairs | 74.36 | 35.13 |
| DishWasher | 54.08 | 61.90 |
| Fridge | 35.71 | 30.95 |
| Closet, cupboard | 28.58 | 48.97 |
| Sink | 86.00 | 88.84 |

With similar performances expressed in specificity and sensitivity, a direct comparison between the overall object classification accuracies of M-SPCOD-OFP and SVM (OFP) becomes meaningful.

However, although the former (95.43%) is slightly better than latter (94.36%), it does not strongly support the expectation that knowing the place will facilitate object detection. This is because of the facts that: a) SVM works well on this binary classification task with adequate features; b) the room to improve current relatively high accuracies is limited; c) extra training and test samples are required to provide more statistically convincing results.

### *C. Overall Performance of SPCOD*

The overall performance of SPCOD, comparing to the SVM (OFP) classification scheme, is shown in TABLE V. The overall classification accuracies (on all target concepts) suggest that the proposed two methods M-SPCOD-IOF and M-SPCOD-OFP are capable of classifying places and objects simultaneously at reasonably high accuracies. In addition, both of them outperform the highest accuracies achieved by SVM on the same data set.

TABLE V
SYSTEM PERFORMANCE EXPRESSED IN OVERALL ACCURACIES

| SVM (OFP) (%) | M-SPCOD-IOF (%) | M-SPCOD-OFP (%) |
|---|---|---|
| 92.88 | 93.29 | 95.06 |

### *D. Precise Object Position Localization*

The current focus of this research is to determine the presence or absence of a certain object, rather than geometrically localizing them. However, the Object Bank is able to provide extra information on the object location. Since this function is not the main focus of this paper, some typical results from object localization relying on the peak value of heat maps are presented in Fig. 5 without further quantitative analysis.


Fig. 5. Typical results from precise object position localization (the figure is better viewed in color)

## VI. Conclusions and Future Work

In this paper, we have proposed and implemented two pairwise CRF based SPCOD methods which were able to classify the robot's environment into 6 place categories and determine the presence of 10 objects in the scene. Simple statistical and geometrical features extracted from laser range data and Object Bank descriptors constructed from image were adopted. Experimental analysis was performed on publicly accessible data sets collected in an indoor environment with different environmental conditions.

Experimental results demonstrated the capabilities and potentials of the proposed methods on the SPCOD task, and they all outperformed standard SVM. The results also confirmed that the Object Bank features are suitable for the object detection tasks. Therefore we suggested using object feature pool rather than individual features for object detection.

The pairwise CRF based framework we proposed does not rely on particular sensory modality or feature set, so that it can be easily extended to have any arbitrary number of object nodes. In the future work, we plan to further improve the object detection ability of the system, incorporate temporal information and move on to online testing.

## References

[1] O. M. Mozos, P. Jensfelt, H. Zender, G. J. M. Kruijff, and W. Burgard, "From labels to semantics: An integrated system for conceptual spatial representations of indoor environments for mobile robots," in *IEEE Int. Conf. Robotics and Automation workshop*, Rome, 2007, pp. 33-40.

[2] L. Shi, S. Kodagoda, and G. Dissanayake, "Multi-class classification for semantic labeling of places," in *Proc. Int. Conf. Control Automation Robotics & Vision*, Singapore, 2010, pp. 2307-2312.

[3] S. Vasudevan and R. Siegwart, "Bayesian space conceptualization and place classification for semantic maps in mobile robotics," *Robotics and Autonomous Systems*, vol. 56, pp. 522-537, 2008.

[4] A. Ranganathan and L. Jongwoo, "Visual place categorization in maps," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2011, pp. 3982-3989.

[5] J. Wu, H. I. Christensen, and J. M. Rehg, "Visual Place Categorization: Problem, dataset, and algorithm," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2009, pp. 4763-4770.

[6] A. Pronobis, L. Jie, and B. Caputo, "The more you learn, the less you store: Memory-controlled incremental SVM for visual place recognition," *Image and Vision Computing*, vol. 28, pp. 1080-1097, 2010.

[7] O. M. Mozos, C. Stachniss, and W. Burgard, "Supervised Learning of Places from Range Data using AdaBoost," in *Proc. IEEE Int. Conf. Robotics and Automation*, Barcelona, 2005, pp. 1730-1735.

[8] P. Sousa, R. Araiijo, and U. Nunes, "Real-Time Labeling of Places using Support Vector Machines," in *Proc. IEEE Int. Symp. Industrial Electronics*, Vigo, 2007, pp. 2022-2027.

[9] L. Shi, S. Kodagoda, and G. Dissanayake, "Laser range data based semantic labeling of places," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Taipei, 2010, pp. 5941-5946.

[10] K. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: a graphical model relating features, objects and scenes," *Advances in Neural Information Processing Systems*, vol. 16, pp. 1499-1506, 2003.

[11] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, "Context-based vision system for place and object recognition," in *Proc. IEEE Int. Conf. Computer Vision,* 2003, pp. 273-280.

[12] L. Ronghua, P. Songhao, and M. Huaqing, "Simultaneous Place and Object Recognition with Mobile Robot Using Pose Encoded Contextual Information," in *Proc. IEEE Int. Conf. Robotics and Automation*, Shanghai, 2011, pp. 2792-2797.

[13] S. Kim and I. S. Kweon, "Simultaneous place and object recognition using collaborative context information," *Image and Vision Computing*, vol. 27, pp. 824-833, 2009.

[14] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning," *Introduction to statistical relational learning*, p. 93, 2007.

[15] M. Schmidt, K. Murphy, G. Fung, and R. Rosales, "Structure learning in random fields for heart motion abnormality detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Anchorage, 2008.

[16] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy, "Accelerated training of conditional random fields with stochastic gradient methods," in *Proc. Int. Conf. Machine Learning*, Pittsburgh, 2006, pp. 969-976.

[17] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 1-27, 2011.

[18] S. Friedman, H. Pasula, and D. Fox, "Voronoi random fields: Extracting the topological structure of indoor environments via place labeling," in *Proc. Int. Joint Conf. Artificial Intelligence*, Hyderabad, 2007, pp. 2109-2114.

[19] A. Poncela, C. Urdiales, B. Fernandez-Espejo, and F. Sandoval, "Place characterization for navigation via behaviour merging for an autonomous mobile robot," in *Proc. IEEE Mediterranean Electrotechnical Conf.*, Ajaccio, 2008, pp. 350-355.

[20] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, pp. 306-310, 1970.

[21] D. Leung, S. Newsam, "Can off-the-shelf object detectors be used to extract geographic information from geo-referenced social multimedia?", in *ACM SIGSPATIAL Int. workshop Location-Based Social Networks*, Redondo Beach, pp. 12-15, 2012.

[22] L.J. Li, H. Su, E.P. Xing, L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification", *Advances in Neural Information Processing Systems*, vol. 23, pp. 1378-1386, 2010.

[23] T. Althoff, H.O. Song, T. Darrell, "Detection Bank: An Object Detection Based Video Representation for Multimedia Event Recognition", in *Proc. the ACM Multimedia Conf.*, Nara, 2012.

[24] A.F. Araujo, P. Weinzaepfel, P. Perez, C. Diot, *'object bank'-based scene classification*, Technicolor tech. report, Stanford Univ.

[25] A. Pronobis and B. Caputo. "COLD: COsy Localization Database". *The Int. Journal of Robotics Research*, vol. 28(5), pp.588-594, 2009.