Preparation of Papers for IEEE Sponsored Conferences & Symposia*

Akshaya Thippur[1], Rares Ambrus[1], Kaushik Desai[1], Adria Gallart[1], Malepati Sai Akhil[2], Gaurav Agrawal[2], Mayank Jha[2], Janardhan HR[2], Nishan Shetty[2], Prasad NR[2], John Folkesson [1] and Patric Jensfelt[1]

***Abstract*—**

Humans subconsciously exploit various strong correlations amidst different object instances, classes and between different object classes and scene types when analysing indoor environments. Correlations in naive, logical object co-occurrences have been exploited along with the extraction of vision based object-intrinsic descriptive features in previous research. In this paper, we present several alternative learning techniques to model and make estimates of scenes based on a variety of spatial relations - geometric extrinsic features with different amounts of discretization, which capture *how* the objects co-occur; and compare their efficacy in the context of object classification in real-world table-top scenes. We investigate the possibilities of using such techniques to refine the results from a traditional vision-perception system. We also contribute a unique, long-term periodic, large 3D dataset of 20 office table-top scenes, manually annotated with 18 object classes. Apart from our current comparison, we foresee that the dataset will be useful for applications such as generalized learning of spatial models, learning object data structuring based on semantic hierarchy, learning best suited semantic abstractions and grammar for long term autonomy, ground truth for vision-perception systems etc.

I. 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 IV), 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 **??**). 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

*This work was supported by STRANDS

[1] KTH Royal Institute of Technology albert.author@papercept.net

[2] MSRIT Bangalore b.d.researcher@ieee.org

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.

Points discussed:

1) **Make this more "Human Oriented"**
2) Overview of the problem - why?
   a) Not only desktops
   b) How do we treat object classification
3) Generalize and Transfer Knowledge
4) How are objects correlated, search becomes easier
5) Understanding a model for inter object influence for inference
6) No benchmarks yet or datasets having 3D spatial information
   - time
   - complete scenes
   - different people
   - different types of people
7) need to structure data – metric is tedious, because of large amounts of data
8) Contributions:
   - A big 3D data set, annotated – benchmark for results, folding data, classification.
   - suggest SR and QSR
   - something that could augment perception systems

II. Related Work

Spatial relations have been used previously to provide contextual information to vision-related work. [**?**]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, [**?**] demonstrate the learning of activity phases in airport videos using spatial relations between tracked objects, and [**?**] 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 [**?**]; and the use of object presence to provide context in activity recognition [**?**]. 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 [**?**]. 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 [**?**]. 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 [**?**]. 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 [**?**]. In [**?**] spatial relations between smaller objects, furniture and locations is used for pruning in object search problems in human environments. In [**?**] 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 [**?**] 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 [**?**], [**?**] 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* [**?**] 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* [**?**] 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* [**?**] 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* [**?**] 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.

1) More on dataset papers

III. Motivating Example

We are investigating systems that operate for a long time in environments populated by humans. As a motivating example we will look at security guard in an office building. The

robot patrols the working environment and should learn the allowed variations of different parts of space. In this work we focus on desktops the robot must look at table-tops of the office inhabitants and learn about the allowed variations of the objects on them. The objects could generally change place, within certain semantic bounds (A chair is generally in front of a table, almost never on a table); some objects could be missing at times (coffee mug) which is normal, but some other objects are rarely moved (monitor). Certain specific instructions might be passed on - like a "clean desk policy" which means that the desk needs to be rid of any paper material post working hours but is allowed otherwise during working hours. All these are many rules and quite specific to circumstances for a programmer to practically specify in the execution code of the robot, hence the need of intelligent robots. The ultimate aim of this aspect of the robot learning would be to be able generalize the learnt knowledge about Table Tops to Rooms or to different kind of Table Tops with objects never seen before. e.g. Learn about table tops in an office environment and be able to generalize to table tops in a cafeteria.

The STRANDS robot, is expected to observe environments continuously and adapt their learnt algorithm based on new allowed variations and/or human reinforcements. It is required to learn general structures of environment settings and object spatial configuration properties and come up with a general structure for say, *Office Room*, *Office Table*, *Cafeteria Table*. At the next level of intelligence it must be able to learn what is normal for particular people in the environment if it is feasible or find generalized structures over time or other different parameters e.g. category of researcher, winter, holiday season etc. At a higher level of semantic mapping, the robot could also learn about the variations in particular scene classes e.g. *Office Rooms*, *Office Kitchen*, *Office Lobby*, *Office Bathroom*.

The project has begun with a focus on Office Tables. The aim is to learn spatial models for general allowed configurations of sets of objects usually found on Office Tables, across variations in time and people. Once there are general spatial models of Office Table, security protocols and action-reaction mechanisms can be specified or learnt. These can also act as a tool to aide vision-perception systems. In an "*Object Search*" task the robot could use the knowledge from the spatial configuration to obtain prior stochastics about where an object could most probably be found. In the "*Object Recognition*" task the learned spatial models could provide prior information to process a portion of the RGB-D image for locating a particular object that is usually present there because of the spatial configurations of the remaining recognized objects, but cannot be currently found because of noise such as: occlusions, sensor infidelity etc. The learnt spatial models can also provide for the posterior probabilities of the recognizer to disambiguate between recognised objects.

The following sections elaborate on mainly two things: the dataset constructed for this purpose and comparisons of different spatial modelling techniques on the developed data to provide suggestive insights on the kind of qualitative, non-descriptive features for human environment modelling.

Points Discussed:

1) To be specific lets look at STRANDS for evaluating concepts
2) What is test case in STRANDS
3) This section should stress why it is important to gather to such datasets
4) what type of data do we really need? - hint it
5) reference to related work section to show that this dataset is unique.

IV. Dataset

It is convenient to perceive a structure of the objects in the environments as if on a "nested table-top system". Table-tops are the most commonly present large objects in any floor map. It is more likely that objects on a table have more correlation amidst them, than when compared to those objects not on that same table e.g. *Pen* has more semantic/functional correlation with a *Laptop* (on the table) than a *Couch* (not on that table). At the next level of semantic abstraction we can imagine all the objects in a room to be on a "table-top" which is the floor. With this kind of perception, the first dataset has been constructed, periodically observing table-tops of researchers at KTH University.

It is required for such a long-term autonomous learning to have entire views of the scenes to model the interrelations amidst the member objects. Apart from being variant over object instances across different scenes, the data needed to be periodic over time at a scale of couple of hours so as to capture the individual and group variations in position and pose when there has been regular/irregular human interaction involved. The currently available datasets either are of individual objects or single instances of entire rooms. The required regularity in instances and time was the main motivation for the construction of this dataset.

A. Design

Points Discussed:

1) Why did we collect it the way we did?
2) What did we collect?
3) Time and people level slicings
4) weekends and weekdays
5) different times of the day
6) We want the data to have these properties and hence we designed the dataset in this way.

B. Dataset Realization

A 3D dataset " *Tables for Spatial Modelling*" has been created by KTH, Royal Institute of Technology. The dataset is a collection of human annotated office tables of researchers at KTH. This is the ground truth data to train models for spatial configuration of Office Tables.

The data was collected using a freely available software called *SCENECT* [] and an *Asus Xtion Pro* RGB-D camera. A *Scene* is defined to be a single instance of a table-top of one person at one time instance. There is one 3D colour image in the form of a point cloud per scene (.pcl 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 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). The scenes were collected for 19 days, 3 times per day and for the same 20 different tables. Depending on who the table belongs to and the date and time of the recording, each table-top scene recording receives a *Scene_ID*. These Scene_IDs help in slicing across the dataset with respect to time of the day {Morning, Afternoon, Evening}, or people{Akshaya, Yuquan, Carl,...}, or day {2013-11-01, 2013-11-06, 2013-11-13,...}.

A *3D Annotation Tool* was developed at KTH for manually segmenting out objects of interest from the point clouds. On an average 12 different objects were labelled per scene. The objects belong to the following super set - {Mouse, Keyboard, Monitor, Laptop, Mobile, Keys, Headphones, Telephone, Pencil, Rubber, Notebook, Papers, Book, Pen, Highlighter, Marker, Folder, Pen-Stand, Lamp, Mug, Flask, Glass, Jug, Bottle}. The information about every scene and each object are available in the .xml and .json formats. Each scene data has a nested list of object data, and each object data has the following information about the object - {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}.

Points Discussed:

1) Annotation Tool
2) Tools for collecting Asus Scenect
3)

C. Dataset Summary

1) Dataset details - days, objects, people, times
2) **figure** showing variations over different times of day, different people
3) details of XML and JSON files
4) **figure** Histograms of object occurrences

V. Analysis

Points Discussed:

1) **figure** Scatter plot for variation within person ID
2) **figure** Scatter plot for variation wrt object type
3) **figure** Scatter plot for variation considering different landmark - trajectors.
4) **figure** Scatter plot of 2D footprints of objects
5) General conclusions about above figures.
6)
7)

VI. Conclusions

VII. Future Work

Points Discussed:

1) Room level data set
2) Which QSR for which purpose?

VIII. Acknowledgements

CVAP-KTH, Accel Partners, STRANDS project