Relational Approaches for Joint Object Classification and Scene Similarity Measurement in Indoor Environments

Marina Alberti, John Folkesson and **Patric Jensfelt**
Computer Vision and Active Perception Lab
The Royal Institute of Technology (KTH)
Stockholm, Sweden
{malberti, johnf, patric}@kth.se

Abstract

The qualitative structure of objects and their spatial distribution, to a large extent, define an indoor human environment scene. This paper presents an approach for indoor scene similarity measurement based on the spatial characteristics and arrangement of the objects in the scene. For this purpose, two main sets of spatial features are computed, from single objects and object pairs. A Gaussian Mixture Model is applied both on the single object features and the object pair features, to learn object class models and relationships of the object pairs, respectively. Given an unknown scene, the object classes are predicted using the probabilistic framework on the learned object class models. From the predicted object classes, object pair features are extracted. A final scene similarity score is obtained using the learned probabilistic models of object pair relationships. Our method is tested on a real world 3D database of desk scenes, using a leave-one-out cross-validation framework. To evaluate the effect of varying conditions on the scene similarity score, we apply our method on mock scenes, generated by removing objects of different categories in the test scenes.

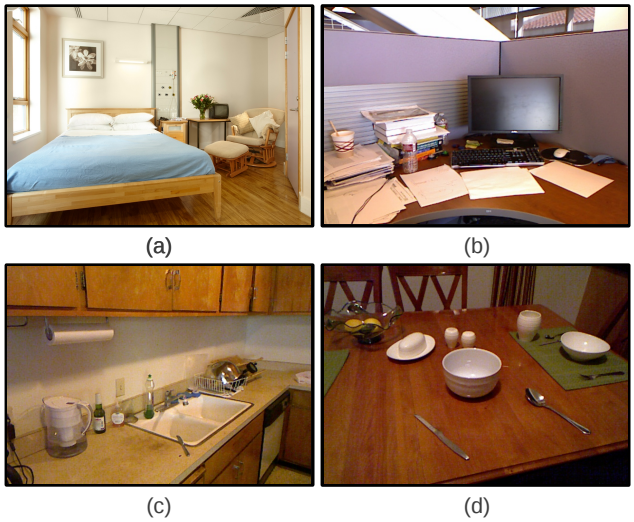


Figure 1: Examples of indoor scenes of different categories: (a) bedroom, (b) office desk, (c) kitchen and (d) living room.

1 Introduction

As robots are becoming more capable of performing a wide range of tasks in applications such as surveillance, service robotics, object search and retrieval, human care, security and object manipulation, and as they increase their interaction with humans, the ability to predict scene categories is becoming more important in robotics. An indoor scene category can be defined as the general concept of a limited region in 3D space that has specific properties or serves a specific purpose. Examples of indoor scene classes are desk area, kitchen, living room and bedroom (see Figure 1).

The traditional object-based computer vision approaches for indoor scene category recognition have limitations (Quattoni and Torralba 2009) due to the variety of object poses and appearance, since the object classifiers (Felzenszwalb et al. 2010) are highly dependent on object pose, colour, texture, camera viewpoint and illumination. Another limitation of traditional computer vision methods is that performance worsens with increasing number of object categories. In spite of having this huge variety in object categories, shapes, poses, texture, etc., it is interesting to note that in our living environments the objects are not placed randomly but tend to maintain a certain ordering and arrangement. In particular, the type of an indoor scene already defines a subset of objects that can be expected to be seen and may also infer a certain relative positioning of the objects. For example, in a desk scene we could expect to see a monitor, a keyboard and a mouse in a certain configuration.

This paper presents an approach for joint object prediction and scene similarity measurement in indoor environments, based on the spatial relations of the objects. A spatial relation specifies how an object is located in space in relation to a reference object (Freeman 1975). Spatial relationships can be roughly divided into two categories, topological and metric relationships. The latter can further be grouped into distance relationships and directional relationships. Additionally, size-based relationships can be defined. In this study, we consider directional, distance and size relations on the

3D geometrical characteristics of the objects.

To compute scene similarity, we learn a probabilistic model for indoor scene category by training on example scenes. A scene is described using a layered representation. In the first layer, the scene is defined as a set of objects, and in the second layer as the set of their spatial relations. To describe these layers, two main feature sets are computed, i) from single objects and ii) object pairs. As single object features, we use pose, bearing with respect to the reference system defined by a landmark, volume and length of the object projection along the X, Y and Z axes. For object pairs, Euclidean distance, vertical displacement, bearing between object centroids and ratio of object volumes are computed. The object categories and the spatial relations of pair of objects of different categories are modeled by fitting a probability distribution - a Gaussian Mixture Model (GMM) - to these sets of features. For an unknown test scene, we first predict the object class categories using the learned object category model. From the predicted object classes, object pair features are extracted. Scene similarity is then computed using the object category based object pair relationships. By applying a threshold on this similarity score, the test scene can be classified.

This paper presents the above mentioned concept in a simplistic dataset, which will be extended in the future. The method is tested on a database of 3D data of office desk scenes acquired with an RGB-Depth sensor. The scene information is represented as a 3D point cloud. Points corresponding to separate objects are segmented by a manual annotation where 3D bounding boxes are used to define the objects. 3D geometrical characteristics are computed from the segmented objects. A leave-one-out cross-validation experiment is performed to evaluate the performance of object classification and scene similarity measurement. The method is also applied on mock scenes, generated by removing objects of different categories, to investigate the effect of varying conditions on the scene similarity score.

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

2 Related work

Recent studies (Southey and Little 2007; Kasper, Jakel, and Dillmann 2011) compute the distribution of spatial relations of objects over a set of scenes, and show how the obtained data and models can be used in typical scenarios for service robotics, such as the recognition of individual objects given the knowledge of the other objects types in the scene. Kasper et al. (Kasper, Jakel, and Dillmann 2011) develop an empirical base for scene understanding by encoding the structure of the scene in the spatial relations between the objects. The distribution of the spatial relations in office desk scenes is learned from a real world 3D dataset and statistical measurements are computed. Different types of spatial relations are considered: size of the 3D object bounding boxes, Euclidean distance between object centers, relative position of an object with respect to the reference system defined by a reference object. 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.

Spatial 3D features and spatial relations between pairs of objects are used in several robotics studies, in the context of navigation planning, object recognition, object position prediction and manipulation (Rosman and Ramamoorthy 2011; Ye and Hua 2013; Burbridge and Dearden 2012). Rosman et al. (Rosman and Ramamoorthy 2011) redescribe a 3D scene in terms of a layered representation, consisting in the skeletonized description of the objects structure as a network of contact points, as well as a symbolic description of the spatial relationships between objects. The scene description is based on the assumption that the objects are in contact. Ye et al. (Ye and Hua 2013) compute 3D directional spatial relations between pairs of object by dividing the 3D space around an object into 26 primitive directions, using a method inspired by ray-tracing. In the work of Burbridge et al. (Burbridge and Dearden 2012), a bi-directional mapping is learned between geometric and symbolic states of object configurations in the context of a manipulation planning system, by using Gaussian Kernel Density estimation. Dee et al. (Dee, Hogg, and Cohn 2009) propose a method for scene modeling and classification from RGB video data, using quantized descriptions of motion. The method learns spatial relationships between parts of frames in videos which correspond to regions experiencing similar motion.

State-of-the-art methods for scene classification include object-based and context-based approaches. In object-based scene categorization, the objects are recognized and used as landmarks. Different vision-based features are proposed in the literature for object recognition, such as SIFT, HOG, SURF and PIRFS (Lowe 1999; Dalal and Triggs 2005; Bay et al. 2008; Kawewong, Tangruamsub, and Hasegawa 2010). Many approaches for visual classification are restricted in their ability to classify large number of objects. Moreover, visual object recognition methods fail on seemingly simple examples, when the objects lack sufficiently distinctive appearance data. In context-based scene recognition, features of the whole scene are described after compression in a low-dimensional space, based on mechanisms that humans use to recognize scenes. Olivia et al. (Oliva and Torralba 2006) propose *Gist* as a feature to describe global characteristics of a scene. *Gist* is used popularly in context-based feature description. Several approaches have also been developed for vision-based semantic scene recognition, from 2D images representing outdoor environments (Boutell, Luo, and Brown 2006; Yao, Fidler, and Urtasun 2012). However, when applying the state-of-the-art outdoor scene classification methods in indoor scenes, it is observed that accuracy drops dramatically (Quattoni and Torralba 2009). On the other hand, the probabilistic approach for indoor scene similarity measurement proposed in this paper does not need the computation of complex features and visual classifiers.

The most recent papers (Ye and Hua 2013; Kasper, Jakel, and Dillmann 2011) on the spatial distribution of objects are typically tested on RGB-Depth databases. Depth infor-

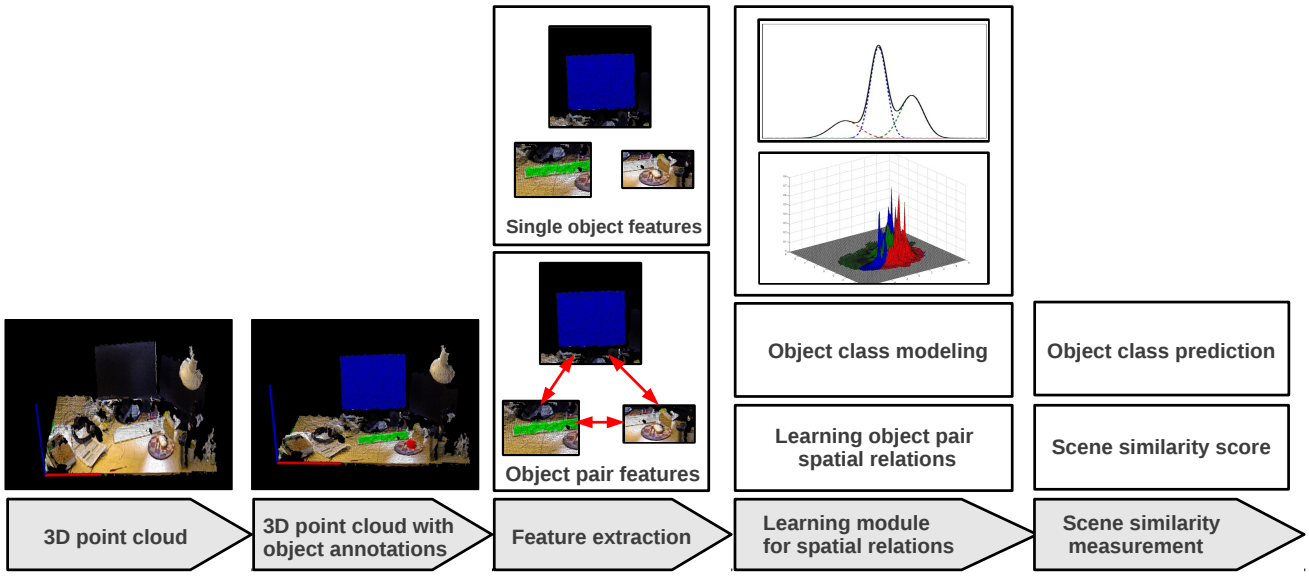


Figure 2: Overview of the proposed system. The indoor scene information is represented in the form of a 3D point cloud. Different objects are segmented from the point cloud. Spatial features are extracted from single objects and from object pairs. A learning module trains a GMM on features of different object categories and on spatial relations of objects of different classes, on a set of indoor scenes of same scene class type. Given an unknown scene, the method predicts object classes and computes a measure of scene similarity with respect to the modeled scene class type.

mation can also be obtained by using multi-camera stereo technology (Rosman and Ramamoorthy 2011). Point clouds are commonly used to represent the 3D scene information (Ye and Hua 2013; Kasper, Jakel, and Dillmann 2011; Rosman and Ramamoorthy 2011). Some studies also use data in simulated environments (Burbridge and Dearden 2012). In our work, we acquire an RGB-D database using the Asus Xtion Pro-Live sensor and we work on the obtained 3D point clouds.

The contributions of this paper are as follows. We introduce a wider set of spatial features, both for single objects and object pairs, along with the features inspired by the state-of-the-art methods. We extend the concept of object class prediction described in Kasper et al. (Kasper, Jakel, and Dillmann 2011) and Southey et al. (Southey and Little 2007) to joint object classification and scene similarity measurement using a probabilistic framework. Finally, we apply the proposed approach on a simple dataset of desk scenes and obtain encouraging results.

3 Proposed method

Let N be the number of examples of indoor human environment scenes of scene class type t, $S_t = \{s_1, s_2, \ldots, s_N\}$. We assume that each example scene, s_i, $i = 1, 2, \ldots, N$, contains a landmark object, o_l, and a set of other objects of different categories, shapes and sizes, $O = \{o_1, o_2, \ldots, o_m\}$. The landmark is used to define a reference system for the other objects. For example, in an office desk scene, the landmark can be the desk, and other objects can be monitor, keyboard, mouse, mug, lamp, pens, etc. We take into account that these object classes can be missing from the scene or be present in multiple instances.

The proposed algorithm for scene similarity measurement is composed of three main steps: feature extraction, modeling of object classes and object pair relationships in a learning phase, and object class prediction and computation of a scene similarity score in test scenes. The pipeline for the proposed method is described in Figure 2.

3.1 Feature extraction

To model the object categories and the relationships between pairs of objects of different categories, it is necessary to obtain a feature set that best captures the object geometry and the spatial distribution of objects in the scene. Following are the proposed feature sets for single objects and object pairs.

Single object features Single object features f_{o_i} are computed from the 3D spatial characteristics of the object itself w.r.t. a landmark object. These features are inspired by the state-of-the-art methods (Kasper, Jakel, and Dillmann 2011; Burbridge and Dearden 2012), but unlike in the previous literature, we use the table as the landmark object and we consider the angles formed by the objects w.r.t. the table centroid and the front-left table corner. The 7 computed features are distributed in 4 main categories:

- Position features: 3D position of the object centroid (1×3).
- Angle features, computed in the extrinsic coordinate system (see Figure 4):
  - 2D (horizontal) bearing of object centroid from front-left table corner, θ_1 (1×1).

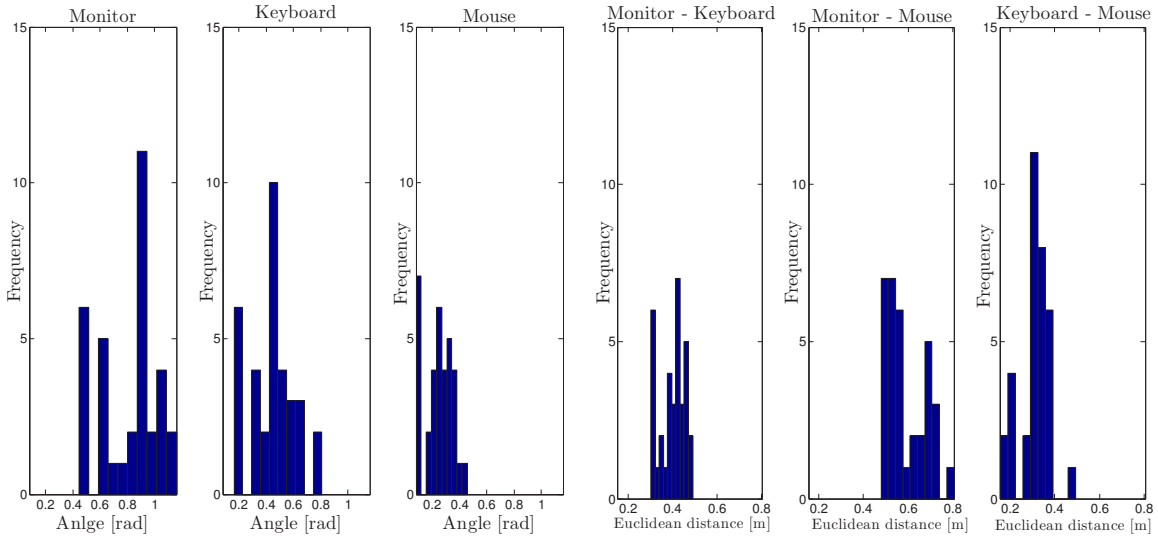


Figure 3: Examples of (a) single object features: the histograms of the bearing of the object centroid from the front-left table corner for the objects monitor, keyboard and mouse and (b) object pair features: the histograms of the Euclidean distance between object centroids, for three pairs of object classes, monitor-keyboard, monitor-mouse and keyboard-mouse.

  - 2D (horizontal) bearing of object centroid from table plane centroid, θ_2 (1×1).
- Volume (1×1).
- Projected length features: length of the object projection along X, Y and Z axes (1×3).

In Figure 4, a scheme representing the desk surface and the described *angle* features computed for an example object is illustrated. Figure 3-a shows the distribution of one of these features, the angle between object centroid and front-left table corner, θ_1, for the object categories monitor, keyboard and mouse, over a set of analyzed office desk scenes.

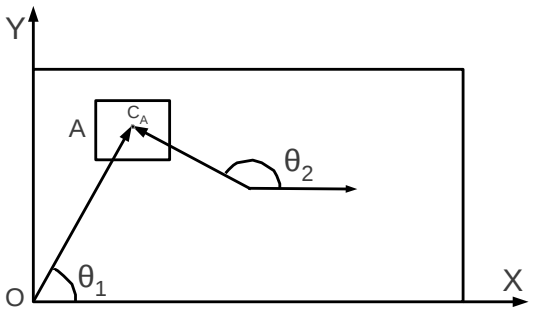


Figure 4: Representation of the computed *angle* features: the bearing of the object centroid C_A from the front-left table corner, θ_1, and the bearing of C_A from the table centroid, θ_2, where the large rectangle represents the desk surface.

Object pair features To capture the 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 as for single object features. The proposed feature set consists of 5 features:

- $d(C_{o_i}, C_{o_j})$, where d is the Euclidean distance and C_{o_i}, C_{o_j} are the centroids of objects o_i and o_j, respectively (1×1) (Kasper, Jakel, and Dillmann 2011).
- $d_{XY}(C_{o_i}, C_{o_j})$ projected on the X-Y plane (1×1) (Kasper, Jakel, and Dillmann 2011).
- 2D (horizontal) bearing from C_{o_i} to C_{o_j} (1×1).
- Ratio of object volumes (1×1).
- $d_Z(C_{o_i}, C_{o_j})$, where d_Z is the vertical displacement (1×1).

The ratio of object volumes and the vertical displacement between object centroids are novel features. In Figure 3-b, a plot of the Euclidean distance between object centroids is illustrated for different object category pairs, computed over a set of desk scenes.

3.2 Learning object categories and object pair relationships

As the object class prediction and the scene similarity measurement (described in Section 3.3) follow the probabilistic framework, we choose to learn both the object class categories and the object pair relationships using a Gaussian Mixture Model based representation. By observing the plots of the features (see Figure 3) we infer that a GMM would be a reasonable choice to model the multivariate probability distribution of both single object and object pair features.

Object class modeling To model the object class category of $o_i \in O$, a set of *single object features* f_{o_i} is extracted from the pre-segmented objects. The object class model is learned

by applying a Gaussian Mixture Model on the single object feature set f_{o_i}:

$$GMM(f_{o_i}, \mu^z_{o_i}, \Sigma^z_{o_i}) = \sum_{z=1}^{n_c} \pi_z \frac{1}{K} exp(-\frac{1}{2}\zeta^z_{o_i}), \quad (1)$$

where: $\zeta^z_{o_i} = (f_{o_i} - \mu^z_{o_i})^T {\Sigma^z_{o_i}}^{-1} (f_{o_i} - \mu^z_{o_i})$,
$K = \sqrt{(2\pi)^{dim} |\Sigma^z_{f_{o_i}}|}$, $\pi_z \geq 0$, $\sum_{z=1}^{d} \pi_z = 1$, n_c is the number of mixtures, π_z is the weight of the z^{th} mixture, $\mu^z_{o_i}$ is the mean of the normal distribution, $\Sigma^z_{o_i}$ is the covariance matrix and dim is the dimensionality of the feature space.

Learning object pair relationships Object pair relationships are the next important factor in the scene similarity measurement. To learn these relationships, a different set of features f_{o_i,o_j} is computed as *object pair features*. For m objects in O we compute object pair features for all the combinations of object classes (o_i, o_j), where $i \neq j$. Only the relations of objects of different categories are learned. For example, in a desk scenario with two monitors, the relations between the two monitors will not be computed, although the relations between each of the two monitors and all the other objects are considered. The probability distribution of these features is modeled in a multi-dimensional feature space by applying a Gaussian Mixture Model on the object pair feature set f_{o_i,o_j}:

$$GMM(f_{o_i,o_j}, \mu^z_{o_i,o_j}, \Sigma^z_{o_i,o_j}) = \sum_{z=1}^{n_c} \pi_z \frac{1}{K} exp(-\frac{1}{2}\zeta^z_{o_i,o_j}), \quad (2)$$

where:
$\zeta^z_{o_i,o_j} = (f_{o_i,o_j} - \mu^z_{o_i,o_j})^T {\Sigma^z_{o_i,o_j}}^{-1} (f_{o_i,o_j} - \mu^z_{o_i,o_j})$,
$K = \sqrt{(2\pi)^{dim} |\Sigma^z_{f_{o_i,o_j}}|}$, $\pi_z \geq 0$, $\sum_{z=1}^{d} \pi_z = 1$, n_c is the number of mixtures, π_z is the weight of the z^{th} mixture, $\mu^z_{o_i,o_j}$ is the mean of the normal distribution, $\Sigma^z_{o_i,o_j}$ is the covariance matrix and dim is the dimensionality of the feature space.

3.3 Scene similarity measurement

After training the models for object classes and object pair spatial relations, the scene similarity score $sim(s_u, S_t)$ of an unknown scene s_u with respect to the modeled scene class type t, S_t, is computed.

Object class prediction As a first step, object classes are predicted using a probabilistic framework in the learned object class models. We assume that the objects are already segmented from the input point cloud data. Given an unknown object $o_u \in s_u$, the set of single object features f_{o_u} are extracted. The probability that o_u is an object of given object class type j can be formulated by applying the Bayes' theorem:

$$Pr(o_j|f_{o_u}) = \frac{Pr(f_{o_u}|o_j) \cdot Pr(o_j)}{Pr(f_{o_u})}, \forall o_j \in O, \quad (3)$$

where $Pr(f_{o_u}|o_j)$ is the likelihood of the trained object category model given the features from o_u, which is equal to the conditional probability of the features given the parameter values of the model, $Pr(o_j)$ is the a-priori probability of the object class j and $Pr(f_{o_u})$ is the probability of the feature set f_{o_u}. Since $Pr(f_{o_u})$ does not depend on the object class, Equation 3 can be simplified as:

$$Pr(o_j|f_{o_u}) \propto Pr(f_{o_u}|o_j) \cdot Pr(o_j), \forall o_j \in O. \quad (4)$$

In our work, the a-priori probability $Pr(o_j)$ is computed based on the frequency of appearance of the object category j in the training dataset:

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

where N_{o_j} is the number of training scenes where o_j is present and N_{tot} is the total number of training scenes. The object class is predicted as:

$$j^* = argmax_j \Big(GMM(f_{o_u}, \mu^z_{o_j}, \Sigma^z_{o_j})) \cdot \frac{max(1, N_{o_i})}{(1 + N_{tot})} \Big), \quad (6)$$

where $o_j \in O$ and $GMM(f_{o_u}, \mu^z_{o_j}, \Sigma^z_{o_j})$ is the learned object class model (Equation 1).

Detecting the *other* object class An additional outliers detection stage is proposed to address the presence of objects which are not learned during training in the test scenes and thus make the framework more general. A threshold υ_j is set on the likelihood of each object category j. If $GMM(f_{o_u}, \mu^z_{o_j}, \Sigma^z_{o_j}) < \upsilon_j$, the category j is not considered in Equation 6. The object o_u is classified as *other* object category if $GMM(f_{o_u}, \mu^z_{o_j}, \Sigma^z_{o_j}) < \upsilon_j, \forall j$. In this case the object is not considered in the successive scene similarity score computations. This threshold, υ_j, is set to the minimum value of the likelihood $GMM(f_{o_j}, \mu^z_{o_j}, \Sigma^z_{o_j})$ computed over the training data.

Scene similarity score From the predicted object classes (Equation 6), object pairs are identified and the features from object pairs are computed, $f_{o_i,o_j}, \forall o_i, o_j \in O, i \neq j$.

A final scene similarity score is obtained using the learned probabilistic models of object pair relationships (Equation 2). The similarity score, $sim(s_u, S_t)$, is computed as a sum of the likelihoods of the trained object pair relation models, for all the object pairs identified in s_u, weighted by the probability of co-occurrence of the corresponding object categories. This weight is introduced because relationships between objects that are more frequently found together should have a larger impact on the similarity score:

$$sim(s_u, S_t) = \sum_{\substack{o_i,o_j \in s_u \\ i \neq j}} Pr(f_{o_i,o_j}|S_t) \cdot Pr(o_i, o_j), \quad (7)$$

where $Pr(f_{o_i,o_j}|S_t)$ is the likelihood of the trained object pair spatial relation model given the features from that

object pair computed in s_u, which corresponds to the conditional probability of the features from s_u given the parameter values of the model and $Pr(o_i, o_j)$ is the probability of co-occurrence of the object categories i and j. This second term represents how often this type of relation has been observed and is an important information about the relation. In our work, $Pr(o_i, o_j)$ is computed based on the frequency of co-occurrence of the object categories in the training dataset:

$$Pr(o_i, o_j) = \frac{max(1, N_{o_i,o_j})}{(1 + N_{tot})}, \quad (8)$$

where N_{o_i,o_j} is the number of training scenes where both o_i and o_j are present and N_{tot} is the total number of training scenes. Equation 7 can be expressed as:

$$sim(s_u, S_t) = \sum_{\substack{o_i,o_j \in s_u \\ i \neq j}} GMM(f_{o_i,o_j}, \mu^z_{o_i,o_j}, \Sigma^z_{o_i,o_j}) \cdot \frac{max(1, N_{o_i,o_j})}{(1 + N_{tot})}, \quad (9)$$

where $GMM(f_{o_i,o_j}, \mu^z_{o_i,o_j}, \Sigma^z_{o_i,o_j})$ is the learned object pair relation model.

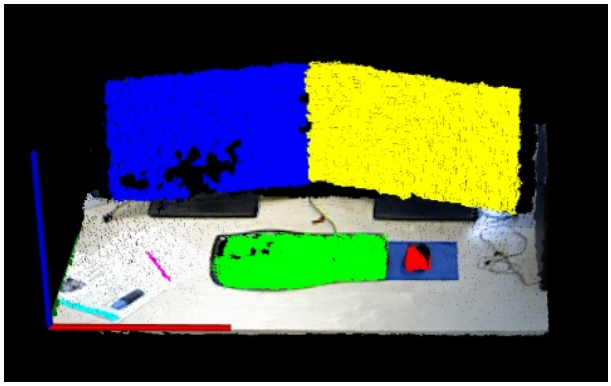
The presented approach for scene similarity measurement can be viewed as labeling each object in the test scene as one of the training object categories (or as the *other* class) and computing a scene similarity score based on the object pair spatial relations. Only a particular set of object class labels assigned to the objects in the test scene will give the maximum scene similarity score. Exhaustive search by considering all the possible object category labels is an NP problem. To make this problem solvable in polynomial time, we apply object class prediction (Equation 6) to obtain an initial guess of the object class labels and then use the object pair features to compute the final scene similarity score (Equation 9).

4 Experimental results

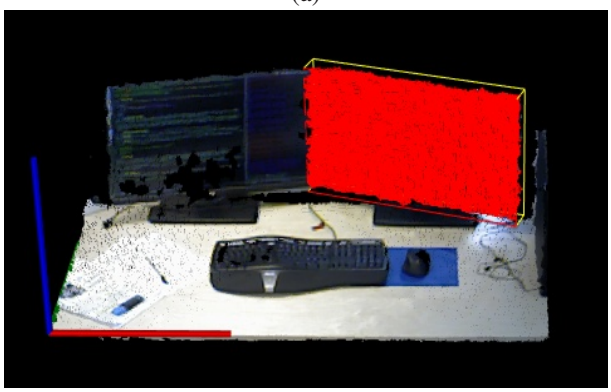
The proposed method is trained on a database of desk scenes where seven object categories, namely monitor, keyboard, mouse, mug, lamp, laptop and pen/pencil, are manually annotated with their bounding boxes and object category labels. Performance is evaluated by using leave-one-out cross-validation and by performing tests with mock scenes.

4.1 Database description

Our experiments are performed on a database of 3D office desk scenes which contains 42 scene examples. The database scenes contain 57 monitors, 42 keyboards, 37 computer mice, 15 mugs, 26 lamps, 5 laptops and 43 pens/pencils. The data are acquired using an RGB-D sensor, namely the Asus Xtion Pro Live sensor. Single snapshots are captured and stored as point clouds. The database is manually annotated by labeling the desk and objects on the desk using 3D cuboidal bounding boxes, as shown in Figure 5. In the annotated scenes, a local reference frame is defined by the desk, having its Cartesian axes aligned with the two axes of the desk and with the vertical direction, respectively, and having its origin in the front-left desk corner. The output of the annotation for each scene is an XML file which stores the information on the object types, the geometry of the bounding boxes and the point cloud indexes of each object.



(a)



(b)

Figure 5: (a) Point cloud with object annotations for two monitors (blue and yellow), a keyboard (green), a mouse (red) and two pens (cyan and magenta) and (b) the 3D bounding box of a monitor is highlighted in the same scene.

4.2 Object classification performance

A leave-one-out cross validation experiment is performed on the whole dataset and object classification accuracy is computed for each fold. In the experiment, n_c = 2 mixture components are used for the GMM of each object class.

In a first experiment we investigate the influence and relevance of the different *single object features*, by training and testing the models on different subsets of features. Table 1 shows the F-Measure[1] (Rijsbergen 1979) computed for each of the seven object categories by using all the proposed features and by using different subsets of features: position, angles, volume, projected object lengths along the X, Y and Z axes and 2D (horizontal) position + angles. It can be observed that the set of all proposed features (column [a]) and the subset of *projected object lengths* features (column [e])

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

Table 1: Comparison of average object classification F-Measure (%) for each of the considered object categories, by using all the proposed *single object features* f_{o_i} (column [a]) and by using different subsets of features: position ([b]), angles ([c]), volume ([d]), projected object lengths along the X, Y and Z axes ([e]), horizontal position + angle ([f]).

	[a]	[b]	[c]	[d]	[e]	[f]
Monitor	96.49	93.1	78.74	85.95	94.01	81.53
Keyboard	100	75	50	87.8	100	61.36
Mouse	98.63	81.08	56	100	100	53.76
Mug	96.77	96.55	0	82.35	100	0
Lamp	92.59	90.56	42.1	83.33	82.35	47.36
Laptop	75	57.14	46.15	0	75	66.66
Pen/pencil	100	85.71	0	100	100	21.91

Table 2: Object classification performance in terms of precision (P), recall (R) and F-Measure (F) (%) without the outliers detection stage (columns 2-4) and by setting a threshold on the likelihood of the single object features to discard outliers (columns 5-7).

	Without outliers detection			With outliers detection		
	P	R	F	P	R	F
Monitor	96.49	96.49	96.49	100	92.98	96.36
Keyboard	100	100	100	100	97.61	98.79
Mouse	100	97.29	98.63	100	91.89	95.77
Mug	93.75	100	96.77	100	86.66	92.85
Lamp	89.28	96.15	92.59	88.88	92.30	90.56
Laptop	100	60	75	100	20	33.33
Pen/pencil	100	100	100	100	88.37	93.82

yield the best results. Both the *angle* features alone (column [c]) and the *2D (horizontal) position + angle* features (column [f]) fail in recognizing mugs and pens/pencils. This latter is a predictable result, since the 2D location of mugs, pens and pencils on the desk typically has high variability. These results show that, in this simple scenario, the object size is sufficient to identify the category of the objects. However, we expect that in a more realistic scenario the relative relevance of the other features will increase, because, with a higher number of possible object categories, more of them will present similar volume.

In a second experiment, the effect of the proposed strategy to detect the *other* object class is explored. Table 2 shows the performance scores of precision, recall and F-Measure computed in the cross-validation (considering all the features) both without and with the presented outliers detection strategy. As expected, outliers detection increases the precision, while the recall and the F-Measure slightly decrease, and it can be observed that the system remains robust. For the *laptop* class, much lower scores are obtained than for the other object classes, because the number of training samples is particularly low: only 5 laptops are present. We additionally test the outliers detection strategy on two scenes where objects of two other classes are annotated: a notebook and a telephone. Both objects are correctly detected as outliers.

4.3 Scene similarity measurement performance

In the same cross-validation framework, the scene similarity score $sim(s_u, S_t)$ is computed for each scene in the dataset, by using the probabilistic models of object pair relationships learned on the training scenes (Equation 9). In the experiments, $n_c = 2$ mixture components are used for the GMM of each object pair relation. It is observed that the GMM model is not robust when few samples are used for training. To obtain meaningful likelihood values for the scene similarity score computation, the GMM models are trained only if a sufficient number of samples is present, which in these experiments is set to 10 samples. This constraint excludes all object pair relationships involving the laptop, which is present in only 5 scenes.

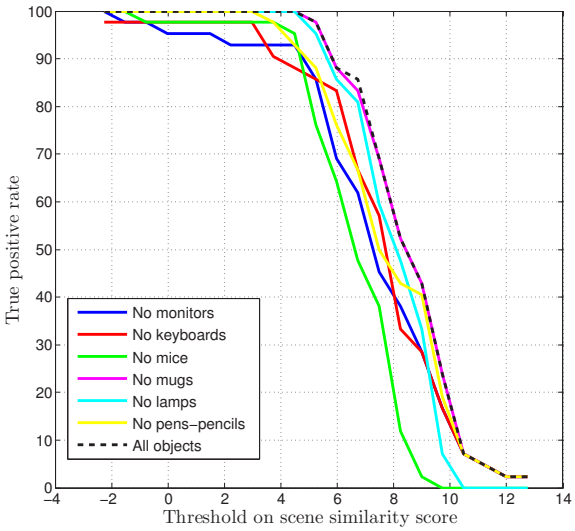


Figure 6: Plots of the average True Positive rate for scene class prediction as a function of the threshold on the scene similarity score (τ), for the original scenes (dotted black line) and scenes where removed objects. The values on the horizontal axis are in logarithmic scale.

To evaluate the effect of varying conditions on the scene similarity score, our method is applied on mock scenes generated by removing all object instances of each of the seven objects categories from the test scenes. This experiment reflects a typical situation of a real world sensor acquisition, when some objects may not be correctly segmented, or not all of the expected object classes may be present. The scene category can be predicted by defining a threshold τ on the similarity score $sim(s_u, S_t)$ and classifying the scene as $s_u \in S_t$ if $sim(s_u, S_t) > \tau$. Figure 6 shows the plots of the average True Positive (TP) rate for scene class prediction as a function of τ, for both the original scenes and the scenes obtained by removing different categories of objects. It can be observed that when removing an object, $sim(s_u, S_t)$ tends to decrease. For example, when τ is set to have 97% TP on the test scenes with all object categories, the TP score becomes 85% for 'no monitors', 85% for 'no keyboards', 76% for 'no computer mice', 97% for 'no mugs', 95% for 'no lamps' and 88% for 'no pens/pencils'. These

results, obtained by testing the method on a limited dataset of only one scene category (an office desk scene), are reasonable and in line with our expectations. The extension of our dataset to more scene categories will allow further tests on scene similarity measurement.

5 Conclusion and Future work

This paper presents an approach for indoor scene similarity measurement and object classification using 3D spatial relations of objects. The proposed method does not require the computation of complex features. Moreover, since the method is based on the spatial distribution of the objects, variations in object pose, appearance and texture are not a limitation, as it happens for visual object classifiers. The method is tested in a simple scenario covering a single scene type, i.e. office desk scenes, with encouraging experimental results. The probabilistic object category classification achieves high accuracy. Our object class prediction framework can provide a prior probability for visual object classification. We perform an analysis of the influence of the different features on object class prediction, which will be more relevant in a more complex dataset. An outliers detection strategy is integrated to identify the presence of unseen object categories in the test scenes. We evaluate the effect of varying conditions on the scene similarity score by removing objects of different categories in the test scene and we observe the expected behaviour of the scene similarity score. Acknowledging the limitations of our dataset, future work will include the acquisition of more scene categories, with a higher number of object classes per scene category and a higher number of scene examples, to make the training data more representative. This will also help us to test the generality of the proposed concept. Moreover, future work will address the introduction of pairwise object features in the object classification phase. In fact, the need for a reliable reference object such as the table can be overcome by using object pair features and volume in the object classification phase. It would also be interesting to integrate the proposed spatial relation based scene classification method with state-of-the-art computer vision based features for scene classification. Finally, the presented approach could be extended to obtain a set of qualitative spatial relations between object pairs, such as *right of*, *bigger* and *near* etc.

6 Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 600623 ("STRANDS").

References

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.

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.

Dee, H. M.; Hogg, D. C.; and Cohn, A. G. 2009. Scene modelling and classification using learned spatial relations. In *Proceedings of the 9th international conference on Spatial information theory*, 295–311.

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.

Freeman, J. 1975. The modelling of spatial relations. *Computer Graphics and Image Processing* 4:156–171.

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.

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.

Oliva, A., and Torralba, A. 2006. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*.

Quattoni, A., and Torralba, A. 2009. Recognizing indoor scenes. In *CVPR 2009: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 413–420.

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.