A Comparison of Spatial Relation Models for Scene Understanding

Abstract

Object recognition approaches can be unreliable when run in isolation, 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 also contribute a novel, large, periodic, long-term dataset of 20 3D table-top scenes, collected over 19 days, three times a day and manually annotated with 18 object classes. 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.

1 Introduction

Objects pervade human environments. 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 the extraction features to be matched against models built through machine learning techniques. As - the number of objects a given system is trained to recognise - increases, the uncertainty of individual recognition results tends to increase as greater number of objects increases the chance of overlapping features existence. The reliability of such recognisers is also affected when used on 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 unseen 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 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 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 arms 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 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 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 novel data set of table-top scenes (described in Section 3), in this paper we explore the performance of a variety of representations for relative object position, plus inference techniquess for operating on them, on the task of table-top scene understanding (Section 5). In particular we investigate representations that use varying forms of spatial relations, from geometric ones such as distances and angles to more qualitative spatial relations 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 inference

on object spatial configurations using spatial relations. (2) An evaluation of the use of these mechanisms for augmenting a robot's vision based *perceptual system* (PS). (3) A new large 3D annotated table-top benchmark dataset.

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. 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.

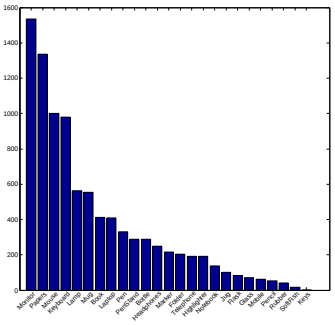
Apart from using the statistics of co-occurrence, a lot of information can be exploited from *how* the objects co-occur in the scene, i.e. the extrinsic, geometric spatial relations between the objects. Recent work in 3D semantic labelling has used such geometric information along with descriptive intrinsic 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. Whilst our techniques are comparable to those in the literature, 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 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.

Our work is evaluated on a new 3D long-term dataset. Other datasets exist: The *B3DO dataset* (Janoch et al. 2011) which contains many single-snapshot instances of indoor human environments having a variety in viewpoints, object-classes, scene-classes and instances. This dataset is in the form of RGB and depth image pairs with manual 2D annotations of object classes, capturing many unique scenes with the sole aim of finding more realistic scenes which are difficult for PSs to perform scene classification. *NYU Depth V1-2* (Nathan Silberman and Fergus 2012) datasets contain different instance examples of object-classes and scene-classes. Each image instance is a combo of synchronous RGB and D images of a different scene-class with semantic annotation provided to every pixel. This dataset is aimed at helping PSs with automatic semantic segmentation and scene classification. The *3D IKEA database* (Swadzba and Wachsmuth 2012) has been collected using robotic maneuvering in different scene-class instances. The aim is to test scene-classification algorithms based on large furniture level objects. The *WRGBD dataset* (Lai et al. 2011) is aimed to support object classification methods and contains many instances of isolated objects in .pcd format. Annotation is done by assigning every pixel a correct semantic label.None of these datasets contain periodically collected data or easily usable spatial annotations of objects which are key for long-term autonomous scene-learning, based on spatial relations.

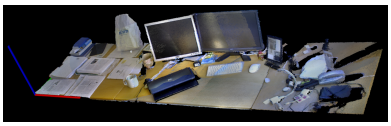
3 Dataset

To enable us to develop and test our techniquess for spatial-relation-based scene understanding, we have created a dataset of table-top scenes. To reflect our interest in long-term autonomy, this dataset contains views of the same tables captured periodically, across intervals of a few hours, over many days. The dataset therefore captures the individual and group variation in object position and pose due to human interaction. The required regularity in instances and time was the main motivation for the construction of this dataset, as currently available datasets either are of individual objects or singular instances of entire rooms.

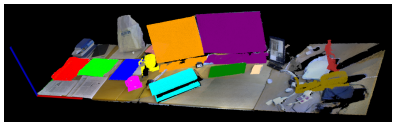
The dataset is a collection of human-annotated office table scenes from a computer science research institute. The data was collected using the *SCENECT* software (Buerkler 2012) and an *Asus Xtion Pro* RGB-D camera. Within the data, a *Scene* is defined to be a single instance of a table-top of one person at a single instance in time. There is one 3D colour point cloud per scene (.pcd format). Every scene is a reconstructed version of the raw data stream obtained by a person scanning a table-top with real time visual feedback



(a)



(b)



(c)

Figure 1: (a) Objects annotated in 3D long-term dataset, sorted in descending order of count of occurrences. X-axis=Object Name, Y-axis=Occurrence Count. (b) Screenshot of one table scene, along with it's annotations in (c).

using SCENECT. The software has in-built real-time sampling, registration and de-noising algorithms to output the final high resolution point cloud.

The scenes were recorded as periodically as possible and at three fixed time instances of the day: *Morning* (09:00 hrs), *Afternoon* (13:00 hrs) and *Evening* (18:00 hrs). Scenes for 20 different tables were collected for 19 days, 3 times per day. A *Scene_ID* is attached to each scene to indicate who the table belongs to and the date and time of the recording. These Scene_IDs help in slicing across the dataset with respect to time of the day {Morning, Afternoon, Evening}, person {Anna, Bob, Carl, ...}, or day {2013-11-01, 2013-11-06, 2013-11-13, ...}.

A 3D annotation tool was developed for manually segmenting out objects of interest from the point clouds. On average, 12 different objects were labelled per scene. The objects belong to the following super set - {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}. The information about every scene and object is available in XML and JSON formats. Each scene has a nested list of object data containing {Position, Orientation, Size, Date and Time of recording, Person ID, Point Indices of the point cloud that have been labelled as belonging to the Object}.

4 Evaluation Scenario

Our evaluation scenario stems from our interest in long-term autonomy in real environments. Our techniques will run on a mobile robot capable of patrolling a collection of offices a couple of times during the day[1]. 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 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 the new dataset described above 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 6.1) is constrained to conditions that reflect the operation of the full robot system (e.g. the order in which the observations are gathered).

5 Metric and Qualitative Spatial Context

The following sections present two approaches for learning spatial relation models (SRM) from observations of object configurations. The first approach (Section 5.1) is based on metric information, the second (Section 5.2) 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 object class inference. The qualitative approach creates a representation which can be reused for a broader a range of spatial tasks (including grounding language), but its coarser representational structure may prevent it from making certain distinctions during inference.

5.1 Metric Spatial Relations

Joint object classification is performed using a voting scheme based strategy that captures the spatial and semantic coherence of object arrangements in an indoor scene environment by exploring the spatial relations of objects.

[1]Whilst we have the majority of this system implemented and running, we exclude details here for blind review and space constraints

Spatial 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. *Single object features* (SOF) f_{o_i}, where o_i is the i^{th} object, are computed from the 3D spatial characteristics of the object w.r.t. a reference frame and origin (here the table and its front-left corner). The set features consists of: the projected lengths of the object along the X, Y and Z reference axes; 3D coordinates of the object centroid; horizontal bearing of the object centroid from front-left table corner. *Object pair features* (OPF) represent the pairwise spatial distribution of the objects, f_{o_i,o_j} as: Euclidean distance between object centroids and it's 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, a set of models for each of the object class categories are learned by using a Gaussian Mixture Model (GMM)-based representation to encode the multivariate probability distribution of *SOF*. The relationship of the different object category pairs are modelled by applying a GMM on the multi-dimensional feature space of *OPF* 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 reference system and with the other objects and on typical object occurrence and co-occurrences. $Score_A(o_i, c_p)$ is computed as the sum of *pairwise scores* that involve the considered assignment:

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

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

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

The scores $Score_{SO}(o_i, c_p)$ and $Score_{OP}((o_i, o_j), (c_p, c_q))$ take into account the likelihood values of the category models and the likelihood value of the category pair model – given the extracted features, corresponding to the conditional probability of the features – given the trained models. Additionally, the scores integrate, as a-priori weights, the occurrence probability of the individual object categories and the co-occurrence probability of the object category pairs, estimated using frequency counts on the training database. The confidence or probability value provided by a vision based PS, $C_{perc}(o_i, c_p)$, is also considered when it is available, as follows:

$$Score_{SO}(o_i, c_p) = Pr(f_{o_i}|c_p) \cdot \frac{max(1, N_{c_p})}{(1 + N_{tot})} \cdot C_{perc}(o_i, c_p), \quad (3)$$

$$Score_{OP}((o_i, o_j), (c_p, c_q)) = Pr(f_{o_i,o_j}|c_p, c_q) \cdot \frac{max(1, N_{c_p,c_q})}{(1 + N_{tot})}, \quad (4)$$

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

5.2 QSR-based technique

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 generating spatial relation predicates which combines a geometric calculus with clustering methods. This produces a symbolic description of a geometric configuration constructed from 12 predicates: 4 directional, 3 distance, 3 size and 2 projective.

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. In the following we denote these positions by *robot*, *landmark*, and *object*. *Robot* and *landmark* define the reference axis which partitions the surrounding 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:

$$\phi_{rel} = \tan^{-1} \frac{y_{obj} - y_{land}}{x_{robj} - x_{land}} - \tan^{-1} \frac{y_{land} - y_{robot}}{x_{land} - x_{robot}} \quad (5)$$

ϕ_{rel}, is the angle between the reference axis, defined by *robot* and *landmark*, and the *object* point. Dependent on this angle we assign directional relations (*behind, in-front-of, left-of, right-of*) to pairs of objects.

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 the bounding box dimensions of two objects. Each axis is compared individually, leading to the three predicates *shorter-than*, *narrower-than*, and *thinner-than*.

Projective connectivity between two objects uses 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.

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

$$S = C_1 \wedge C_2 \wedge ... \wedge C_n \tag{6}$$

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

$$C_n = (R \;\; O_A \;\; O_B), \tag{7}$$

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:

$$p(R_n^{AB}|L_A, L_B) = \frac{N_{R_n,L_A,L_B}}{N_{L_A,L_B}} \tag{8}$$

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 recursively to find the labels of all objects:

$$p(L|R_1, R_2...R_n) \propto \prod_{i=1..n} p(R_i|L)p(L) \tag{9}$$

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:

$$p(L) = \prod_{i=1..n} p(L_n) \tag{10}$$

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 9. To avoid computational arithmetic problems when dealing with very small unnormalised probabilities, we replace the product in Equation 9 with a sum of logarithms:

$$\hat{L} = \arg\max_L \sum_{i=1..n} \log p(R_i|L) \log p(L) \tag{11}$$

We performed this maximisation using gradient ascent.

6 Experimental Evaluation

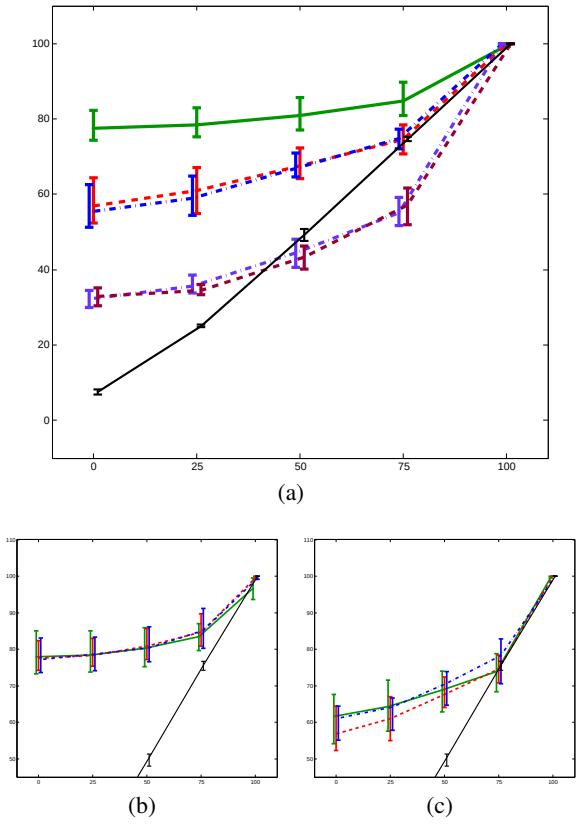
6.1 Experimental Setup

To compare the above approaches on the task of improving object labelling using spatial context 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 (PA), which describes the percentage of objects which are assigned the correct label. We varied this percentage to explore how different perceptual 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* (LOOF) and *single table foldings* (STF). The LOOF 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 STF experiments evaluate how our approaches perform on a single person's table over time. 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 LOOF experiments we split the data 70/30 into train and test sets (14/6 tables or 688/295 scenes), performing 4 folds, each fold the 6 left out tables randomly selected. For STF we split the data 60/40, working with 36/24 scenes per table, with results averaged over 6 tables. In all cases we tested PA values of 0%, 25%, 50%, 75% and 100%, and TP values of 10%, 60% and 100% (of the assigned training set). For each experiment we apply both the metric spatial model from Section 5.1 (labelled SM below) and the qualitative model from Section 5.2. For this latter technique we explore the effects of different combinations of qualitative relations. 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. TDRC uses all relations).

6.2 Results and Analysis

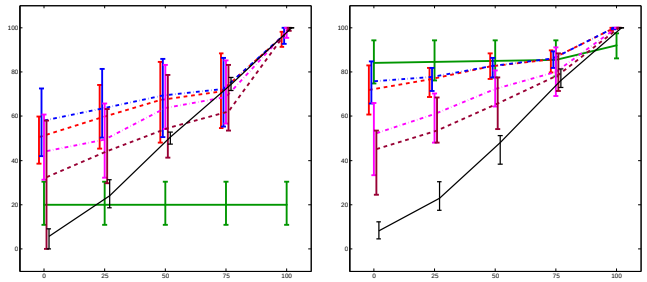
Figure 2a presents the comparison of all our techniques with respect to changes in the accuracy of the perceptual system (PA). The results show that for low perceptual accuracy, all approaches offer an improvement beyond perception, but as perceptual accuracy increases the effect lessens, with only



(a) (b) (c)

Figure 2: Comparison of our spatial relation models (SRM) with different testing parameters. X-axis=PA, Y-axis=PS+SRM accuracy. (a) SM(●), TDRC(●), TDR(●), TD(●) and T(●) techniquess trained at 60% of training fold.(b) SM technique at different TPs. (c) TDRC technique at different TPs. TP = 10%(●) 60%(●) 100%(●) 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 (●).

the SM approach offering any improvement when the perceptual accuracy reaches 75%. Our different techniques encode different amounts of information about scene context. SM has the most information (including spatial relations plus size-object size and location features) and commensurately makes the most improvement over raw perception. 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 data. With perceptual accuracy at 0% it is clear that spatial information alone is sufficient to achieve reasonable classification accuracy for some techniques (SM, TDR(C)), and at 100% all of the techniques correctly trust perception and do not reduce the combined 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 perception. Figures 2b and 2c demonstrate the influence of the amount of training data on our techniques (TP). There is very small



(a) (b)

Figure 3: Comparison of our spatial relation models (SRM) with different testing parameters. X-axis=PA, Y-axis=PS+SRM accuracy. (a) SM(●), TDRC(●), TDR(●), TD(●) and T(●) techniquess trained at 60% of training data and (b) 10% of training data. 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 (●).

incremental change in performance as the training set ranges from 10% to 100% of the available training data.

Figure 3a, shows the results of the STF 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 (Figure 3b) the SM model is unable to generalise, whereas the QSR models perform adequately, benefiting from the natural generalisation of their coarser representation. In cases where mode data is available, the results show a similar pattern to the LOOF experiments, with the more detailed model of SM outperforming the QSR approaches. Overall, these are good results for the real robot where it must start from little or no knowledge and learn online from its observations. It seems reasonable that the robot could start from a QSR-based model while gathering initial data, before switching to an SM model when its performance starts to overtake that of QSRs.

7 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 perceptual system on an object classification task. Our techniques were evaluated on a novel 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, and that different models can play different roles in a robot system: more complex metric models using both single-object and relational features appear to ultimately have better performance when enough training data is available to allow them to generalise, but coarser qualitative relational still perform when only few training samples are available. In the future we plan to extend this research to beyond table-top scenes to full rooms, and evaluate similar techniques in an online, active learning setting to the real robot.

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.

Buerkler, O. 2012. Scenect - faro technologies : http://www.faro.com/scenect.

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.

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.

Janoch, A.; Karayev, S.; Jia, Y.; Barron, J.; Fritz, M.; Saenko, K.; and Darrell, T. 2011. A category-level 3-d object dataset: Putting the kinect to work. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 1168–1174.

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.

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.

Lai, K.; Bo, L.; Ren, X.; and Fox, D. 2011. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 1817–1824.

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.

Nathan Silberman, Derek Hoiem, P. K., and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *ECCV*.

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

Swadzba, A., and Wachsmuth, S. 2012. A detailed analysis of a new 3d spatial feature vector for indoor scene classification. *Robotics and Autonomous Systems*.

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.