Preparation of Papers for IEEE Sponsored Conferences & Symposia*

Albert Author[1] and Bernard D. Researcher[2]

***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

Robotics is beginning to pervade into human life as a direct consequence of the recent amalgamation of computers into almost all avenues of human life. Research in intelligent robotics is ardently sought after by many current academics because of the utility that our society can expect to obtain from augmenting itself with robots. The main drawback of naive computer programs is that they perform *only* instructed tasks, withal discounting the incredibly utilitarian character of accurate repetitive performance with virtually no signs of fatiguing. The research in intelligent robotics is to tackle this issue, by which, new age robots can learn about human nature, activities, needs, actions, reactions and seamlessly integrate into their environment to make for more efficient living.

With the above concept in mind, consider the need for robots to learn, understand and generalize about spatial configurations of human environments and the many animate and inanimate objects they contain. Ideally, given a collection of supervised observations and provided with the intelligence of one of the relevant state-of-the-art machine learning techniques the problem seems straight forward. However some of these ideal conditions are very hard to fulfil. Consider a human environment viz. office, supermarket, bank, cafeteria etc. the general description of the task of the robot installed there is to check on the working humans, monitor the environment, raise alarms if something out of the ordinary is observed and also react and display proactivity toward human occupations as an ultimate goal. For this to be effected, the robot is required to use the signals from it's sensors, extract semantics from the signals, understand configuration, activities and normality of both from the semantics and ultimately try to gain view of human intent. Other profitable attributes that such robots could have are that of generalization of gained knowledge and transferability of such knowledge to it's peers.

Robots thusly, need to gain knowledge about the objects in the human environment. This primarily entails the recognition of the class of the many objects in the environment in the presence of instance and pose variations and occlusions. Classification of objects in the environment based on structure or function can have different utilities such as passive or active surveillance and activity observations. Post object classification, a scene classification analysis needs to be done which means to consider all the recognized object class instances together to understand the configuration of the scene as a whole. Understanding that the presence and pose of particular object as an effect of the scene in question; understanding that the configuration of objects in a scene are not independent events but is because of interrelated causality amongst the objects in the scene, is quite powerful, to obtain a holistic view for the performance demanded from these robots.

A technical challenge that comes in the way of learning spatial models that represent objects in a human environment encompassing the interrelations, variations over different spans of time and space and instances involves handling a plethora of data received from the sensors. Consider only an RGB-D sensor such as Asus Xtion mounted on a patrolling robot that sweeps the environment once every hour. Even, if a perfect vision system for object recognition is in use, consider the amount of representative metric data generated per object, per scene, per time instance and hence for different instances of the same scene (e.g. Office Desk or Cafeteria Table) over spans of time, grows at a large order. Using this data to get one generalized model of *Office Table* or *Cafeteria Table* capturing the essences of variation over

*This work was not supported by any organization

[1]Albert Author is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands albert.author@papercept.net

[2]Bernard D. Researcheris with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA b.d.researcher@ieee.org

time and instances of the environment and it's components is quite close to impossible without extracting some intelligent features over the objects/scene or mapping them onto a lesser dimensional subspace.

The state-of-the-art stand-alone vision-perception based systems extract descriptive features from images to recognize objects. The characteristic parametrizations of these descriptive features typical and intrinsic to each object is what helps the perception system classify the objects. However, a lot of non-descriptive features are also available but rarely exploited e.g. spatial configurations of a group of different objects which usually occur together, functional information, contextual information, pose and placement with respect to a global origin etc. Such features extracted from object constellations, given their semantic data, are called *Spatial Relations* (SR) or *Qualitative Spatial Relations* (QSR) based on amount of discretization and similarity to human perception. A few such techniques are proposed in this paper to provide openings toward further such research for long-term autonomous learning robots.

These extrinsic or spatial features based systems can be further developed for activity recognition, and understanding the human environment and it's components at a higher level of semantic abstraction. Theses systems that do mainly object or scene-classification could currently work alongside already present vision-perception systems and augment the performance to obtain better accuracy. They could provide prior probabilities for object labels or finding them in a scene. They could also help calculate the posterior probabilities of the object labels to disambiguate a recognition system, once it has actually detected the objects or even labelled them (verification).

Training such long-term intelligent systems to learn about human environments from such non-descriptive features requires large amounts of data. As current robotic platforms usually use RGB-D sensors and perform analyses on 3D data, large amounts of periodic *point-cloud* images of the same scenes over periodic instances of time is required. These point clouds need to be annotated for ground truth data about objects in the environment. Such a large dataset has been developed for one category of scenes and is available for use. This data set is one of it's kind and can be used by the researchers doing vision-perception and/or robotics.

Points discussed:

1) Overview of the problem - why?
   a) Not only desktops
   b) How do we treat object classification
2) Generalize and Transfer Knowledge
3) How are objects correlated, search becomes easier
4) Understanding a model for inter object influence for inference
5) No benchmarks yet or datasets having 3D spatial information
   - time
   - complete scenes
   - different people
   - different types of people
6) need to structure data – metric is tedious, because of large amounts of data
7) 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) QSR - LEEDS activity recognition

III. Test Case / Working Example

The members of this research team are working for a EU-Research project[**?**] by the name of *Spatio-Temporal Representations and Activities for Cognitive Control in Long-Term Scenarios* (STRANDS) which involves developing a long-term autonomous robot with a SCITOS G5 [], with an RGB-D camera as a vision sensor, so that this robot can be used to augment tedious observe-learn-serve tasks in human environments.

Security Guard for Office is a test case for the project and by the end of the project it is required to develop a robotic intelligence which would learn the map of the area to be patrolled, the objects present in the office, the human activities in the office at different times of the day, week, month and year, normality and abnormality about these aspects and alarm raising protocols if abnormalities have been detected.

This translates to an action plan for the robot in which it patrols the working environment and learns the allowed variations of different room-level objects such as chairs, tables, shelves etc. As a nested action plan, 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

- See structure

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. Dataset Details

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}.

B. 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

C. Implementation or execution

Points Discussed:

1) Annotation Tool
2) Tools for collecting Asus Scenect
3)
4)

V. Analysis

Points Discussed:

1) Scatter plots
2) Histograms
3)
4)

VI. Conclusions

VII. Future Work

Points Discussed:

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

VIII. Acknowledgements

Adria, Kaushik Desai, Malepati Sai Akhil, Prasad NR, Gaurav Agrawal, Janardhan N, Mayank Jha, Nishan Shetty, MSRIT Bangalore, CVAP-KTH, Accel Partners, STRANDS project