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], John Folkesson [1] and Patric Jensfelt[1]

***Abstract*—**

AK: Not yet written, this is from the previous paper. 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

AK: Needs shortening, has been modified from AAAI paper. PJ: Shorted and removed some things that are central to this paper. I think that the aim should be that this paper is cited not only for the dataset BUT also for showing, with hard data and not hand waving, that there are structures that can be exploited when building models and performing inference.

In the last decades, computers have become part of human activities in all avenues of life and the field of intelligent autonomous robotics shows signs to follow in a similar way in the near future. In this paper, we concentrate our discussions on understanding and learning about indoor human environments, particularly office environments and suggestively toward understanding about human interactions with such environments.

Our research work is performed in the context of developing a mobile service robot for long-term autonomy in indoor human environments, from offices to hospitals (Section III). 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 learning capabilities. In particular, the robot would expected to learn to perform an assigned task that is repetitive with weaning supervision and human interaction. The contextual knowledge the robot can gain from the repeated attempts can make it learn so that it's subsequent attempts on the same task are improved in accuracy and efficiency. In this paper we study *table-top scene understanding* as a first step to such robotic behaviour. Whilst the objects present on a single table may change in position, their overall arrangement has some regularity over time as influenced by the context and functionality of the table-top. For example, there is a general structure in office employee table-tops to that of cafeteria table-tops or kitchen table- tops. The differences derive from the variety of objects and their group-configurations on the table-tops with respect to time – hours, days, weeks, months etc. For example: Office tables typically have monitors, keyboards, mouse along with papers and pens and coffee mugs configured such that the monitors can be viewed, the keyboard typed at and the mouse easily accessed from the keyboard. The spatial configuration vary in particular instances and the arrangement change gradually over many days and minutely at different times of the day; However, cafeteria tables have cutlery, jugs, food, napkins etc. configurations of which don't vary over many days, weeks or years but vary in content and presence according to different times of every day. It is this structure and it's variants that we aim to exploit in order to improve the understanding of table-top scenes.

Bluntly modelling the absolute positions of objects on tables, is likely to fail because such structure is improbable to generalise across a range of different types and instances of tables over numerous observations w.r.t. time. We are instead investigating qualitative *relational* models of space, i.e. ways of encoding the qualitative position of a target object relative to the position of one or more landmark objects. For example: Given – "A Keyboard is *In Front* of a Monitor"; Depending on the size and span of the table and given knowledge about the functionalities and sizes of monitors and keyboards, humans can generally estimate a probable location for both objects on an unseen table. We believe that representations that are to support long term autonomy and ability to generalize knowledge require such rough discretizations of metric measurement space inherently encoding the generality of structure in the information.

As a mean to study this, we have constructed a novel benchmark 3D data set of office type table-top scenes

*This work was supported by STRANDS

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

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

called ***3D T**able-**T**ops Dataset for Long **T**erm **A**utonomous **L**earning* (3D- TOTAL)[1] observing the same set of select tables, over many times a day and over many days in a month. Objects-of-interest have been manually annotated in 3D to provide ground truth data.

The main contributions of this paper are 3D-TOTAL dataset and the accompanying analysis that clearly show the potential presented by the structures that are inherent in the scenes for building models and for performing inference. The rest of the paper is structured in the following way: We survey the related work in Section II and present an example scenario for in Section III. In Section IV we exhibit 3D-TOTAL in depth and provide an analysis of it in Section V. We summarize and conclude in Section VI.

II. Related Work

AK: Maybe needs shortening. PJ: Edited and shorted a bit.

The dataset constructed in this work is motivated mainly by research pertaining to automatic learning about indoor human environments for long term autonomous robots. This requires certain characteristics of a dataset.

The *B3DO dataset* [1] 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. It captures single snapshots of unique scenes for which scene classification and object classification is difficult for vision based perception systems (VPS). The *NYU Depth V1-2*[2] datasets contain different instance examples of object-classes and scene-classes captured with RGB+D images. Automated pixel clustering is conducted by using features in the RGB and D images separately and annotation of the clusters has been done manually, thereby assigning a semantic label to every pixel. The dataset contains a wide range of singular snapshots of indoor scenes from commercial and residential buildings. This dataset is primarily aimed at evaluating VPS aiming at automatic semantic segmentation and scene classification.

The *3D IKEA database* [3] has been collected using a robot moving in different scene-class instances. The aim is to test scene-classification algorithms based on large, furniture level objects. 3D point clouds are formed by stitching a series of RGB+D images. There are very few small objects in the scenes and the annotations are provided at the scene-class level. The *WRGBD dataset* [4] is aimed to support object classification methods and contains many scene instances of isolated objects in point cloud format. Annotation is done by assigning every pixel a semantic label in each scene. Each point cloud is created from a series of RGB+D images.

The system constructed by [5] can be used to automatically generate 3D datasets of scenes using rough human annotations on 2D images as input. The system infers 3D information from the scene using the semantics of the annotated properties of important planes in the image. The generated dataset thus includes a large set of singular scenes, indoor and outdoor from very particular viewpoints, with annotations to the image components provided manually.

The Kinect@Home project [**?**] has collected a large set of crowd sourced data. People having access to a kinect like sensor have captured and uploaded data from their own environments. The data is in the form of RGB-D video sequences. The data covers a large variety of different scene types and range from single objects to whole rooms. No annotation is availble.

The dataset provided by [6] contains a collection of 3D images of a few table-top objects in clear view and cluttered view. This dataset has been constructed to aid VPS for object classification and segmentation functioning on 3D data. Other datasets that have been developed have mainly been for training VPS for robust indoor object classification on 3D data [7], [8]. However, none of these datasets contain the following three properties in their scenes:

- complete 3D scene instances of a particular scene type (e.g. office table-tops, office rooms, cafeteria tables, hospital ward- rooms etc.)
- instances of subsets of objects-of-interest co-occurring in the scenes – manually annotated for ground truth.
- long term periodic observations of the same set of scenes.

which are key for long-term autonomous scene model learning.

Scene learning and understanding can become a very expensive task if the raw data acquired by the sensors is used in terms of metric measurements. Many recent works have investigated how learning and modelling human environments increase in efficacy if qualitative spatial features are used instead of metric ones. Spatial relations have been used previously to provide contextual information to vision-related work; [9] 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 [10], [11]. Recent work has used object co-occurrence to provide context in visual tasks such as activity recognition [12]. Apart from using the mere statistics of co-occurrence, a lot of information can be exploited from *how* the objects co-occur in the scene. Recent work in 3D semantic labelling has used such geometric information along with descriptive intrinsic appearance features [13]. 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 [14]. In [15] spatial relations between smaller objects, furniture and locations is used for pruning in object search problems in human environments. In [16], [17] the authors utilise both geometric features on objects and spatial relations between objects for scene understanding.

In our future work we aim to provide different representations of spatial context for a novel, long-term, learning task. Additionally the usage of such features leads to increased

[1] http://www.cas.kth.se/data/3D-TOTAL/

compatibility for robots to function using semi-supervised learning and in this way bootstrap the system using expert knowledge from humans.

III. Motivating Scenario

AK: I do not know if Patric wrote this part :-o. PJ: I did but it is not very grand in its current form

We are investigating systems that operate for long periods of time in environments populated by humans. As a motivating scenario we will look at security guard in an office building. The robot patrols the working environment and should learn models of what the environment normally looks like and what variations there are. In an implementation of such a system the robot should be able to raise an alarm when something is not as expected. Initially we will focus on office table-top scenes. We are interested in models for individual desks and as well as general models of desks. Our working hypothesis is that there are some structures that govern how desks are organized. We want to extract and later exploit these when building the models. As mentioned above, we expect most desks to have a monitor, a keyboard and a mouse and that these are arranged in a certain way. Many people will have the mouse ot the right of the keyboard. This is an example of a spatial structure we want to learn. Some people have a laptop on their desk and they bring this laptop home in the evening. This is an example of a dynamic property of a desk that we want to learn.

Another aim is to be able to transfer knowledge from one desk to another and ultimately from one environment to to the next. That is, if the robot encounters a desk that it has not seen before, it should be able to make use of the models other desks to bootstrap the learning. This way, the robot would be able to perform some its functions even after a single observation, making it significantly more useful. The model of a newly observed desk would start out with relatively large uncertainty regarding what can be considered normal for that desk but this would improve over time as the model is adapted by new observations. To summarize, we strive to develop representations that caters for the knowledge transfer, the ability to learn from few samples and adapting existing models.

To study these questions we need data to learn from. The data need to capture both variations across different desks but also over time. None of the datasets available (see Section /refsec:Related Work) meet these requirements which is the motivation for the work behind the dataset that we present in this paper.

IV. Dataset

AK: COMPLETED - needs refining.

Most of the human outdoor and especially indoor environments, are characterized by a supporting surface on which relevant objects are placed e.g. Land - buildings, living room floor - furniture, dining table - cutlery and dishes etc. It is hence convenient to perceive a structure for the object arrangement in the human environments to be as if on a "nested table-top system". Table-tops provide for a favourable prototypical example for analysis of object organization structure.

With the aim of exploring possibilities to understand, learn and model these organizational structures amongst objects in human indoor environments, a first, pertinent dataset called 3D-TOTAL, has been composed by periodically capturing observations of a fixed set of entire table-tops in an office environment. In what follows we define a *scene* as a single observation of a table-top instance at a single instance in time. The scenes have been captured as is[2] with intervals of a few hours, over many days and populated with various instances of object classes. The dataset therefore captures the individual and group variation in object pose due to humans and their regular/irregular interaction with the environment. 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 single instances of tables or entire rooms. This regularity in sampling and extent over time is important when studying modeling interrelations amidst the member objects in a long-term autonomous learning.

A. Dataset Design and Concept

As explained in Section III our research intentions are to provide intelligence to an long-term operating, autonomous robot for indoor human environments. Thus, the concept of composing the 3D-TOTAL dataset follows naturally from this research motivation. The dataset has been composed by capturing and manually annotating 3D point clouds of office type table-tops, for a fixed set of people, at fixed times of the day and for a span of weeks. Observing the table-tops of the same set of people at different times of the day gives insight about the daily interactions a human has with his table and over many days gives an understanding of the gradual variances of their table-tops. If the data is observed for an entire week, including weekends, features in the table-top configurations, that can be used for estimation of the type of the day of the week, can be extracted (e.g. Weekdays, Fridays, Weekends). Table-top models can also be learnt for all the people put together – which gives a gross functional representation of an office table-top in general in office environments – or for individual people which helps to build functional representations of office table-tops for individuals, for different activities and so on (Figure 1). In summary: When the dataset is partitioned in different ways with respect to time, people or object instances, it richly yields knowledge of table-tops in office environments supporting building representations of the same.

B. Dataset Realization

In 3D-TOTAL, 3D point clouds representing 20 people's desks were captured regularly 3 times a day for 19 days. The data was collected by carefully scanning each table-top using an *Asus Xtion Pro Live* RGB-D camera. The raw RGB-D

[2]People have been asked to step aside not to contribute to the occlusion of the table-top.

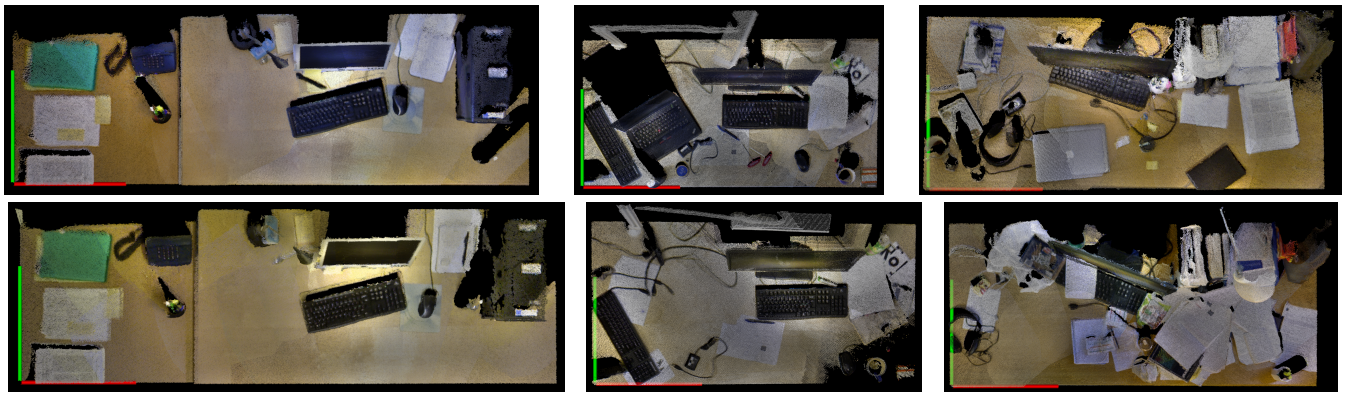


Fig. 1: Each column shows a different person's office table in top view at two different times. The tables in the first two columns are captured in the morning and evening of the same day, whereas the table in the last column is captured 12 days apart. We can see distinct differences between different person's desk but there are also many commonalities that a system should exploit.

data stream is aggregated into a single high resolution point cloud (.pcd format) using the *SCENECT* software [18].

The scenes were recorded as periodically as possible and at three fixed times each day: *Morning* (09:00 hrs), *Afternoon* (13:00 hrs) and *Evening* (18:00 hrs). Scenes contain tables of 20 different people collected over 19 days including weekends. 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 partitioning the dataset with respect to time of the day {Morning, Afternoon, Evening}, person {Carl, Nils, ...}, 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, including repeating instances of the same object class, per scene depending upon feasibility and occurrence. 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 points in the point cloud that have been labelled as belonging to the Object}. This manual annotation provides the required ground truth data for long term autonomous learning.

C. Dataset Summary

This section gives a brief summary of the dataset to serve as a quick reference. 3D-TOTAL has:

- scenes collected from 20 unique tables, 3 times a day for 19 days and hence in total, approximately 1140 scenes.
- each scene manually annotated with 18 possible object classes and an average of 12 object instances, from these classes, annotated per scene.
- has annotation stored in XML and JSON formats containing scene instance and it's object instances' specifications.
- occurrences of object instances as depicted in Figure 2a and annotations as exemplified in Figure 2b,2c

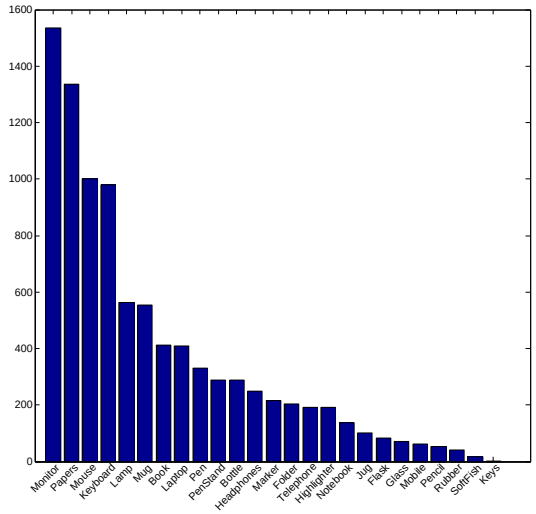
V. Analysis

AK: Patric wrote this part, I am yet to look at it.

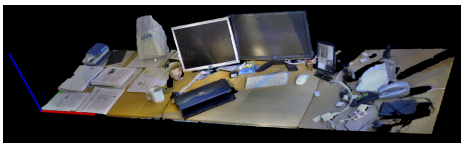
In this section we perform an analysis of the data to highlight some interesting aspects of it. In particular we want to show that there are structures in the data that can be exploited by a system for more efficient representations and better reasoning with less data.

Figure 1 shows three desktop scenes. Each column shows the same desk at two different times. The leftmost two columns contain scenes from the same day, whereas the two scenes in the third column were sampled 12 days apart in time. Notice in Column 1: the slight changes in position of the keyboard, mouse, papers and pen; Column 2: the relatively big changes in position of laptop, mouse, papers, pen, keyboard, lamp. When objects in columns 1,2,3 are compared there is a certain generality in structure (keyboards are in front of monitors), but also a specificity for each person (occurrence of headphones, position of mouse w.r.t. keyboard, etc.). Studying the scenes could also allow a system to infer the activity of people. We can see that there are changes to the first desk suggesting that someone was there during the day and that this person is quite well-ordered. The second table misses the laptop in the second observation. This might suggest that the person has left the desk, maybe for the day. The third desk seems to be occupied by someone that is less sensitive to clutter. With only two observations of each desks these are just speculations but by looking at the data over 19 days stronger claims coud be made.

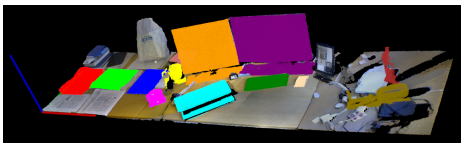
One of our hypotheses is that a qualitative model will be needed to achieve efficient and powerful representations of space, at least if the amount of data is limited as it will be



(a)



(b)



(c)

Fig. 2: (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) Screen-shot of one table scene, along with it's annotations in (c).

in realistic cases. Such qualitative models could allow some of the inherent structure in the environment be encoded in the representation itself. We have already seen in Figure 1 that monitors are typically in the back while a keyboard is usually in front of it and the leftmost table shows an example of a mouse being to the right of a keyboard. Figure 3 shows a scatter plot over the position of keyboards and mice when found in the same scene. The table outline gives an example of a prototypical desk to make it easier to interpret the data. In the top part of the figure, the green circles mark the position of the centroid of each keyboard that exist in a scene where there is at least one mouse. The red square shows the mean position of all these centroids. The black crosses show the position of all mice in scenes with at least one keyboard. In the bottom part of the figure the position of the mouse relative to the keyboard is shown for each observed pair in the data. As expected most mice are qualitatively to the right of the keyboard. There are some outliers. However, encoding the position of the mouse as being to the right of a keyboard would capture most of the information. The largest cluster of

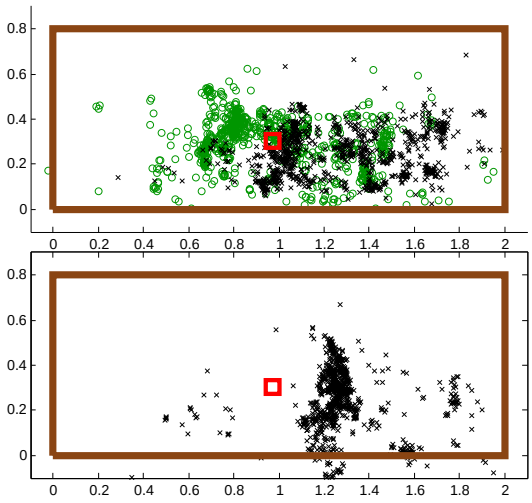


Fig. 3: Top: The green circles shows the positions keyboards(red square shows mean of keyboards). The black crosses show the position of all mice in these scenes. Bottom: Mean position of keyboards (red square) and relative position of mouse relative to it. The brown rectangle gives a rough idea about where the borders of a typical desk would be.

points in the lower part of Figure 3 contains 95% (CHECK THIS!!!!) of all data points. Notice how this structure in the data is lost, at least visually, when looking at the position of the keyboard and mouse in the table frame (top figure) and how it pops out when looking at the relative positions (bottom figure). We want our representation to be able to capitalize of this structure. Clearly, in this case, the position of the keyboard contains almost all information that is needed to represent the position of the mouse as well.

To further investigate the correlation between different object classes, and thus look for other inherent structures in the data, we look at the relative position of all objects of class C_j w.r.t. to objects of class C_i present in the same scene. To get a quick overview of the type of distributions this results in we look at the entropy of these distributions. A large entropy (closer to uniform distribution) would suggest that the objects are largely uncorrelated and a small entropy (a more peaky distribution) a stronger correlation between the objects. We calculate the entropy as

$$E = -\sum(\frac{n_i}{N})ln(\frac{n_i}{N}) \quad (1)$$

where n_i is the number of samples that fall in cell i in a grid discretization of the table and N is the total number of samples. Each sample correspond to one object pair in one scene. The true entropy will only be estimated well when N is large. We therefore limit this investigation to pairs of objects that occur more than a certain number of times in the data. Figure 4 shows these entropies for 10 of the objects in the dataset. From this figure we can, for example, see the low entropy in the relation between keyboard and mouse (element 1,3 and 3,1 in the matrix). The relative positions of monitors and keyboards also have a fairly low entropy. We

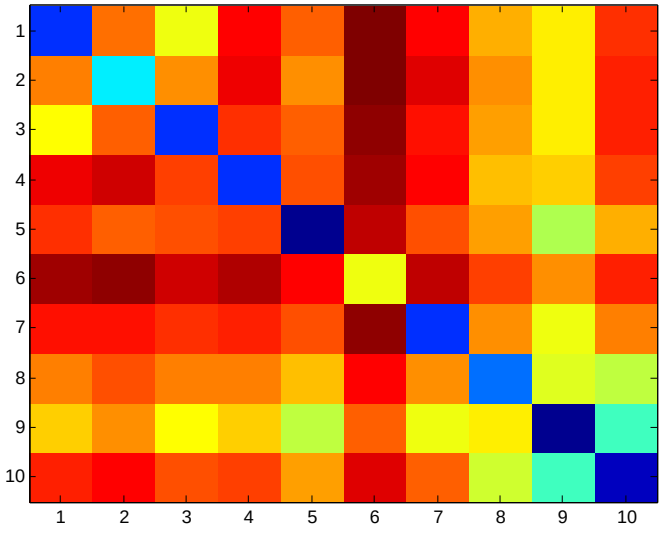


Fig. 4: The figure illustrates the entropy in the distribution of relative positions of one objects (column) w.r.t. to another objects (row). Dark red indicates high entropy (more uniform distr.) and dark blue low entropy (peakier distr.). We show this for a subset of objects from the database; 1:keyboard, 2:monitor, 3:mouse, 4:mug, 5:laptop, 6:papers, 7:book, 8:bottle, 9:jug, 10:notebook.

also see that the position of papers is largely uncorrelated with many other objects (uniform distribution gives high entropy).

In Figure 5 we look closer at some of these relations. Figure 5a shows the position of the keyboard w.r.t. to the monitors. Our intuition that keyboards are placed mostly in front of monitors is supported by data. Figure 5b shows the position of the mug w.r.t the keyboard. We see that the mug is rarely very close to the center of the keyboard but rather positioned around the keyboard. Taking function into account the data might suggest that the mug is in fact often placed at arms length from the person working on the desk to keep it at safe distance from the keyboard but still within reach. We see a bias towards the right side, probably a result of most people being right-handed. In Figure 5c we see that the distribution for the relative position of papers w.r.t. to the keyboard is almost uniform, suggesting that they are largely uncorrelated. The fact that there are often many papers in a scene adds to the uniformity.

To summarize the analysis, we have shown that the data has many structural properties that a method for representing and reasoning about space should exploit. If the aim is to represent typical configuration of objects, this preliminary analysis suggests that a significant part of such knowledge can be encoded well with qualitative spatial relations, such as the mouse is to the right of the keyboard, while keeping in mind that this represents the typical case and not the only possible situation. We can also see that a system that observes these desks for an extended period of time will be able to learn quite a lot about the habits of the owners of the desks and even the current activity in many cases. It is important

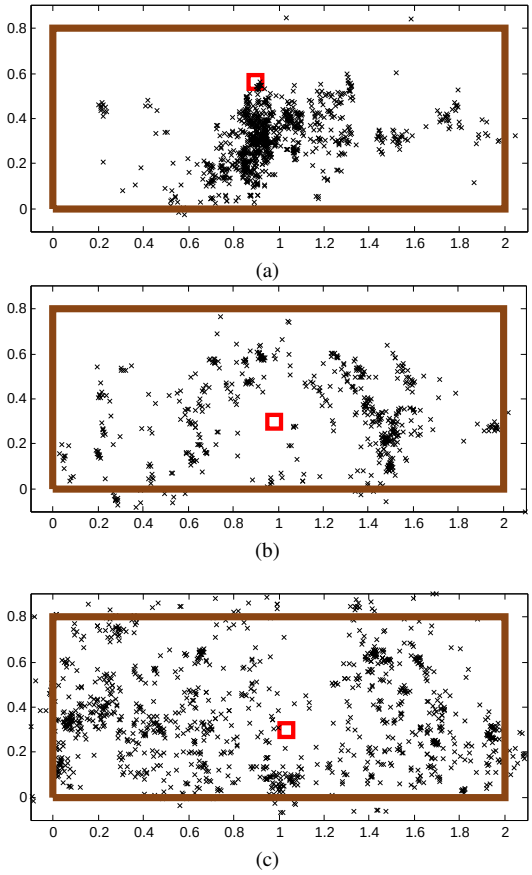


Fig. 5: The figures show the relative position of a) keyboard w.r.t. monitor, b) mug w.r.t. keyboard and c) papers w.r.t. keyboard. Qualitatively keyboards are mostly infront of monitors, mugs are around the keyboard and the position of papers is mostly independent on the position of the keyboard.

to differentiate between typical knowledge, i.e., knowledge about what the world typically is like and specific instance knowledge, i.e., knowledge about a particular scene at a specific time. It is to capture the first kind of knowledge that we believe qualitative spatial representations will be most beneficial.

VI. Conclusions

AK: Not yet written, this is from the previous paper.

VII. Future Work

AK: Not yet written, this is from the previous paper.

VIII. Acknowledgements

AK: Not yet written, put the content in the correct format.
CVAP-KTH, Accel Partners, STRANDS project

References

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

[2] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

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

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

[5] B. Russell and A. Torralba, "Building a database of 3d scenes from user annotations," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2711–2718, June 2009.

[6] M. Sun, G. Bradski, B.-X. Xu, and S. Savarese, "Depth-encoded hough voting for joint object detection and shape recovery," in *Computer Vision ECCV 2010* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), vol. 6315 of *Lecture Notes in Computer Science*, pp. 658–671, Springer Berlin Heidelberg, 2010.

[7] G. Bradski and T. Hong, "Solution in perception challenge.," 2011.

[8] S. Helmer, D. Meger, M. Muja, J. Little, and D. Lowe, "Multiple viewpoint recognition and localization," in *Computer Vision ACCV 2010* (R. Kimmel, R. Klette, and A. Sugimoto, eds.), vol. 6492 of *Lecture Notes in Computer Science*, pp. 464–477, Springer Berlin Heidelberg, 2011.

[9] M. J. Choi, J. Lim, A. Torralba, and A. Willsky, "Exploiting hierarchical context on a large database of object categories," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 129–136, June 2010.

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

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

[12] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as attributes for scene classification," in *Trends and Topics in Computer Vision* (K. Kutulakos, ed.), vol. 6553 of *Lecture Notes in Computer Science*, pp. 57–69, Springer Berlin Heidelberg, 2012.

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

[14] M. Fisher, M. Savva, and P. Hanrahan, "Characterizing structural relationships in scenes using graph kernels," *ACM Trans. Graph.*, vol. 30, pp. 34:1–34:12, July 2011.

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

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

[17] A. Kasper, R. Jakel, and R. Dillmann, "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*, 2011.

[18] O. Buerkler, "Scenect - faro technologies : http://www.faro.com/scenect," 2012.