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*—**

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.

II. Related Work

The dataset constructed by this work is motivated mainly by research pertaining to automatic learning about indoor human environments for long term autonomous, human activity augmenting robots. This requires certain characteristics datasets that were explored for in our survey of state-of-the-art datasets. 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 many singular snapshots of unique scenes with the sole aim of collecting such realistic scenes on which scene classification and object classification is difficult for vision based perception systems (VPS). The objects have been manually annotated in the RGB and D frames correspondingly, using 2D bounding boxes. *NYU Depth V1-2* [2] datasets contain different instance examples of object-classes and scene-classes. Each image instance is a combo of synchronous RGB+D images of a different scene-class, with labelling provided by manual annotation. Moreover, 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 robotic manoeuvring in different scene-class instances. The aim is to test scene-classification algorithms based on large, furniture level objects. These scene data are 3D point clouds formed by stitching a series of time synchronized RGB+D images. There are very few small objects in the scenes and the annotations are provided at the scene-class level. This dataset is aimed to support the research in the area of automatic scene labelling.

The *WRGBD dataset* [4] is aimed to support object classification methods and contains many scene instances of isolated objects in .pcd format. Annotation is done by assigning every pixel a correct semantic label in each scene. Each point cloud is developed from a series of synchronized 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. The system infers 3D information from the scene using the semantics of the annotated properties of the important planes in the image. The 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 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.

Long-term autonomous characteristic and spatial human indoor scene understanding is the aim of our immediate research work (Section III). Scene learning and understanding can become a very expensive task if the raw data grabbed by the sensors is used in terms of metric measurements, as available. Many recent researchers have tried to show that learning and modelling human environments increase in efficacy if the scene is perceived by not just using metric descriptive features extracted on the scene components but a lot also from the geometric, qualitative spatial features extracted from them. 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 compatibility for robots to function using semi-supervised learning and in this way bootstrap the system using expert knowledge from humans.

III. Motivating Scenario

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 tell when something differs from the ordinary too much and then raise an alarm. Initially we will focus on desktop scenes. We are interested in models for individual desks and as well as general models of desks. Our working hypotheses is that there is some general rules for how desks are organized that we want to be able to extract and later exploit when building the models. We expect that object will change place, within certain semantic bounds. Some objects could be missing at times (coffee mug) which is normal, but some other objects are rarely moved (monitor).

Another aim is to be able to transfer knowledge from one environment to the next. This would allow a robot that just entered a new environment to be functional from the start. Concretely this would correspond to having a reasonable prior which can then be adapted when new observations are available. What information is general and what is environment specific? How do we represent the knowledge in a way that caters for the knowledge transfer, the ability to learn from few samples and adapting existing models? These are some further examples of questions that we want to study.

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.

Points Discussed:

1) This section should stress why it is important to gather to such datasets
2) what type of data do we really need? - hint it
3) reference to related work section to show that this dataset is unique.

IV. DATASET

Most of the human outdoor and especially indoor environments, are characterized by a supporting surface on which all 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 are the most commonly present large objects in any floor map. They provide for a favourable prototypical example for analysis of object organization structure. Moreover, it is highly likely that objects on a ”table” have more semantic and organizational 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, here the ”table” references to the floor of the room).

With the aim of exploring possibilities to understand, learn and model these organizational formalities amongst objects in human indoor environments, a first, pertinent dataset called ***3D T****able-***T****ops Dataset for Long* ***T****erm* ***A****utonomous* ***L****earning* (3D-TOTAL), has been composed by periodically capturing observations of entire table-tops of a fixed set of researchers at a computer science research facility. A singular observation of a table-top of one person at a single instance in time, captured in image format, is termed as a *scene*. These scenes have been captured with intervals of few hours, over many days and various instances of object classes. The dataset therefore captures the individual and group variation in object position and pose due to humans and their 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 singular instances of entire rooms.

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 (Section II).

A. Dataset Design and Concept

The target research groups to benefit from this dataset are of the kind that develops artificial intelligence for autonomous robots augmenting human activities (especially if they are fatiguing) within indoor human environments. Hence, it becomes essential for robotic learning to pay attention to variances in human environments wrt. scales of time, space and instances in the same space. The dataset has been composed by capturing and manually annotating 3D images of office type table-tops, for a fixed set of people, at fixed times of the day and for a span of many days.

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 in their table-top setups. 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 a typical table-top in general for research employees in office environments – or for individual people which helps to gain functional representations of office table-tops for individuals, seniority, gender and so on (Figure 1). Finally, when trying to find general models over any of these types of data partition, the model learns to be able to generalize over different instances of a fixed set of object classes. In summary: When the dataset is partitioned in different ways with respect to

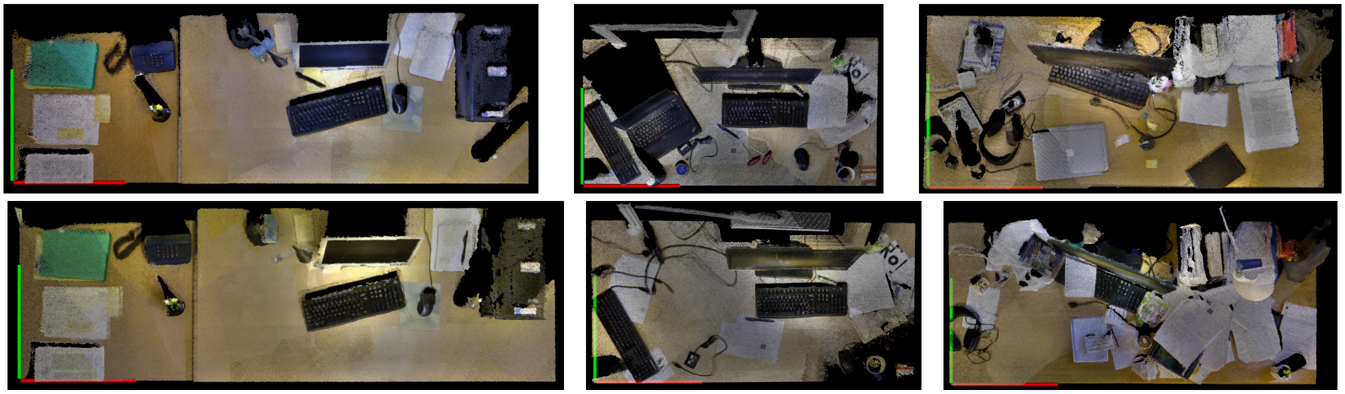


Fig. 1: Each column shows a different person's office table in top view at two different times. The tables in the first two columns are captured in the morning and evening of the same day, whereas the table in the last column is captured 12 says apart. We can see distinct differences between different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

time, people or instances, it richly yields knowledge and hence representations of table-tops in office environments.

As explained in Section III our research intentions are to provide intelligence to an long-term operating, autonomous, human activity augmenting robot for indoor human environments. Thusly, the concept of composing the 3D-TOTAL dataset follows naturally from this research motivation.

B. Dataset Realization

In 3D-TOTAL, 3D images of scenes were captured regularly at 3 times a day for 19 days for 20 people. Each scene has been manually annotated to obtain information about common objects-of-interest generally and regularly present, when observed across the many scenes.

The data was collected using the *SCENECT* software [18] and an *Asus Xtion Pro* RGB-D camera. There is one 3D colour point cloud per scene (.pcd format). Every scene is a reconstructed version of the raw data stream obtained by manual detailed scanning of 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). 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 {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, 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 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

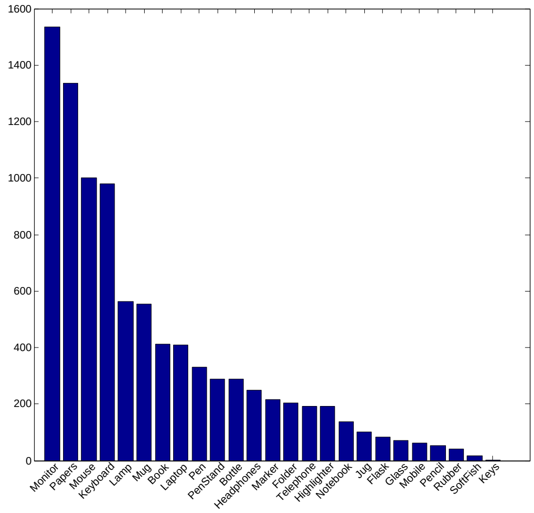
This section gives a brief summary of the dataset to serve as a quick reference.
3D-TOTAL has:

- 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

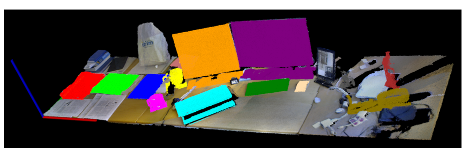
In this section we perform a first analysis of the data to highlight some ainteresting spects of it. In particular we want to show that there are structures in the data that can be exploited by a system for more efficient representations and better reasoning using less data.

Figure 1 shows three desktop scenes. Each column shows a desk at two different times. The leftmost two columns contain scenes from the same day, whereas the third column shows scenes 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



(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) Screenshot of one table scene, along with it's annotations in (c).

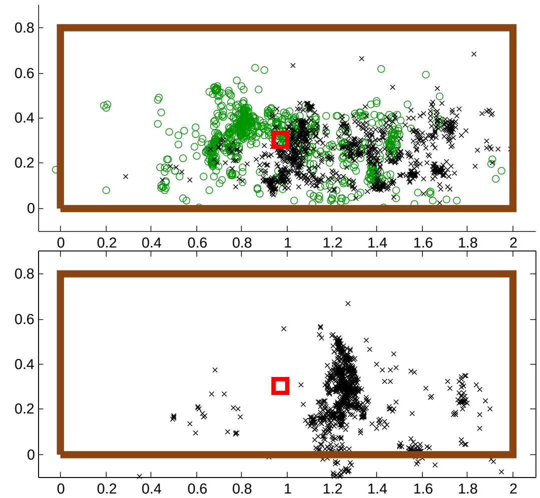


Fig. 3: Top: The green circles shows the positions keyboards(red square shows mean of keyboards). The black crosses show the position of all mice in these scenes. Bottom: Mean position of keyboards (red square) and relative position of mouse relative to it.

laptop, mouse, papers, pen, keyboard, lamp. When objects in columns 1,2,3 are compared there is a certain generality in structure (keyboards are always in front of monitors), but also a specificity for each person (occurrence of headphones, position of mouse wrt. keyboard). Studying the scenes could allow a system to infer things related to behaviour of people as well as activity. We can see that there are changes to the first desk suggesting that someone was there during and that ths person is quite well-ordered. The second table misses the laptop in the second observation. This suggest that the person has left the desk, maybe for the day if it is late. The third desk seems to be occupied by someone that is less sensitive to clutter.

As mentioned before, 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 in most 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 the keyboard is in front of it. The leftmost table shows an example of a mouse being to the right of a keyboard. This is something that we would also expect to hold in many cases. Figure 3 shows a scatter plot over the position of keyboard and mouse. 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 shows 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. As expected most mice are qualitatively to the right of the keyboard. There are some outliers but simply encoding the position of the mouse as being to the right of a keyboard would capture most of the information. 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 relatiev positions (bottom figure). We want our representation to be able to capitalize of this structure and 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 another 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 distrubution and likely 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 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 played mostly infront of monitors is supported by data. In 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 along a circle around the keyboard. Taking function into account in the analysis could 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 fromthe keyboard but still within reach. We see a bias towards the right side, a result of most people being right-handed. In Figure 5c we see that the position of paper is almost completely uncorrelated with the position of the keyboard.

To summarize the analysis, we have shown that the data has many structural properties that a method for representing and reasoning about pace should make use of. If the aim is to represent typical configuration of objects, this preliminary analysis suggest 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 is 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 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.

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)

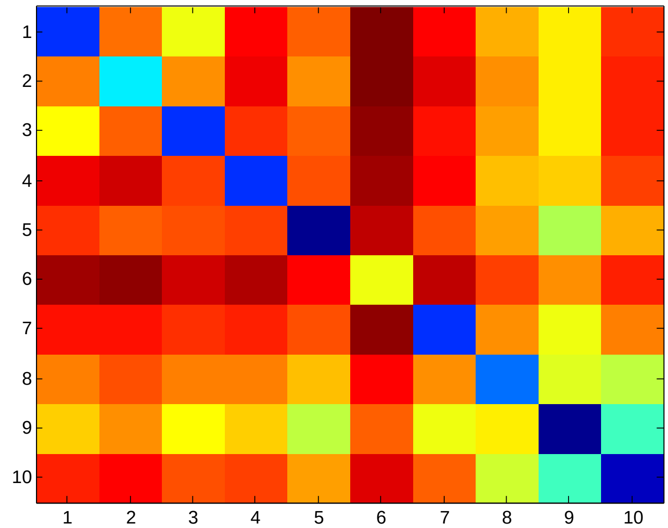


Fig. 4: The figure illustrates the entropy in the distribution of relative positions of one objects (column) w.r.t. to another objects (row). Dark red indicates high entropy (more uniform distr.) and dark blue low entropy (peakier distr.). We show this for a subset of objects from the database; 1:keyboard,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

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

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.

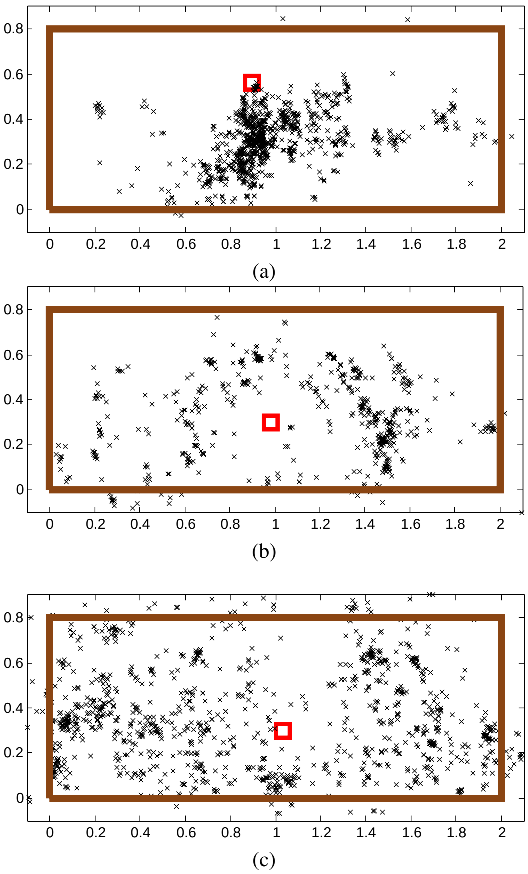


Fig. 5: The figures show the relative position of a) keyboard w.r.t. monitor, b) mug w.r.t. keyboard and c) papers w.r.t. keyboard. Qualitatively keyboards are mostly infront of monitors, mugs are around the keyboard and the position of papers is mostly independent on the position of the keyboard.

[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.