A Comparison of Qualitative and Metric Spatial Relation Models for Scene Understanding

Akshaya Thippur*, Chris Burbridge†, Lars Kunze†, Marina Alberti*
John Folkesson*, Patric Jensfelt*, Nick Hawes†
*Centre for Autonomous Systems, KTH Royal Institute of Technology, Sweden
†Intelligent Robotics Lab, School of Computer Science, University of Birmingham, United Kingdom

Abstract

Object recognition systems can be unreliable when run in isolation depending on only image based features, but their performance can be improved when taking *scene context* into account. In this paper, we present techniques to model and infer object labels in real scenes based on a variety of *spatial relations* – geometric features which capture *how* objects co-occur – and compare their efficacy in the context of augmenting perception based object classification in real-world table-top scenes. We utilise a long-term dataset of office table-tops for qualitatively comparing the performances of these techniques. On this dataset, we show that more intricate techniques, have a superior performance but do not generalise well on small training data. We also show that techniques using coarser information perform crudely but sufficiently well in standalone scenarios and generalise well on small training data. We conclude the paper, expanding on the insights we have gained through these comparisons and comment on a few fundamental topics with respect to long-term autonomous robots.

1 Introduction

Human environments are characterized by the objects they contain and if robots are to perform useful service tasks for humans, it is crucial that they are able to locate and identify a wide variety of objects in everyday environments. State-of-the-art object recognition/classification typically relies on matching the features extracted with models built through machine learning techniques. As the number of object classes a given system is trained to recognise increases, the uncertainty of individual object recognition tends to increase - due to the high chance of existence of overlapping features between the training examples of such classes. The reliability of such recognisers is also affected when used by real robots in everyday environments, as objects may be partially occluded by scene clutter or only visible from certain angles, both potentially reducing the visibility of features for their trained models. In this paper we argue that the performance of a robot on an object recognition task can be increased by the addition of *contextual knowledge* about the scene the objects are found in. In particular we demonstrate how models of the *spatial configuration* of objects, learnt over prior observations of real scenes, can allow a robot to recognise the objects in novel scenes more reliably.

Our work is performed in the context of developing a mobile service robot for long-term autonomy in indoor human environments, from offices to hospitals. The ability for a robot to run for weeks or months in its task environment opens up a new range of possibilities in terms of capabilities. In particular, any task the robot performs will be done in an environment it may have visited many times before, and we wish to find ways to capture the contextual knowledge gained from previous visits in a way that enables subsequent behaviour to be improved. The use of context to improve object recognition is just one example of this new intelligent robotics paradigm.

In this paper we focus on the task of *table-top scene understanding*, and more specifically what objects are present on a table-top. Whilst the objects present on a single table may change in position, their overall arrangement has some regularity over time as influenced by the amount and type of use to which the table is put. For example, if this table is used for computing, then a (relatively static) monitor will be present, with a keyboard in front of it and mouse to one side. A drink, or paper and a pen, may be within an arm's length of the keyboard, as may headphones or a cellphone. This arrangement may vary across different tables in the same building, but the overall pattern of arrangements will contain some common underlying structure. It is this structure we aim to exploit in order to improve the recognition of table-top objects, e.g. knowing that the object to the right of a keyboard is more likely to be a mouse than a cellphone.

As the absolute positions of objects on a table (or their relative positions with respect to some fixed part of the table) is quite unlikely to generalise across a range of different tables, we are investigating *relational* models of space, i.e. ways of encoding the position of a target object relative to the position of one or more landmark objects. Using a data set of table-top scenes (Thippur et al. 2014) (briefly described in Section 3.1), in this paper we explore the performance of a variety of representations for relative object position, plus inference techniques for operating on them, on the task of table-top scene understanding (Section 4). In particular we investigate representations that use varying forms of spatial

relations, from actual geometric and metric spatial relations (MSR) such as distances and angles to more qualitative spatial relations (QSR) such as *Left* and *Behind* as a means for capturing observations of object configurations over time.

The contributions this paper makes are: (1) A novel comparison between mechanisms for representing, learning and inferring on object spatial configurations using spatial relations. (2) An evaluation of the use of these mechanisms for augmenting a robot's vision based *perception system* (PS). (3) Our insight on the performances of the various context based recognition systems.

It is unfortunate that we cannot avoid many acronyms in this paper. However, to help the reader, we summarize all of the acronyms in a table in Section (7).

2 Related Work

2.1 Spatial Relations Based Techniques

Spatial relations have been used previously to provide contextual information to vision-related work. Choi et al. (2010) used a hierarchy of spatial relations alongside descriptive features to support multiple object detections in a single image. The work in Divvala et al. (2009) uses a set of contextual cues either depending on image properties such as pixel neighbourhoods, adjacent frames in a video and other image-based features; and on metadata properties such as geotags, photographer ID, camera specifications etc. Most importantly, they use spatial context configurations projected onto typical 2D images, size relationships between object representations in the image and such, to obtain a better semantic segmentation and object annotation in the image. Our QSR models, in contrast to this work, are depending solely on spatial relations in 3D while they use a metric stochastic approach. We also only focus on a group-label assignment problem for already detected objects.

Spatial relations and contextual information are commonly used in activity recognition from video streams. For example, Dubba, Cohn, and Hogg (2010) demonstrate the learning of activity phases in airport videos using spatial relations between tracked objects, and Behera, Cohn, and Hogg (2012) use spatial relations to monitor objects and activities in videos of a constrained workflow environment. Recent work has used object co-occurrence to provide context in visual tasks. Examples in 2D include object co-occurrence statistics in class-based image segmentation (Ladicky et al. 2013); and the use of object presence to provide context in activity recognition (Li et al. 2012). However, all this previous work is restricted to 2D images, whereas our approaches work with spatial context in 3D (RGB-D) data. Authors have also worked with spatial context in 3D, including parsing a 2D image of a 3D scene into a simulated 3D field before extracting geometric and contextual features between the objects (Xiao et al. 2012). Our approaches to encoding 3D spatial context could be applied in these cases, and we use richer, structured models of object relations.

Apart from using the statistics of co-occurrence, a lot of information can be exploited from *how* the objects co-occur in the scene, in particular the extrinsic geometric spatial relations between the objects. Recent work in 3D semantic labelling has used such geometric information along with descriptive intrinsic appearance features (Koppula et al. 2011). They achieve a high classification accuracy for a large set of object-classes belonging to home and office environments. Scene similarity measurement and classification based on contextual information is conducted by Fisher, Savva, and Hanrahan (2011). They also use spatial information for context-based object search using Graph Kernel Methods. The method is further developed to provide synthetic scene examples using spatial relations (Fisher et al. 2012). In (Aydemir et al. 2011) spatial relations between smaller objects, furniture and locations is used for pruning in object search problems in human environments. In (Lin, Fidler, and Urtasun 2013) a technique is developed for automatic annotation of 3D objects. It uses intrinsic appearance features and geometric features and is employed to build an object and scene classifier using conditional random fields. In (Kasper, Jakel, and Dillmann 2011) the authors utilise both geometric single object features and pair-wise spatial relations between objects to develop an empirical base for scene understanding. Recent studies (Southey and Little 2007; Kasper, Jakel, and Dillmann 2011) compute statistics of spatial relations of objects and use it for conditional object recognition for service robotics. Finally, Ruiz-del Solar, Loncomilla, and Saavedra (2013) improve on (Aydemir et al. 2011) by introducing the use of search-masks based on primitive threshold-based distance based spatial relations ranging from "*very far*" to "*very near*". The authors in this work also test only on a simulated environment.

Whilst our techniques are comparable to those in the literature (as above), our contribution comes from the explicit comparison of different representations of spatial context (metric vs qualitative) on a novel, long-term learning task. Additionally our qualitative approach relies on many different kinds of spatial relationships which could be provided through other mechanisms than unsupervised machine learning (e.g. through a human tutor describing a spatial scene), and in this way bootstrap the system using expert knowledge. The spatial relations discriminate object pairs based on location, distance, size etc. We also evaluate our comparison experiments on a recently developed real-world dataset which could evolve into a benchmark for testing such methods.

3 Evaluation Scenario

Our evaluation scenario stems from our interest in long-term autonomy in real environments. The aim of our project is to make long-term intelligent observation/surveillance robots to assist humans working as security guards or personnel in elderly care scenarios. Our techniques will run on a mobile robot capable of patrolling a collection of offices a couple of times during the day. The role of this robot is to inspect the tables in these offices for certain conditions (e.g. checking that laptops or confidential papers have not been left unattended in an insecure area). This robot is equipped with a 3D *perception system* (PS) which can segment potential objects from table-top backgrounds as *object clusters*. It can also use a pre-trained classifiers to assign a distribution over

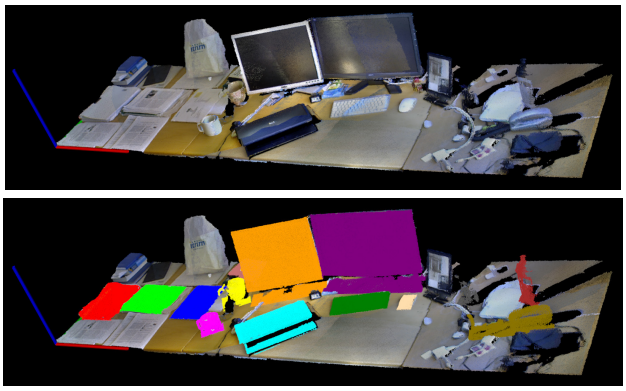

Figure 1. *One table-top scene from the dataset (top), along with its corresponding annotations (bottom).*

category labels to each cluster (Aldoma et al. 2012). The task we now address is – how can the spatial properties of the observed object clusters be used to improve the performance of the robot on this categorisation task beyond the baseline provided by the unaugmented classifier? We assume that the robot has access to a database of labelled table-top scenes, as would be captured on its previous patrols of the same offices, and must create a model from this data which it can exploit for the labelling.

To provide a detailed exploration of approaches for representing and reasoning with spatial context, we remove the robot from the design and evaluation of techniques, and instead use KTH-3D-TOTAL in place of the observations that would have, in principle, been gathered by the robot on previous patrols. Evaluating on a dataset rather than a real robot allows us to vary parameters in the experimental configuration (e.g. initial classifier labelling accuracy) in order to consider a broader range of conditions that would otherwise be possible. However, our evaluation setup (Section 5.1) is constrained to conditions that reflect the operation of the full robot system (e.g. the order in which the observations are gathered) (Kunze et al. 2014).

For our experiments, we focus on the problem of *joint object classification*; this is the task in which a set of detected objects (at least 2 objects) are assigned class labels simultaneously, based on extracted features which could be single object features, relational features, co-occurrence features etc. It addresses the question of - "What is the best group of label-fits, *collectively*, to a group of objects?" rather than "What is the collection of best label-fits to individual objects, all taken *separately* in the same group of objects?" In all our experiments we are interested in observing the performance of our methods (described in Section 4) on joint object classification using only spatial relational features (metric or qualitative). At this point we would like to emphasize the meanings of two important terms w.l.o.g for the methods described later (in Section 4)

- *Relatum*: This is the object which is a *landmark* or with respect to which a spatial relation could be measured.
- *Referant*: This is the object to which a spatial relation is measured with respect to the relatum.

For example, if we measure: "*the mug is 22cm shorter than the monitor*" then, the spatial relation is "*shorter than*"; "*monitor*" is the relatum and "*mug*" is the referant.

3.1 Dataset

To enable us to develop and test our techniques for spatial-relation based scene understanding, we have used a dataset of real world table-top scenes called KTH-3D-TOTAL (Thippur et al. 2014). This dataset contains views of the same set of tables captured repeatedly, across intervals of a few hours, to over many days. There are some examples of the table-tops from the dataset in Figure (2). Clearly, the configuration of the table has a lot of variation between scenes and between different instances of the same scene. Despite the credible variations in instances and structure, there are still inherent contextual/semantic structures that humans recognise – which is exactly what we want to model and exploit.

The dataset contains the individual and group variation in object position and pose due to human interaction. Each scene in the dataset is a registered point cloud obtained from a continuous scan and hence contains a richer and more robust snapshot of the table-top. These scenes contain manually annotated objects on the table-tops with multiple instances from about 10 object categories per scene. The objects belong to the following super set of categories- {Mouse, Keyboard, Monitor, Laptop, Cellphone, Keys, Headphones, Telephone, Pencil, Eraser, Notebook, Papers, Book, Pen, Highlighter, Marker, Folder, Pen-Stand, Lamp, Mug, Flask, Glass, Jug, Bottle}.

4 Metric and Qualitative Spatial Context

The following sections present two approaches for learning spatial relation models (SRMs) from observations of object configurations. The first approach (Section 4.1) is based on metric information, the second (Section 4.2) abstracts from metric information to qualitative relations.

Whilst the metric approach is based on state-of-the-art features, it produces a model which is useful only for joint object classification or collection of single object classifications. In contrast, the qualitative approach creates a representation which can be used for a broader range of spatial tasks: supervised object search, quick structure learning, and even including language grounding and communicating for knowledge transfer. However, its coarser representational structure may prevent it from making highly accurate distinctions during object classification. This can be understood better by seeing QSR as coarser or more discrete versions of the actual measurement based MSR. The classifier trained on QSR could thus be limited in accuracy because of nature of the range of features available for it to discriminate upon.

4.1 Metric Spatial Relations

Metric spatial relations (MSR) considered here are features extracted based on relationships between actual measurements observed about the objects.

Joint object classification is performed using a voting scheme based strategy that captures the spatial and semantic

Figure 2. *Column-1 (L to R) shows the robot ROSIE (SCITOS G5 platform with additional 3D sensors) that is being used for deploying and testing our current systems. Column-2 shows the table of the same person with changes in arrangement between the morning and evening on the same day. Column-3 shows the table configurations of a different person 12 days apart (Thippur et al. 2014).*

coherence of object arrangements in an indoor scene environment by exploring the spatial relations of objects. Depending on what combination of relatums and referants are considered and the spatial relations between them, we may quite possibly obtain many label assignments to the same referant with respect to the different relatums considered. However, these label assignments can be weighted and these count as weighted votes which, as detailed below and in (Alberti, Folkesson, and Jensfelt 2014), determine the final label assignment to the referant.

Features and Spatial Relation Based Features: To model the object categories and the relationships between pairs of object categories, we use the sets proposed in (Alberti, Folkesson, and Jensfelt 2014) to capture the object geometry and the spatial distribution of objects in the scene. *Object pair features* represent the pairwise spatial distribution of the objects, f_{o_i,o_j} as: Euclidean distance between object centroids and its projection in the X-Y plane; bearing between the two object centroids; ratio of object volumes; vertical displacement between object centroids.

Learning Spatial Models: In the training phase, the relationship of the different object category pairs are modelled by applying a Gaussian Mixture Model on the multi-dimensional feature space of object pair features set.

The Voting Scheme: In the inference phase, a voting scheme is applied and a score $Score_A(o_i, c_p)$, is computed for the assignment of each test object, o_i, to each of the possible categories, c_p, based on the spatial relations with the other objects. $Score_A(o_i, c_p)$ is computed as the sum of *pairwise scores* that involve the considered assignment:

$$Score_A(o_i, c_p) = \sum_{\substack{j\in\{1,\ldots,n\}\\ j\neq i}} \sum_{\substack{q\in\{1,\ldots,m\}\\ q\neq p}} Score_P((o_i, c_p), (o_j, c_q)), \quad (1)$$

where n is the number of test objects and m is the number of object categories. The *pairwise score* is defined as:

$$Score_P((o_i, c_p), (o_j, c_q)) = Score_{OP}((o_i, o_j), (c_p, c_q)) \quad (2)$$

The score $Score_{OP}((o_i, o_j), (c_p, c_q))$ takes into account the likelihood value of the category pair model – given the extracted features, corresponding to the conditional probability of the features – given the trained models. The confidence or probability value provided by a vision-based PS, $C_{perc}(o_i, c_p)$, is also considered when it is available, in the following manner:

$$Score_{OP}((o_i, o_j), (c_p, c_q)) = p(f_{o_i,o_j}|c_p, c_q) \cdot \frac{max(1, N_{c_p,c_q})}{(1 + N_{tot})}, \quad (3)$$

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

4.2 QSR-based techniques

Qualitative relational approaches abstract away the geometric information of a scene such as relative angles, relative distances, and relative sizes, and instead represent a scene using first-order predicates such as *left-of*, *close-to*, and *smaller-than*. Our work first generates these first-order predicates from geometric descriptions, then builds a probabilistic model to reason about the class of an object, without knowing the geometric grounding of the state.

Qualitative Relations In this work we adopt a semi-supervised approach to produce a symbolic description of a geometric configuration constructed from 12 predicates: 4 directional, 3 distance, 3 size and 2 projective. We chose these directional and distance predicates as they seem linguistically most common and are suggested as sufficient to

describe the world (Freeman 1975). Unless mentioned otherwise, the predicates are defined using arbitrary thresholds which were hand tuned. Also the predicates are independent of object category. For instance, *close-to* relationship w.r.t Monitor *is not* different from close-to w.r.t Mouse.

Directional predicates are created using the *ternary point calculus* (Moratz, Nebel, and Freksa 2003). The three positions in the calculus are the *origin*, *relatum* and *referent*. In this work, origin corresponds to the position of the *robot*, relatum to a *landmark* object, and the referent to another *object* under consideration. *Robot* and *landmark* define the reference axis which partitions the surrounding horizontal space. Then, the spatial relation is defined by the partition in which *object* lies with respect to the reference axis. In order to determine the partition, i.e. the directional relation, we calculate the relative angle ϕ_{rel} as follows:

$$\phi_{rel} = \tan^{-1} \frac{y_{obj} - y_{land}}{x_{robj} - x_{land}} - \tan^{-1} \frac{y_{land} - y_{robot}}{x_{land} - x_{robot}} \quad (4)$$

ϕ_{rel}, is the angle between the reference axis, defined by *robot* and *landmark*, and the *object* point. Depending on this angle we assign directional relations (*behind, in-front-of, left-of, right-of*) to pairs of objects. When multiple assignments (such as front-right) are possible the predicate with the most score is assigned and the rest discarded for this experiment. This scoring scheme can also be used to maintain such ambiguous assignments. Distributions can be built using QSR scores independently in every predicate dimension.

Distance relations are determined by clustering observed geometric examples. A training set of object scenes is used to derive cluster boundaries between a previously defined number of clusters, each of which will correspond to a qualitative relation. Based on the membership of a geometric relation to a cluster, the associated qualitative predicate is then assigned to a pair of objects. In our technique we use three different predicates: *very-close-to, close-to, distant-to*.

Size predicates compare dimensions of two objects independently along each axis leading to the three predicates *shorter-than*, *narrower-than*, and *thinner-than*.

Projective connectivity between two objects uses Allen's interval calculus (Allen and Allen 1983) on the projection of the objects' axis-aligned bounding boxes onto the x or y axis. The *overlaps* predicate is then extracted for each axis.

In every scene, given a set of objects, every object is considered as a landmark in turns and all remaining objects are considered for extracting measurements for all of these predicates.

Probabilistic QSR-based Reasoning Our objective is to infer the types of all objects given a symbolic scene description

$$S = C_1 \wedge C_2 \wedge ... \wedge C_n \quad (5)$$

where C_n is a clause in a description comprising of a relation predicate R between two objects O_A and O_B:

$$C_n = (R \;\; O_A \;\; O_B), \quad (6)$$

for example `(shorter-than object15 object7)`.

To achieve this we formulate the problem probabilistically: from a training set of scene descriptions for which object types are labelled, we use the occurrence count for each relation to estimate the probability that it will hold given the object types of its arguments:

$$p(R_n^{AB}|L_A, L_B) = \frac{N_{R_n,L_A,L_B}}{N_{L_A,L_B}} \quad (7)$$

where R_n^{AB} is one of the 12 symbolic relations between two objects O_A and O_B with class labels L_A, L_B, N_{L_A,L_B} is the number of times that objects of types L_A and L_B have co-occurred across all training scenes, and N_{R_n,L_A,L_B} is the number of times that relation R_n has occurred between objects of types L_A and L_B across all training scenes.

Then, given a new scene description S for which the object types are only known from perception with a certain confidence, we find all object labels simultaneously. Assuming that one relation holding is independent of another holding, we can apply Bayes theorem successively to find the labels of all objects:

$$p(L|R_1, R_2...R_n) \propto \prod_{i=1..n} p(R_i|L)p(L) \quad (8)$$

where L is a vector of class labels for the objects in S, and R_i is the ith relation in S. The prior probability of the labels $p(L)$ comes from the robot's perception model:

$$p(L) = \prod_{i=1..n} p(L_n) \quad (9)$$

where all n object class labels are independent and provided with their confidences $p(L_n)$.

Finding the optimum class labelling estimate $\hat{L}$ for the objects is then equivalent to finding the maximum posterior in Equation (8). To avoid computational arithmetic problems when dealing with very small unnormalised probabilities, we replace the product in Equation 8 with a sum of logarithms:

$$\hat{L} = \arg\max_L \sum_{i=1..n} \log p(R_i|L) + \log p(L) \quad (10)$$

We performed this maximisation using gradient ascent.

4.3 Spatial Relations Example

Consider the many instances of mice (referant) on table-tops w.r.t their corresponding keyboards (relatum) assuming that the robot (origin) is facing the table just as a human would sit at the table and work.

For every instance of the object-pair, MSR and QSR features be extracted. A MSR-based feature vector would be comprised of real numbers from the actual geometric calculations for distance, orientations, volume ratios, overlap measures etc. In contrast, a QSR-based feature vector would have every dimension range in $[0, 1]$ reflecting the extent of alignment with the QSR.

Using this type of data we could model a distribution of these feature vectors for both SRMs for "mice w.r.t keyboard". For the MSR-based system, this is characterized by the parameters of the learnt Gaussian Mixture Model

(means and covariances), whereas for the QSR-based system, it could be a discrete distribution. An example of such a distribution could be mouse w.r.t a keyboard is 0.75 right-of, 0.20 in-front-of, 0.03 for left-of and 0.02 behind-of.

5 Experimental Evaluation

5.1 Experimental Setup

To compare the above approaches we chose the task of improving object labelling using spatial contexts (Metric or QSR) and we use the following experimental setup. The annotated dataset is split into training and test sets as described below. For the test data we use a simulated PS to assign class labels to objects along with confidence values. We can configure this simulated PS with a perceptual accuracy percentage, which describes the percentage of objects which are assigned the correct label. We varied this percentage to explore how different perception systems will benefit from our work. We also varied the percentage of the available training data (TP) we used to train our models, to explore how sensitive the approaches are to data availability – as this is crucial in online learning applications (such as our long-term autonomous patrolling robot).

Using this setup we performed two experiments with different foldings of the data: *leave-one-out-foldings* and *single-table-foldings*. The leave-one-out-foldings experiments evaluate how our approaches operate on unseen tables (the ones left out) after training on all other tables. This is to replicate the condition of a trained robot encountering a new table, a likely situation in our application. The single-table-foldings experiments evaluate how our approaches perform only on a single person's table and more importantly – less number of training samples. This is also an important use case on our robot, where data from individual tables can be separated by position in a global map. For the leave-one-out-foldings experiments we split the data 70/30 into train and test sets (i.e. 14/6 tables or 330/141 scenes), performing 4 folds, and in each fold the 6 left out tables were randomly selected. For single-table-foldings we split the data 60/40, working with approximately 18/12 scenes per table, with results averaged over 6 tables. In all cases we tested with raw perception system accuracy values of 0%, 25%, 50%, 75% and 100% and training percentage (TP) values of 10%, 20% ... 100% (of the assigned training set). When the accuracy of PS is 100% it means that the perception system classifies the object to the correct class all the time with 100% confidence. When the accuracy of PS is 0% it means that the classification is random with 100% confidence.

For each experiment we apply both the metric spatial model from Section (4.1) (labelled MSR below) and the qualitative models from Section (4.2). As discussed above, the MSR-based technique extracts features which can accept real or continuous values such as orientation angle, euclidean distance etc. of the object w.r.t a landmark and hence utilises a suitable inference mechanism based on Gaussian Mixture Models. However, the QSR-based techniques can be perceived as systems that operate on features obtained according to pre-defined discretizations of the corresponding type of features used by the MSR-based technique. Therefore the QSR-based model requires a more appropriate inference system, in this case *probabilistic reasoning*.

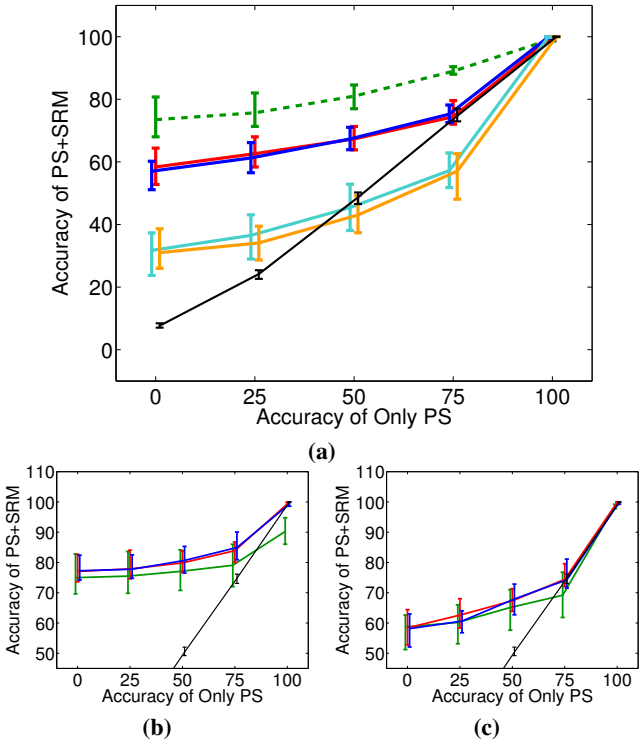

Figure 3. *Leave-one-out-foldings Experiments: Comparison of our SRMs with different testing parameters. (a) MSR(●)-dotted line, TDRC(●), TDR(●), TD(●) and T(●) techniques trained at 60% of training fold. (b) MSR-based technique at different TPs. (c) TDRC technique at different TPs. TP = 10%(●) 60%(●) 100%(●) of 330 scenes – in (b,c). The error bars in all, capture the* max *and* min *value of the PS+SRM system accuracy when checked for different foldings. Raw PS system accuracy is in (●).*

For the QSR-based techniques we explore the effects of different combinations of qualitative relations. The QSR techniques were successively expanded in order to function on larger feature sets by including one extra type of more complex relation in every successive version of the technique. Individual relations are labelled as follows: ternary point calculus (T), qualitative distance (D), relative size (R), projective connectivity (C). Combinations of labels reflect the combination of relations that were employed (e.g. T uses only ternary point calculus QSR and TDRC uses all relations). Similarly, the MSR-based technique was tested with different combinations of the features; However, we present only the best comparable version of the MSR-based technique which works with all the features considered. To keep fair contest between the techniques, the final set of features for the MSR-based technique are: relative angle, relative distance, relative size along the 3 axes and projected connectivity along the 2 axes which correspond to the feature set used by the QSR techniques.

5.2 Results and Analysis

The plots in Figure (3) are the results of the leave-one-out-foldings experiments. Figure (3a) presents the comparison

of all our techniques with respect to changes in the accuracy of the raw PS. The results show that for low perceptual accuracy, all approaches offer an improvement beyond perception, but as perceptual accuracy increases the effect lessens, with only the MSR-based technique offering any improvement when the perceptual accuracy reaches 75%.

The different techniques presented encode different types and amounts of information about scene context. The MSR-based technique operates on more continuous features and hence more detail thereby commensurately makes the most improvement over raw PS compared to the QSR system, given a substantial amount of training data. As additional relations (and thus information) are added to the qualitative relational approaches, a corresponding performance increase occurs, although it appears that the connectivity relation does not have any effect on the results.

Spatial information alone is sufficient to achieve a useful classification accuracy of 50% or higher for some techniques (MSR, TDR(C)) – look at the performances in the leave-one-out-foldings experiments at accuracy of PS at 0%. At 100% all of the techniques are implemented such that they are completely overshadowed by PS and do not affect the PS+SRM accuracy. This is not the case for the T(D) approaches at 50% and 75% perceptual accuracy, where they actually reduce the combined result below raw PS.

Notice that the accuracy of the TDR and TDRC techniques are better than T or TD techniques by about 10-20% consistently. The T and TD techniques use direction and distance only and with the variety of configurations in present in the dataset, and the few coarse features used, it becomes very hard to come up with distinguishing models for different objects based only on these. For example, the distribution of mugs, books and papers w.r.t monitor would look similar if only T and D predicates are used. Adding the three predicates for size removes this ambiguity because mugs, books and papers always have a well defined set of dimensions compared to each other. Then again the connectivity predicates do not add too much more of a distinguishing capacity because they also, with many examples, can have similar looking distributions of measures. This result also makes it evident that the main discriminative feature comes from the object size relations. This occurs because even though the accuracy of the perception system accuracy has increased the accuracy of MSR and QSR systems have not increased as they access training data of similar quantity and quality.

Figures (3b) and (3c) hence demonstrate the influence of TP on our techniques. However, there is very small incremental change in performance as the training set ranges from 10% to 100% of the available training data. This experiment shows that even though we have increased the number of training data from about 35 scenes to 330 scenes the effect of availability of data is minimal. This is because the multiple instances of most of the object categories occurring in these scenes make the training of the relational models quite flat. For instance, the amount of spread of occurrences of mugs with respect to monitors becomes so wide that training a Gaussian Mixture Model for modelling that distribution yields one mixture component with a very large footprint. This might even confuse the SRM and degrade the performance of the classifier.

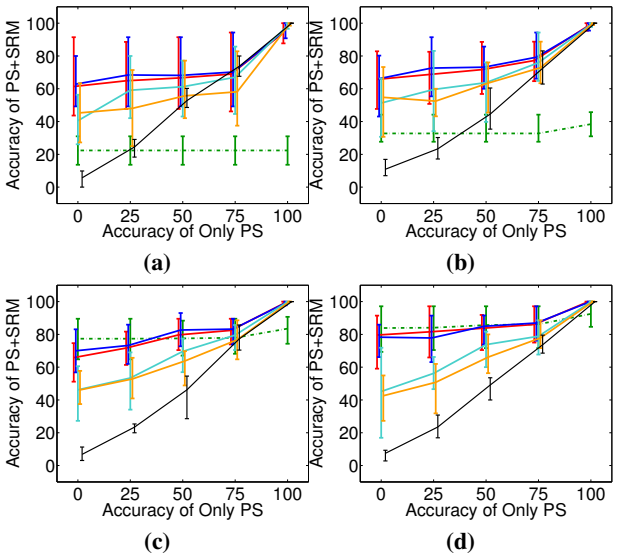

Figure 4. *Single-table-foldings Experiments: Comparison of our SRMs with different testing parameters. MSR(●)-dotted line, TDRC(●), TDR(●), TD(●) and T(●) techniques trained at TP of (a) 10% (b) 30% (c) 70% (d) 100% of training data which is 36 scenes. The error bars capture the* max *and* min *value of the PS+SRM system accuracy when checked for different tables. Raw PS system accuracy is in (●).*

Figure (4), shows the results of the single-table-foldings experiments. As these were performed on only data from a single table, much less data was available in each case (100% TP is 36 scenes). When only very few training samples are available the MSR-based model is unable to generalise, whereas the QSR models perform at about 50% (Figures 4a, 4b), benefiting from the natural generalisation contained in their coarser representation (e.g. *right-of*, *nearer-to* which are ranges of values available to the MSR-based technique)(Section 6). In cases where more data is available (Figures 4c, 4d), the results show a similar pattern to the leave-one-out-foldings experiments, with the more detailed model of MSR-based technique outperforming the QSR techniques.

6 Discussion

The experiment tests the SRM techniques in an object classification task, differing mainly in the variety of features they utilise. The features used by the MSR-based technique are mainly extracted from raw data on which spatial contextual models are built. However, the QSR techniques use human defined feature sets with which we can well define the scenes in the world. They "extract" features by calculating the alignment of the environment instance with these feature definitions. In other words, the QSR techniques "learn how likely it is that a given qualitative structure matches with the world through the creation of a probabilistic model. This is learning structure, but not actually learning spatial structure.

Let us take a look at the main advantages of QSR over MSR. For an extreme example, consider a single instance of mug with respect to the monitor. Let us say we know:

its centroid location (for MSR) and the fact that it is *right-of* monitor (for QSR)(having previously defined this spatial relation *right-of*). Subsequently, let us develop separate generic models using both of these techniques. The MSR-based model learns "all mugs are always at *that particular* single point and pose w.r.t the monitor" because of the single data point available. The QSR technique learns "all mugs are *right-of* monitor" and *right-of* defines a region (containing an inherent generalisation) with respect to the monitor and not a single point. Hence the QSR system can accommodate to estimate the position of a mug in a new scene (assuming that majority of the people are right handed and the mug is usually to the right of the monitor on tables of right handed people). It is thus evident that the latter will be able to generalise with just a few data points, because of the generalisation contained in QSR feature *right-of* in contrast to the MSR-based technique.

The MSR-based technique learns from features capable of higher precision (more continuous values in comparison to the QSR technique) and this needs a substantial amount of data (about 40-50 instances) before a generic model can be successfully learned. The features used by the QSR techniques have an inherent generalisation and thus with very low amounts (about 3-6 instances) of available data they are still able to generate spatial relations models which generalise crudely enough for lower accuracy functionalities (hypotheses bootstrapping).

The MSR-based technique, with sufficient data, (assuming the data it has observed is drawn from the same distribution it is yet to observe) will always perform better than the QSR technique, because it will build a more specific model that is best suited for the kind of environments it has seen and is expecting to see. The QSR technique will build a good model but still confound itself with the generalisations contained in the definitions of the features. The QSR features, though a boon for generalising in situations of less data, act as unneeded hurdles when there is sufficient data to build more accurate models. In case we are uncertain if the training data is a good representation of the test data, then the QSR techniques can be expected to offer a comparatively more robust performance than the MSR-based techniques because many outliers get subsumed within the coarseness of representation in the QSR features which could otherwise hurt the training of MSR-based models.

We think the main situation where pre-defined, linguistically suitable QSR are useful are when we suspect there is structure present that the robot should learn, but we don't have the data to support it yet (Section 5.2). QSR are also of utility when the robot should report its knowledge to a human and when it needs to use human guidance for understanding structure in the environment.

In summary, these results give us indications of what to use when a real robot must start from little or no knowledge about the environment and bootstrap by learning online from its observations. This suggests that a long-term autonomous environment-generic robot could begin operating and learning using a QSR-based technique out-of-the-box and gradually adopt an MSR-based technique, once it (MSR-based technique) is ready to deliver robustly and better than the QSR-based technique. Then again, this is very application specific.

QSR-based techniques have an edge over MSR-based techniques if there is a need to transfer knowledge: *robot* $\rightarrow$ *human* or *human* $\rightarrow$ *robot*. We strongly believe that when qualitative knowledge (in contrast to quantitative knowledge) is shared between robots, in the scenario that they have very little training data - then the inherent generalisations in such descriptions lead to better generalising capabilities. Thus, they (QSR techniques) might be better even in the *robot* $\rightarrow$ *robot* transfer of knowledge scenario. For example, 5 table-top scenes from robot-operation-site A and robot-operation-site B each could help generalise better about topological models than actual metric measurements of all the object configurations from these 10 scenes.

7 Acronyms and Descriptions

Acronym	Description
MSR	Metric Spatial Relations
QSR	Qualitative Spatial Relations
SRM	Spatial Relation Model, which could be based on MSR or QSR
PS	Vision-based Perception System. This is unaided by any SRMs.
TP	Training Percentage. This is the percentage of available data used as Training Data in the experiments. e.g. 10%, 70%
T	(Ternary point calculus) based QSR technique.
TD	(Ternary point calculus + Qualitative distance) based QSR technique.
TDR	(Ternary point calculus + Qualitative distance + Relative size) based QSR technique.
TDRC	(Ternary point calculus + Qualitative distance + Relative size + Projective Connectivity) based QSR technique.

8 Conclusions

We presented two techniques for learning spatial context from observations of collections of objects, and for using this learnt context to improve the performance of a perception system on an object classification task. Our techniques were evaluated on a long-term 3D table-top dataset. The results showed that spatial context knowledge can be used to improve classification results beyond that of raw perception systems (i.e. 3D-vision-based object classifiers which operates only on visual cues extracted from point cloud images). Results also showed that different models can play different

roles in a robot system: more complex metric models using learnt relational features appear to have better performance when enough training data is available to allow them to generalise, but coarser qualitative relational models perform when only few training samples are available and the robot needs to start functioning and learning in an online manner. However, when there is need for any kind of knowledge transfer, QSR-based techniques could be more efficient. In the future we plan to extend this research to beyond tabletop scenes to full rooms, over longer periods of time, and evaluate similar techniques in an online, active learning setting on the robot operating in real-world scenarios. We are also interested in delving the actual structure learning problem that stems from using such SRM-based description of scenes.

9 Acknowledgements

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, the Swedish Research Council and the Swedish Foundation for Strategic Research through its Centre for Autonomous Systems.

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*.

Aldoma, A.; Marton, Z.-C.; Tombari, F.; Wohlkinger, W.; Potthast, C.; Zeisl, B.; Rusu, R. B.; and Gedikli, S. 2012. Using the point cloud library for 3d object recognition and 6dof pose estimation. *IEEE Robotics & Automation Magazine* September 2012:12.

Allen, J. F., and Allen, L. F. 1983. Maintaining knowledge about temporal intervals. *Communication of ACM* 832–843.

Aydemir, A.; Sjoo, K.; Folkesson, J.; Pronobis, A.; and Jensfelt, P. 2011. Search in the real world: Active visual object search based on spatial relations. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2818–2824.

Behera, A.; Cohn, A. G.; and Hogg, D. C. 2012. Workflow activity monitoring using dynamics of pair-wise qualitative spatial relations. In *Advances in Multimedia Modeling*. Springer. 196–209.

Choi, M. J.; Lim, J.; Torralba, A.; and Willsky, A. 2010. Exploiting hierarchical context on a large database of object categories. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 129–136.

Divvala, S.; Hoiem, D.; Hays, J.; Efros, A.; and Hebert, M. 2009. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1271–1278.

Dubba, K. S. R.; Cohn, A. G.; and Hogg, D. C. 2010. Event model learning from complex videos using ilp. In *Proc. ECAI*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, 93–98. IOS Press.

Fisher, M.; Ritchie, D.; Savva, M.; Funkhouser, T.; and Hanrahan, P. 2012. Example-based synthesis of 3d object arrangements. *ACM Trans. Graph.* 31(6):135:1–135:11.

Fisher, M.; Savva, M.; and Hanrahan, P. 2011. Characterizing structural relationships in scenes using graph kernels. *ACM Trans. Graph.* 30(4):34:1–34:12.

Freeman, J. 1975. The modelling of spatial relations. *Computer Graphics and Image Processing* 4(2):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*.

Koppula, H. S.; Anand, A.; Joachims, T.; and Saxena, A. 2011. Semantic labeling of 3d point clouds for indoor scenes. In *Advances in Neural Information Processing Systems*, 244–252.

Kunze, L.; Burbridge, C.; Alberti, M.; Thippur, A.; Folkesson, J.; Jensfelt, P.; and Hawes, N. 2014. Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Ladicky, L.; Russell, C.; Kohli, P.; and Torr, P. 2013. Inference methods for crfs with co-occurrence statistics. *International Journal of Computer Vision* 103(2):213–225.

Li, L.-J.; Su, H.; Lim, Y.; and Fei-Fei, L. 2012. Objects as attributes for scene classification. In Kutulakos, K., ed., *Trends and Topics in Computer Vision*, volume 6553 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 57–69.

Lin, D.; Fidler, S.; and Urtasun, R. 2013. Holistic scene understanding for 3d object detection with rgbd cameras. *ICCV, December*.

Moratz, R.; Nebel, B.; and Freksa, C. 2003. Qualitative spatial reasoning about relative position. *Spatial cognition III* 1034–1034.

Ruiz-del Solar, J.; Loncomilla, P.; and Saavedra, M. 2013. A bayesian framework for informed search using convolutions between observation likelihoods and spatial relation masks. In *Advanced Robotics (ICAR), 2013 16th International Conference on*, 1–8.

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

Thippur, A.; Ambrus, R.; Folkesson, J.; Jensfelt, P.; Burgo, A.; Agrawal, G.; Ramesh, J.; Jha, M.; Akhil, M.; and Shetty, N. 2014. Kth-3d-total: A 3d dataset for discovering spatial structures for long-term autonomous learning. In *(To appear (may the force be with us)) Control Automation Robotics Vision (ICARCV), 2014 13th International Conference on*.

Xiao, J.; Russell, B. C.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2012. Basic level scene understanding: From labels to structure and beyond. In *SIGGRAPH Asia 2012 Technical Briefs*, SA '12, 36:1–36:4. New York, NY, USA: ACM.