Joint Object Classification in 3D Indoor Scenes using Spatial Features and Spatial Relations

Abstract

This paper presents a novel approach for joint object classification that converts the scene into a graph, where each node corresponds to an object with assigned a possible category label. The assignment problem, i.e. to find the correct object category label for each unknown object, is framed as a maximum score clique search in this graph. The clique score is computed using a voting scheme, which considers spatial relation based features of the object pairs and spatial characteristics of the individual objects. The typical spatial properties and the pairwise spatial relations of the objects are learned in a training phase using a Gaussian Mixture Model representation, along with the object occurrence and co-occurrence probability. Experiments on two real world 3D desk scene datasets show the discriminative power of our approach among object class categories.

1 Introduction

Automatic object classification plays a central role in numerous robotics applications such as surveillance, service robotics, object search and retrieval, human care and object manipulation. Object classification is still a challenging problem in Computer Vision, as the vision-based object classifiers are highly dependent on object pose, colour, texture, camera viewpoint and illumination (Felzenszwalb et al. 2010). Another limitation of traditional core Computer Vision based approaches to this problem is that the performance decreases with increasing number of object categories. Despite the huge variety in object categories, shapes, poses, texture, etc., indoor environments often exhibit a coherent composition of objects, in terms of relative spatial arrangement and co-occurrence frequency (Kasper, Jakel, and Dillmann 2011). This type of knowledge can be a strong cue for collectively determining object labels in a scene.

This paper presents a novel approach to joint object classification in indoor environments by effectively modeling these coherent inter-relationships in the object arrangement observed in the indoor scenes. An unknown test scene, consisting of a set of objects, each of which can correspond to a set of possible object categories, is transformed to a graph, a natural data structure to model relationships. Each node represents a different assignment of an object instance to a category and the edges represent pairwise relations between objects having assigned a possible category label. Only the nodes corresponding to different object instances are adjacent. A clique of this graph corresponds to the assignment of category labels to a set of test objects.

The joint object classification task is addressed by finding the clique which includes all test objects and maximizes a clique score. The clique score is computed using a voting strategy, where each pair of nodes in the clique has a contribution. In the training phase, a set of Gaussian Mixture Models (GMMs) are learned to represent the object categories and the relations between pairs of object categories, and encode the coherence of the inter-object arrangement in the observed scenes, as proposed in (Alberti, Folkesson, and Jensfelt 2014). In the inference phase, the likelihood values of the learned models are used in the clique score computation, along with the information about characteristic object occurrence and co-occurrence probabilities.

The method is tested on two datasets of 3D office desk scenes acquired with an RGB-Depth sensor. The scene information is represented as a 3D point cloud and the objects are segmented by manual annotation, using 3D bounding boxes. A leave-one-out cross-validation framework is applied to evaluate the joint object classification performance.

Note that the proposed framework could be integrated with vision based approaches to object classification, to complement them and enhance their performance. To achieve a reasonable performance, a vision based approach to this problem typically needs a large dataset (Zhu et al. 2012) compared to our 3D scene datasets. For this reason, instead of testing our method as a complement to a vision based classification, we investigate the relevance of different features in the experimental results section and show how relational features can aid classification when compared to spatial features of the individual objects (features not based on relations). It is reasonable to believe that we would see a similar complementary effect on a vision based classifier.

The remainder of this paper is organized as follows. A survey of the related work is presented in Section 2. The proposed method for joint object classification is described in Section 3. The experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2 Related work

Joint object classification in Robotics focuses more on the relational features which capture the inter-relationships and the coherent composition of objects in indoor scenes (Southey and Little 2007; Kasper, Jakel, and Dillmann 2011; Choi et al. 2013; Lin, Fidler, and Urtasun 2013), than the core Computer Vision features, such as SIFT, HOG, SURF and PIRFS, commonly applied to general object classification tasks (Lowe 1999; Dalal and Triggs 2005; Bay et al. 2008; Kawewong, Tangruamsub, and Hasegawa 2010).

One important limitation of many core Computer Vision based object classification methods is their dependence on colour, texture, appearance and pose, which makes them vulnerable on seemingly simple examples, when the objects lack sufficiently distinctive appearance data. To overcome this limitation, it is important to introduce new independent and complementary features that explore the coherent arrangement of objects in indoor environments.

Recent studies (Southey and Little 2007; Kasper, Jakel, and Dillmann 2011) compute the distribution of 3D spatial relations of objects over a set of scenes, and show how the obtained data and models can be used in a typical scenario for service robotics such as the categorization of individual objects given the knowledge of the category labels of other objects in the scene. Southey et al. (Southey and Little 2007) learn a maximum entropy model of 3D spatial relations between objects from artificial indoor scenes of a video game and test the model in an object recognition task. In the work of Kasper et al. (Kasper, Jakel, and Dillmann 2011), an empirical base for scene understanding is developed, by encoding the structure of the scene in the spatial relations between the objects. Spatial 3D features and spatial relations between pairs of objects are also used in several other robotics studies, in the context of navigation, planning, object position prediction and manipulation (Rosman and Ramamoorthy 2011; Ye and Hua 2013; Burbridge and Dearden 2012; Aydemir et al. 2011). The method that we present in this paper builds on previous work on the use of 3D spatial relations for object classification and scene similarity measurement in indoor scenes (Alberti, Folkesson, and Jensfelt 2014), and extends the proposed concept by using pairwise object features, along with features from the individual objects, for object classification and by proposing a novel graph search based method.

To represent the spatial and semantic coherence of regions in a scene, graphical models are often used in the literature (Boutell, Luo, and Brown 2006; Galleguillos, Rabinovich, and Belongie 2008; Yao, Fidler, and Urtasun 2012; Lin, Fidler, and Urtasun 2013). Boutell et al. (Boutell, Luo, and Brown 2006) develop a generative model of 2D outdoor scenes based on characteristic objects in the scene and spatial relationships between them. Scene probabilities are estimated using loopy belief propagation on a factor graph. Galleguillos et al. (Galleguillos, Rabinovich, and Belongie 2008) propose a method of object categorization for outdoor 2D images that incorporates both semantic context (co-occurrence) and spatial context into a unified framework. The approach uses a conditional random field (CRF) formulation to maximize contextual constraints over the object labels. The method proposed by Yao et al. (Yao, Fidler, and Urtasun 2012) reasons jointly about regions, location, class and spatial extent of objects, presence of a class in the image, as well as the scene type. The holistic problem is framed as a structure prediction problem in a graphical model. In the work of Lin et al. (Lin, Fidler, and Urtasun 2013), a holistic approach exploits 2D segmentation, 3D geometry, and contextual relations between scenes and objects, using a CRF.

The contributions of this paper are as follows. We propose a method to solve the joint object classification problem, based on the graph representation of a scene which reflects the spatial and semantic relations among objects in a 3D scene. We develop an optimization technique to search for the maximum score clique in the graph, which denotes the best category assignment for each object in the scene. The clique score computation is based on a voting scheme, which takes into account the coherence of the objects arrangement in the scene using spatial relation based features. Finally, in the experiments we investigate the feature influence and show the impact of the relational features on the object classification performance.

3 Proposed method

We assume that each example scene S of the same type (e.g. office desk) contains a set of objects, $O = \{o_1, o_2, \ldots, o_m\}$ (where m can vary among scenes), of different shapes and sizes and corresponding to various possible object categories. For example, in an office desk scene, possible object classes can be monitor, keyboard, mouse, books, mugs, etc. We also assume that a finite set of possible categories, $C = \{c_1, c_2, \ldots, c_k\}$, can be identified in the scene. The problem of assigning a category label to each unknown object instance is an NP problem, having k^m possible solutions. In the proposed method each of these solutions is represented as a clique of a graph and the optimal solution, i.e. the correct assignment of a category label to each of the unknown objects, is identified by finding the maximum score clique using an optimization technique. An overview of the method is given in Figure 1.

3.1 Scene to graph conversion

The approach is based on the representation of the scene $S = (O, C)$ as an undirected graph, $G = (V, E)$. Each node of the graph, $v \in V$, consists of a pair of random variables, *object instance* and *category label*, $v = (o_i, c_p)$, $i \in 1, 2, \ldots, m$, $p \in 1, 2, \ldots, k$, and represents an object instance with an assigned category label. The graph has the property that nodes having the same value of o_i cannot be adjacent. The edges of the graph, $e \in E$, represent pairwise relations between objects with assigned categories, $e = \{(o_i, c_p), (o_j, c_q)\}$.

In this representation, the solution of the joint object classification problem, i.e. the global scene with the assigned object labels, is a clique of the graph G (see Figure 2), defined as a subset of vertices of G such that every two vertices are connected by an edge. Note that the clique cannot contain more than one vertix corresponding to the same object instance. The problem of assigning the correct category label

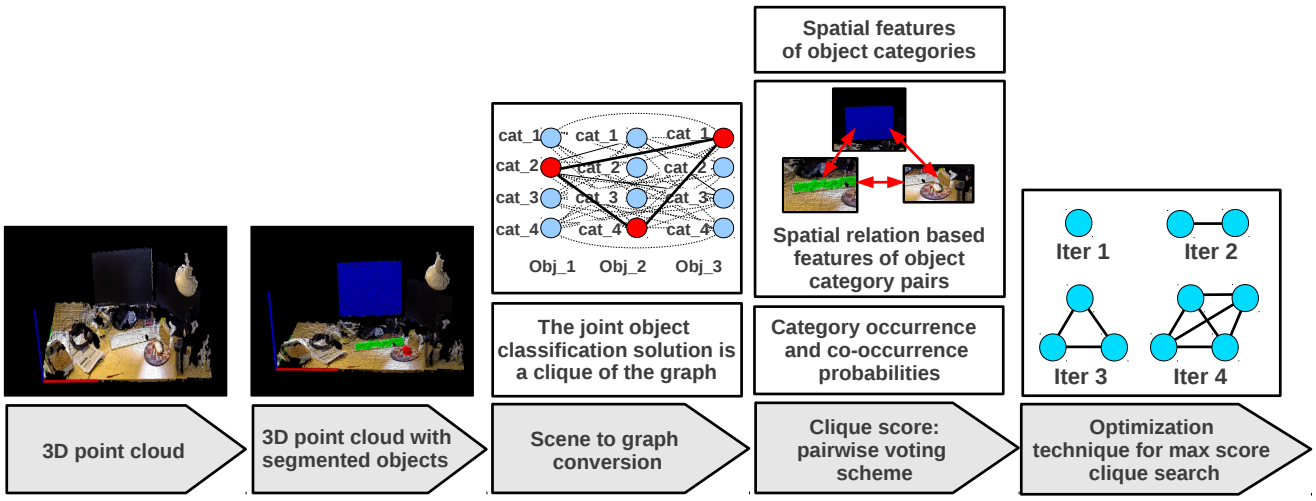

Figure 1: Overview of the proposed joint object classification method. The indoor scene information is represented as a 3D point cloud. Different objects are segmented from the point cloud. The set of objects is represented as a graph and the global assignment of object categories to each of the objects in the scene is a clique in this graph. The optimal clique is identified by applying an optimization search algorithm to find the maximum score clique. The clique score is computed based on a voting scheme, which takes into account spatial characteristics, pairwise spatial features of the objects, probabilities of occurrence and co-occurrence of the object categories.

to each object in the test scene, choosing from a predefined set of possible object categories, is addressed as maximum score clique search.

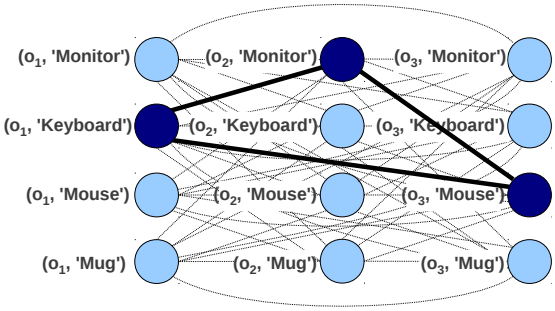

Figure 2: An example of a clique identified in the graph representation of the scene, where each of three objects, o_1, o_2 and o_3, is assigned to a category, chosen among a set of four possible categories: 'Monitor', 'Keyboard', 'Mouse' and 'Mug'.

3.2 The clique score

A clique $\Psi = (v_1, v_2, \ldots, v_n)$ having n vertices represents the assignment of n objects to n category labels. The total clique score, $Clique_S(\Psi)$, is computed based on a voting scheme where each pair of adjacent vertices of the clique contributes to the total score:

$$Clique_S(\Psi) = \sum_{\substack{s \neq t \\ s,t \in \{1,\ldots,n\}}} Pair_S(v_s, v_t) = \sum_{\substack{i \neq j \\ i,j \in \{1,\ldots,m\} \\ p,q \in \{1,\ldots,k\}}} Pair_S((o_i, c_p), (o_j, c_q)), \quad (1)$$

where $Pair_S(v_s, v_t)$ is the *pairwise score* of the adjacent vertices (v_s, v_t); $v_s = (o_i, c_p)$ and $v_t = (o_j, c_q)$. The *pairwise score* considers spatial characteristics of the individual objects, spatial relations between the object pairs, and object classes occurrence and co-occurrence probabilities.

The reason behind the proposed voting scheme is that in the training phase, the method learns the typical arrangement of the object categories, and in the inference phase, the likelihood estimate for each pair is reflected in the overall likelihood of the global assignment (clique). Other solutions for the computation of the clique score, such as a product of pairwise scores, would be too sensitive, for example in scenes containing object categories that appear only in few examples in training, while the sum of pairwise scores is a more robust solution. The following paragraphs explain in detail the quantities used to compute the *pairwise scores*.

Spatial features and spatial relation based features To model the object categories and the relationships between pairs of categories, feature sets are obtained to capture the object geometry and the spatial distribution of objects in the scene (Alberti, Folkesson, and Jensfelt 2014). *Single object features* f_{o_i} are computed from the 3D spatial characteristics of the object, both individually and w.r.t. a reference frame (here the table, with the front-left table corner used as

the origin of an extrinsic reference system aligned with the two horizontal table axes). The set consists of: volume (vol), length of the object projection along the X, Y and Z axes (l_x, l_y and l_z), 3D coordinates of the object centroid in the reference system defined by the table (p_x, p_y and p_z) and horizontal bearing of object centroid from front-left table corner (θ). It is worth noticing that the volume and the length of the object projection along the vertical dimension can be considered as spatial characteristics of the object alone (non-relational features) while the remaining features depend on the table reference system.

To represent the pairwise spatial distribution of the objects, we introduce the feature set f_{o_i,o_j} as *object pair features*, keeping the same extrinsic reference system. The feature set consists of: (1) Euclidean distance between object centroids, (2) Euclidean distance between centroids in the X-Y plane, (3) bearing between the two object centroids computed in the reference system defined by the table, (4) ratio of object volumes and (5) vertical displacement between object centroids.

Learning spatial models of object categories and category pair relations In a training phase, a set of models for each of the object class categories, $c_p \in C$, are learned by using a GMM based representation to encode the multivariate probability distribution of *single object features*. In a similar fashion, the probability distribution of *object pair features* for the different category pairs, f_{c_p,c_q}, are modeled in a multi-dimensional feature space by applying a GMM on the *object pair feature* set (Alberti, Folkesson, and Jensfelt 2014).

Pairwise score computation The *pairwise score*, is computed as the product of the two scores obtained for the assignments of the two individual objects, $o_i : c_p$ and $o_j : c_q$, and the score obtained for the assignment of the pair, $(o_i, o_j) : (c_p, c_q)$, as follows:

$$Pair_S((o_i,c_p),(o_j,c_q)) = Score_{SO}(o_i,c_p)\cdot Score_{SO}(o_j,c_q)\cdot Score_{OP}((o_i,c_p),(o_j,c_q)). \quad (2)$$

These scores, $Score_{SO}$ and $Score_{OP}$, are defined as:

$$Score_{SO}(o_i,c_p) = Pr(f_{o_i}|c_p)\cdot Occr(c_p), \quad (3)$$

$$Score_{OP}((o_i,c_p),(o_j,c_q)) = Pr(f_{o_i,o_j}|c_p,c_q)\cdot Coccr(c_p,c_q), \quad (4)$$

where $Pr(f_{o_i}|c_p)$ is the the likelihood value of the category model given the *single object features*, f_{o_i}, and $Pr(f_{o_i,o_j}|c_p,c_q)$ is the likelihood value of the category pair model given the *object pair features*, f_{o_i,o_j}. The likelihood values correspond to the conditional probability of the computed features given the parameter values of the models. Additionally, the scores integrate, as a-priori weights, the occurrence probability of the individual object categories, $Occr(c_p)$, and the co-occurrence probability of the object category pairs, $Coccr(c_p,c_q)$, which are defined:

$$Occr(c_p) = \frac{max(1,N_{c_p})}{(1+N_{tot})}, \quad (5)$$

$$Coccr(c_p,c_q) = \frac{max(1,N_{c_p,c_q})}{(1+N_{tot})}, \quad (6)$$

where N_{c_p} is the number of training scenes containing c_p, N_{c_p,c_q} is the number of scenes containing both c_p and c_q are present and N_{tot} is the total number of training scenes. The numerator and denominator terms, $max(1,N_{c_p})$, $max(1,N_{c_p,c_q})$ and $(1+N_{tot})$, ensure that occurrence and co-occurrence weights are never 0 or 1 (to avoid borderline cases).

3.3 Optimization technique

A greedy optimization search algorithm is applied to find the maximum score clique, as an alternative to exhaustive search which is not a tractable option in a realistic scenario due to its high computational cost. The search is initialized by selecting the starting node $v^* = (o_{i^*}, c_{p^*})$ with the maximum individual object score:

$$(i^*,p^*) = argmax_{(i,p)} Score_{SO}(o_i,c_p). \quad (7)$$

Then, at each iteration, a node, $v_s = (o_i, c_p)$, is added, by selecting, among all the not-traversed object instances, the node which maximizes the clique score for the obtained clique (Equation 1). A scheme that exemplifies a series of successive iterations is shown in Figure 3.

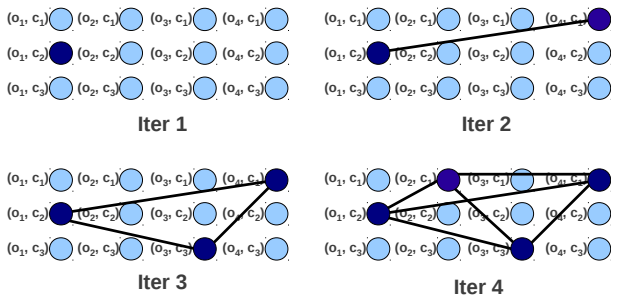

Figure 3: Representation of four successive iterations of the greedy optimization algorithm showing the clique identified at each iteration. Note that all the other edges are not shown.

4 Experimental results

We apply the proposed approach on two datasets of desk scenes where different sets of object categories are manually annotated with their bounding boxes and category labels. Performance is evaluated by using leave-one-out cross-validation and by performing tests with different subsets of features. Furthermore, the computational time of the proposed optimization technique is presented and compared to the exhaustive search solution.

4.1 Database description

Our scene database consists of two datasets of 3D office desk scenes. The first dataset (Dataset A) contains 42 scenes, acquired from 6 office desks over a period of days, where 6 object categories are present. The second dataset (Dataset B), acquired and labeled in a successive moment, consists of 291 scenes from 20 different desks, and contains 14 object categories. On an average, there are 5.2 objects per scene in Dataset A, and 6.9 objects per scene in Dataset B. Due to the higher number of training samples, Dataset B allows a for better learning of the models. The data are acquired using an RGB-D sensor, the Asus Xtion Pro Live sensor, and stored as point clouds. Both datasets are manually annotated by labeling the desk and the objects on the desk using 3D cuboid bounding boxes, and the features are computed from the object bounding boxes.

4.2 Object classification performance

Leave-one-out cross validation experiments are performed to assess the joint object classification performance. Each cross validation fold contains all the acquisitions of the same desk, taken at different time moments, so there are 6 folds in Dataset A and 20 folds in Dataset B. In the experiments, n_c = 2 mixture components are used for the GMM of each object class and category pair.

Joint object classification in the two datasets In a first experiment, we analyze the joint object classification results on the two datasets. Tables 1 and 2 show the precision, recall and F-Measure[1] (Rijsbergen 1979) computed as an average over the folds for each of the object categories, in Datasets A and B, respectively, and the number of examples of each category present in the datasets. In Table 1, we observe that 100% precision and recall scores are obtained in Dataset A for the categories 'Keyboard', 'Mouse', 'Mug' and 'Pen/Pencil', which can be explained with the simplified scenario of a low number of object categories. The object category 'Lamp' is sometimes misclassified as 'Monitor' mainly due to the similar volume of the annotated bounding boxes and to the limited number of examples. In Dataset B (see Table 2), a higher performance can be noticed for the category set: 'Monitor', 'Keyboard', 'Mouse', 'Mug', 'Lamp', 'Laptop', 'Papers' and 'Bottle', compared to the category set: 'Notebook', 'Book', 'Mobile', 'Glass', 'Jug' and 'Headphones'. To observe the inter-category confusion, we present the confusion matrix for Dataset B in Table 3. The major confusion occurs between 'Notebook' and 'Papers', 'Book' and 'Papers', 'Headphones' and 'Book', 'Mobile' and 'Mouse', 'Glass' and 'Mug', mainly due to their similar size and their similar position on the table.

Baseline system with context independent features vs. Complete system with all features In a second experiment we investigate the feature influence by training and testing the models on different subsets of features, using Dataset B. The goal of this experiment is to show how the proposed relational features can complement context-independent spatial features of the individual objects and improve the performance. First, we test a baseline system where only the object volume and the projection of the object bounding box along the vertical axis are used as *single object features*: $f^1_{o_i} = (vol, l_z)$, and no *object pair feature* is used, thus setting in Equation 2: $Score_{OP}((o_i, c_p), (o_j, c_q)) = 1, i, j \in 1, 2, \ldots, m$. These features, $f^1_{o_i}$, do not require the definition of the extrinsic reference frame determined by the table, so they can be considered as an example of context independent features. Second, the complete system with the feature set: $f^2_{o_i} = (vol, l_x, l_y, l_z, p_x, p_y, p_z, \theta)$ and the *object pair features* is compared to the first system. Third, only relational *single object features*: $f^3_{o_i} = (l_x, l_y, p_x, p_y, p_z, \theta)$ and *object pair features* are used. Table 4 shows the joint object classification F-Measure of these three cases. It can be observed that when the relational features are included in the complete framework (column [b]), there is a significant performance improvement w.r.t. the use of the sole context independent features (column [a]) for the categories: 'Keyboard', 'Lamp', 'Papers', 'Book', 'Glass' and 'Headphones'. The context-independent spatial features yield higher F-Measure for 'Jug' and 'Bottle', while for the classes: 'Monitor',

Table 1: Average joint object classification precision, recall and F-Measure (%) and the number of object examples in Dataset A, for each of the considered object categories.

	precision	recall	F-Measure	#Obj.
Monitor	83.05	85.96	84.48	57
Keyboard	100	100	100	42
Mouse	100	100	100	37
Mug	100	100	100	15
Lamp	66.66	61.53	64	26
Pen/pencil	100	100	100	43

Table 2: Average joint object classification precision, recall and F-Measure (%) and the number of object examples in Dataset B, for each of the considered object categories.

	precision	recall	F-Measure	#Obj.
Monitor	94.45	99.1	96.72	447
Keyboard	93.15	95.1	94.11	286
Mouse	90.37	90.68	90.53	290
Mug	86.25	87.89	87.06	157
Lamp	94.28	84.61	89.19	39
Notebook	3.7	7.69	5	13
Laptop	86.04	61.15	71.49	121
Papers	75	78.21	76.57	303
Book	44.685	39.25	41.79	107
Mobile	39.39	68.42	50	19
Glass	52.38	37.93	44	29
Jug	14.28	8.33	10.52	24
Headphones	27.77	29.41	28.57	17
Bottle	95.52	90.14	92.75	71

[1]The F-Measure is defined as the harmonic mean of precision and recall: $F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$.

Table 3: Confusion matrix in Dataset B. The rows represent the actual class and the columns represent the predicted class.

	Monitor	Keyboard	Mouse	Mug	Lamp	N.book	Laptop	Papers	Book	Mobile	Glass	Jug	H.phn	Bottle
Monitor	99.1	0.44	0	0	0.44	0	0	0	0	0	0	0	0	0
Keyboard	0.69	95.1	1.39	0	0	0.34	0	2.09	0	0	0	0.34	0	0
Mouse	0.68	0.34	90.68	0.68	0	0	0	0.34	0	6.89	0	0.34	0	0
Mug	0	0	1.91	87.89	0	0	0	0	1.91	0	2.54	3.81	2.54	0
Lamp	15.38	0	0	0	84.61	0	0	0	0	0	0	0	0	0
N.book	0	7.69	0	0	0	7.69	0	67.92	7.69	0	0	0	0	0
Laptop	11.57	9.91	0	0	0	0	61.15	12.39	4.91	0	0	0	0	0
Papers	0	0.66	1.32	0	0	6.93	2.08	78.21	9.9	0	0	0	0.33	0
Book	1.86	1.86	3.73	0	0	3.73	3.73	42.99	39.25	0	0	0	2.8	0
Mobile	0	0	31.57	0	0	0	0	0	0	68.42	0	0	0	0
Glass	0	0	20.68	37.93	0	0	0	0	0	0	37.93	0	0	3.44
Jug	0	0	0	25	0	0	0	4.16	12.5	0	20.83	8.33	20.83	8.33
H.phn	0	0	5.88	17.64	0	0	0	0	47.05	0	0	0	29.41	0
Bottle	0	0	0	0	0	0	0	0	1.4	0	1.4	7.04	0	90.14

'Mouse', 'Mug' and 'Mobile', the F-Measure is lower but comparable to that of the complete system, demonstrating the discriminative power of $f^1_{o_i}$ for these object classes. However, we expect that in a realistic scenario with more class categories the relative significance of relational features would increase, since larger sets of objects would show similar geometry. The performance of the framework with only relational features (column [c]) is close to that of the whole system (column [b]), although slightly inferior.

Table 4: Average joint object classification F-Measure $(\%)$ in Dataset B, [a] with only context independent features, [b] with all the proposed features and [c] without context independent features, and the number of object examples in the dataset. The bold font indicates the best score.

	[a]	[b]	[c]	#Obj.
Monitor	94.38	**96.72**	96.51	447
Keyboard	30.51	94.11	**95.48**	286
Mouse	85.97	**90.53**	86.88	290
Mug	84.61	**87.06**	77.3	157
Lamp	55.07	89.19	**90.66**	39
Notebook	0	5	**14.81**	13
Laptop	66.66	**71.49**	70.19	121
Papers	38.7	**76.57**	76.47	303
Book	10	**41.79**	38.22	107
Mobile	48.27	**50**	46.66	19
Glass	30	**44**	33.33	29
Jug	**32.25**	10.52	7.14	24
Headphones	10.52	**28.57**	14.28	17
Bottle	**98.61**	92.75	90.37	71

Computational time The algorithm is implemented in C++, using the OpenCV library (version 2.4.6.1) and the Point Cloud Library (PCL-1.6), and is executed on a machine equipped with a quad-core Intel Core i7 CPU and 16 GB RAM. The average computational time for the joint object classification of one scene with the proposed optimization technique is 0.04 sec in Dataset A and 0.47 sec in Dataset B. The computational cost of the exhaustive search solution increases exponentially with the number of object instances in the scene. We apply exhaustive search only in Dataset A and compute the average computational time as 132 sec/scene, which reaches 750 sec/scene in the scenes with 7 test objects. This computational time clearly indicates that exhaustive search is not a tractable solution even in a simplified scenario.

5 Conclusion and Future work

This paper presents an approach for joint object classification in 3D indoor scenes, based on the representation of the scene as a graph and on a voting strategy that takes into account the typical spatial features and spatial arrangement of the objects, as well as their occurrence and co-occurrence frequencies. The method is tested on two 3D datasets of office desk scenes, where two different sets of object categories are annotated, and the obtained results show that it can distinguish objects among a set of different categories. In our experiments we also observe that context independent features have less impact than the all feature set (including relational features), which indicates that the method could complement a vision based approach. Finally, we demonstrate the advantage of the proposed optimization technique over exhaustive search in terms of computational time. Future work will include the acquisition of a dataset with a higher number of object categories and data from different types of indoor scenes. In our ongoing work, we aim to combine the proposed approach with vision-based systems for object and scene classification, where the present object class prediction framework could provide a prior probability for visual object classification, as well as use visual object classification as an input.

References

Alberti, M.; Folkesson, J.; and Jensfelt, P. 2014. Relational approaches for joint object classification and scene similarity measurement in indoor environments. In *AAAI 2014 Spring Symposia: Qualitative Representations for Robots*.

Aydemir, A.; Sjöö, K.; Folkesson, J.; and Jensfelt, P. 2011. Search in the real world: Active visual object search based on spatial relations. In *ICRA 2011: Proceedings of the IEEE International Conference on Robotics and Automation*.

Bay, H.; Ess, A.; Tuytelaars, T.; and Gool, L. 2008. Surf: Speeded up robust features. *Computer Vision and Image Understanding* 110:346359.

Boutell, M. R.; Luo, J.; and Brown, C. M. 2006. Factor graphs for region-based whole-scene classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*.

Burbridge, C., and Dearden, R. 2012. Learning the geometric meaning of symbolic abstractions for manipulation planning. In *TAROS 2012: Proceedings of Towards Autonomous Robotic Systems*, 220–231.

Choi, W.; Chao, Y.-W.; Pantofaru, C.; and Savarese, S. 2013. Understanding indoor scenes using 3d geometric phrases. In *CVPR 2013: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR 2005: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 886–893.

Felzenszwalb, P.; Girshick, R.; McAllester, D.; and Ramana, D. 2010. Object detection with discriminatively trained part based models. *Pattern Analysis and Machine Intelligence* 32:1627–1645.

Galleguillos, C.; Rabinovich, A.; and Belongie, S. 2008. Object categorization using co-occurrence, location and appearance. In *CVPR 2008: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Kasper, A.; Jakel, R.; and Dillmann, R. 2011. Using spatial relations of objects in real world scenes for scene structuring and scene understanding. In *ICAR 2011: Proceedings of the 15th International Conference on Advanced Robotics*.

Kawewong, A.; Tangruamsub, S.; and Hasegawa, O. 2010. Position-invariant robust features for long-term recognition of dynamic outdoor scenes. *EICE Transactions on Information and Systems* 93(9):2587–2601.

Lin, D.; Fidler, S.; and Urtasun, R. 2013. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV 2013: Proceedings of the 14th International Conference on Computer Vision*.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *ICCV 1999: Proceedings of the 7th International Conference on Computer Vision*, 1150–1157.

Rijsbergen, C. J. V. 1979. *Information retrieval*. London: Butterworth, second edition.

Rosman, B., and Ramamoorthy, S. 2011. Learning spatial relationships between objects. *International Journal of Robotics Research* 30:1328–1342.

Southey, T., and Little, J. J. 2007. Learning qualitative spatial relations for object classification. In *IROS 2007 Workshop: From Sensors to Human Spatial Concepts*.

Yao, J.; Fidler, S.; and Urtasun, R. 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR 2012: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 702–709.

Ye, J., and Hua, K. A. 2013. Exploiting depth camera for 3d spatial relationship interpretation. In *MMSys 2013: Proceeding of Multimedia Systems Conference*, 151–161.

Zhu, X.; Vondrick, C.; Ramanan, D.; and Fowlkes, C. C. 2012. Do we need more training data or better models for object detection? In *British Machine Vision Conference (BMVC)*.