Using spatial relations of objects in real world scenes for scene structuring and scene understanding

Alexander Kasper and Rainer Jäkel and Rüdiger Dillmann

Abstract— **Given a room full of individual objects in a generic household scene, one can observe that the objects are mostly not placed randomly but in a certain order. Because of this each object can be described by the surrounding objects and the spatial relations to those objects. This paper presents several types of spatial relationships that can be deduced using object positions in 3D as well as an approach to retrieve these relations from real world scenes via annotation of colored 3D pointclouds gathered with a sensor. Finally, a way to use this data to make predictions about an unknown object based on its surrounding objects is presented.**

I. INTRODUCTION

When observing man made everyday environments one can find a big variety of different objects that belong to various object classes in a multitude of sizes and shapes. Despite this huge variety in most cases a certain ordering of things can be found. In our work we concentrate on scenes that a service robot might encounter in a typical household. In most cases this will be a specific room in a house or a part of a room like a dining table and its near vicinity. The type of scene already defines a subset of objects that can be expected to be seen and may also infer a certain arrangement of the objects. For example if we look at a typical office we will see a desk with at least one chair close to it and probably a computer monitor, keyboard and mouse as well as a telephone all located on the desk. There can also be drawers with folders and books and a trash can. For a human this kind of knowledge is trivial because he learns this during growing up by observing a vast number of similar scenes and learning about the functional connections between the objects of a given scene type. In this work we try to create an empirical basis for this kind of knowledge about scenes and the contained objects by analyzing real world examples thus deriving relational information about objects and searching for statistical measurements, like appearance probabilities or average object sizes, that contain the most information about the scene and the objects therein. The paper is structured as follows: first of all we present similar works in the area to give a context for our work. After that, a short definition of the terms used in this paper is given. Furthermore, the acquisition of the real world data is shown, which is then the basis for the following annotation process. Based on the annotated data we propose a set of relations and statistical measurements that are promising candidates to reproduce the intuitive scene knowledge of a human. Using these relations and the training data, an approach for using the data to predict the class of an unknown object based on the surrounding objects is presented. Finally some results are provided that illustrate the validity of the approach.

II. RELATED WORK

Using annotated data sets to increase performance of classification is certainly not new and there exist a multitude of publications, especially in the field of object recognition and localization, that use an annotated set of examples to deduce various features about the annotated objects. The annotation in those works is usually a contour drawn around objects in the image. Russel et al. [7] have implemented a special application that can be used to annotate images in this fashion which is then used to build a database of object classes. Image data that is annotated in this way allows the automatic extraction of features that describe this specific instance of the object to recognize it later in a different image or categorize it. But not only features of the objects itself are used, some works also analyze the spatial relations of the various objects in the image to allow for a better categorization or recognition [3]. For 2D data the annotation of scenes by marking contained objects is a wide spread method. In 3D data however this is not the case. Especially for whole scenes containing more than one object this process to generate empirical data is rarely used. Fisher and Hanrahan [4] exploit statistical data about the spatial relations between objects in artificial 3D scenes to improve the search for 3D models. They use a large number of manually modeled scenes from Google Warehouse and analyze the geometric structure of the scenes. For every scene, the Euclidean distance of object centers between all objects of the scene is calculated which finally provides a statistical distribution of distances to other objects. These distributions are then later used to limit the search space for a query that consists of a volume in a given scene and a keyword. This publication shows that the structure of a type of scene is at least to some extent encoded in the spatial relations between the objects of the scene. While the artificial nature of the scenes allowed the authors to easily analyze a large number of scenes it is not sure how well the results map to real world scenes. The authors also only use the Euclidean distance between objects centers while other relations could certainly yield interesting results as well.

A. Kasper is with Institute of Anthropomatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany alexander.kasper@kit.edu

R. Jäkel is with Institute of Anthropomatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany rainer.jaekel@kit.edu

R. Dillmann is with Institute of Anthropomatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany ruediger.dillmann@kit.edu

III. SCENES AND OBJECTS

Since the terms object and scene are used very differently depending on the field of research and application we feel it is necessary to define what in our work is considered to be a scene and what is an object.

A. Object and object class

An object class is the general concept that describes objects encountered in daily life that have common properties and functionality, that are limited in their spatial dimensions and are *movable* (location in space can change over time). Examples for object classes are *cup*, *chair* and *monitor*. An object is then an instantiation of an object class.

B. Scene and scene class

A scene class is the general concept of a limited region in 3D space that has specific properties and/or serves a specific purpose. A scene is then an instantiation of a scene class and consists of an arbitrary number of objects where the limitations might be supporting structures like a floor or walls. A scene class in this sense is equivalent to a type of room in a building or a part of that type of room. Examples for scene classes are *kitchen*, *bathroom*, *cubicle* or *desk area*.

IV. DATA ACQUISITION AND ANNOTATION

To be able to annotate scenes and analyze the spatial relations, we need to have a 3D representation of real world scenes where the objects can be marked and annotated to allow for an automatic calculation of the relations. Since we need to maximize the number of equivalent scenes, the data acquisition process needs to be fast, easy and as accurate as possible. In addition the digitization of real world scenes requires a sensor setup that can easily be transported to various locations. Several different approaches were evaluated and we finally decided to use an active 3D sensor, namely the Microsoft Kinect, to gather the 3D data. This sensor allows the acquisition of colored high resolution pointclouds in a handy and portable form factor. Since the complete scene cannot be captured with a single shot from the sensor, a method for the registration and fusion of several measurements is needed. To achieve this an ARToolkit marker is placed in the scene which acts as the reference coordinate system that all data is transformed into. The 2D image captured by the Kinect sensor is used to track the marker, using the ARToolkit library [1]. Every few seconds a snapshot of the current pointcloud is made and preregistered with the camera pose from the tracking. A fine registration via ICP[1] is then afterwards conducted to improve data quality. Finally the resulting pointcloud is filtered to reduce the amount of data and make the point distribution more homogeneous across the entire scene. All the operations on the pointcloud, like registration and filtering were done using the Point Cloud Library [8]. A sample of the resulting pointcloud can be seen in fig. 1. This setup allows the digitization of small scenes (around 3x3 meters) in a manner of minutes and is highly portable. The disadvantage of the system is that the marker always needs to be in the view of the sensor which restricts the area that can be digitized. Further work is needed to incorporate additional markers placed in the scene to allow the digitization of larger scenes, ideally the whole room.

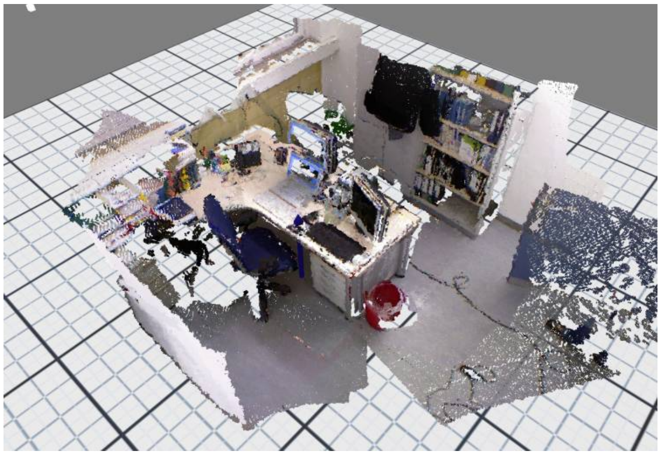

Fig. 1. Sample pointcloud generated with the sensor setup used for annotation.

After the scene has been digitized, the objects of interest in the scene have to be marked and annotated. This is done by placing oriented bounding boxes around the objects which are then labeled with the object class they belong into. This is done in a custom application created for this purpose. Obviously, the categorization of an object is entirely up to the annotator which means that results can be inhomogeneous across different scenes. This is reduced by a growing list of earlier used categories to choose from, but at the same time is not really an issue, since the class names are later on only used to make results easier interpretable for a human. A more serious problem is definition of a consistent reference frame for a given object. For the time being this is addressed by using a convention that the annotators should choose the orientation based on the functionality of the object. So for example a computer monitor's direction would be the normal of the screen surface pointing out of the monitor, whereas the "direction" of a cup would be pointing out of the opening of the cup. The result of the annotation is a number of oriented bounding boxes, enclosing all the objects in the scene the annotator could identify. The position, orientation and size of the boxes are then stored in a database for later use in the evaluation. Figure 2 shows a finished sample annotation. The direction of the annotation box is indicated by a yellow arrow pointing out of the box.

V. SPATIAL RELATIONS AND STATISTICAL DATA

We formulate the hypothesis that the structure of a scene and additional semantic information about an object is encoded in the spatial relations between the various objects contained in the scene. Several spatial relations could be candidates to support this hypothesis and contain information about the scene structure. Careful observation of everyday scenes reveals that spatial proximity is something found

[1] Iterative Closest Point

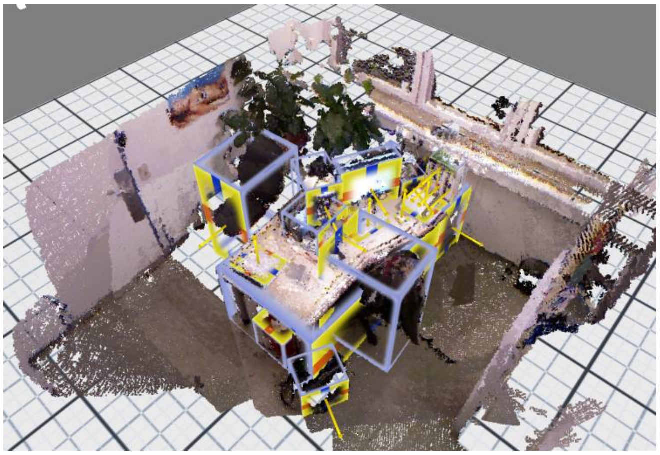

Fig. 2. Sample annotation consisting of a number of oriented bounding boxes around the objects in the scene.

regularly between distinct object classes. For example in an office setting, computer monitor and keyboard or desk and chairs or books oftentimes can be observed in close proximity to each other. The results for contextual search in [4] support this assumption as well. Aside from distance calculations, the position of objects relative to the facing of another object, which is specified by its orientation, is looked at. This is important information for relations like "in front of".

A. Relation calculations

Given two objects, O_1 and O_2, we calculate the Euclidean distance between two objects simply as the norm of the connection vector of the object centers:

$$D = \vec{O_1} - \vec{O_2}, \vec{O_i} \in \mathbb{R}^3 \tag{1}$$

The Euclidean distance in the x-y-plane is calculated similarly, but with the object centers' coordinates projected onto the plane. The bounding boxes of the annotated objects are oriented which gives each object a facing. This information can be used to calculate the position of all objects relative to a given one. To get the relative position, the position of all other objects is simply transformed into the local coordinate system of the queried object. Currently, we evaluate the orientation information only for the query object and only in x-y-plane because for most of the observed scene types this relation seems to contain the most information. Given a query O, the direction in which another object O_i can be found is calculated as an angle between the direction of O and the relative position of O_i where a value of 0 means that the object is directly in front of the query object and a value of 180 means directly behind.

$$\vec{P} = \vec{O_i} - \vec{O} \tag{2}$$

$$A = tan^{-1}(\frac{P_y}{P_x}), A \in [-\pi, \pi] \tag{3}$$

Other spatial relations (e.g. distance between boxes, enclosure, etc.) could also yield interesting information about the objects and will be investigated in future work.

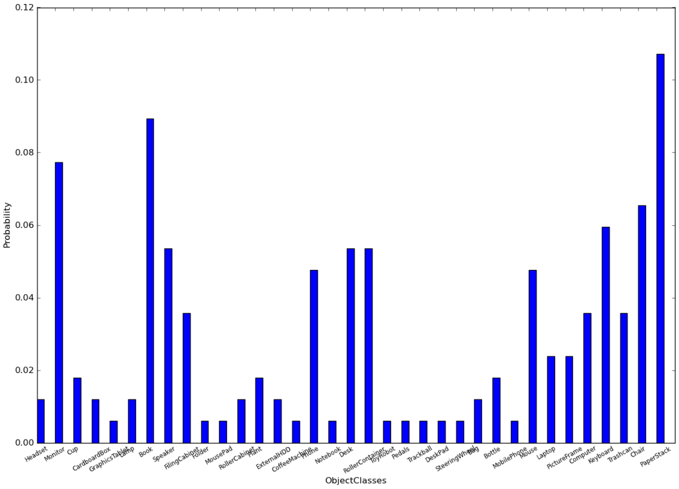

Fig. 3. Appearance probabilities of object classes. From left to right: Headset, Monitor, Cup, CardboardBox, GraphicsTablet, Lamp, Book, Speaker, FilingCabinet, Folder, MousePad, RollerCabinet, Plant, ExternalHDD, CoffeeMachine, Phone, Notebook, Desk, RollerContainer, ToyRobot, Pedals, Trackball, DeskPad, SteeringWheel, Bag, Bottle, MobilePhone, Mouse, Laptop, PictureFrame, Computer, Keyboard, Trashcan, Chair, PaperStack

B. Statistical measurements

When looking at a set of similar scenes that contain similar objects it is of course of interest what objects are found in the scenes, how likely the objects are to appear in the scene, and how many of the objects are to be observed in a scene in average. Aside from the probability of appearance, averages and variances of object sizes are calculated.

VI. DATA SET

For this first evaluation we chose an office space scene, focusing on the desk and the surrounding area. The setting was chosen because the variety in the kind of objects to be observed and the spatial distribution of these objects can be assumed to be rather uniform across observations. Because of the complex and time consuming data acquisition and annotation process, the amount of resulting data sets will inevitably be rather small. Choosing a setting with small variance alleviates this problem by reducing the amount of expected outliers. At the time of this writing, nine different office space settings have been digitized using the sensor setup described earlier and have been annotated using our custom software. Throughout the nine scenes a total of 168 objects in 35 object classes have been annotated.

VII. RESULTS

Using the calculi described above, the data set has been evaluated. The results of this evaluation are presented in the following paragraphs.

A. Probability of appearance

Figure 3 shows the probability distribution of all encountered object classes over all annotated scenes. Even though the amount of annotated objects cannot be considered to be representative, the distribution is reasonable and indicates

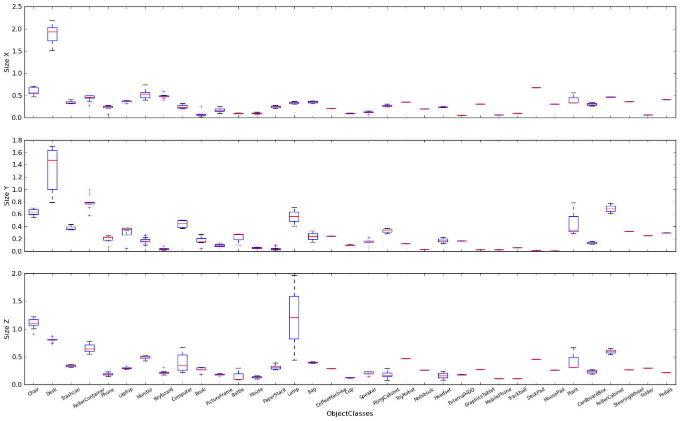

Fig. 4. Box plot of object sizes in X-, Y- and Z-direction. From left to right: Chair, Desk, Trashcan, RollerContainer, Phone, Laptop, Monitor, Keyboard, Computer, Book, PictureFrame, Bottle, Mouse, PaperStack, Lamp, Bag, CoffeeMachine, Cup, Speaker, FilingCabinet, ToyRobot, Notebook, Headset, ExternalHDD, GraphicsTablet, MobilePhone, Trackball, DeskPad, MousePad, Plant, CardboardBox, RollerCabinet, SteeringWheel, Folder, Pedals

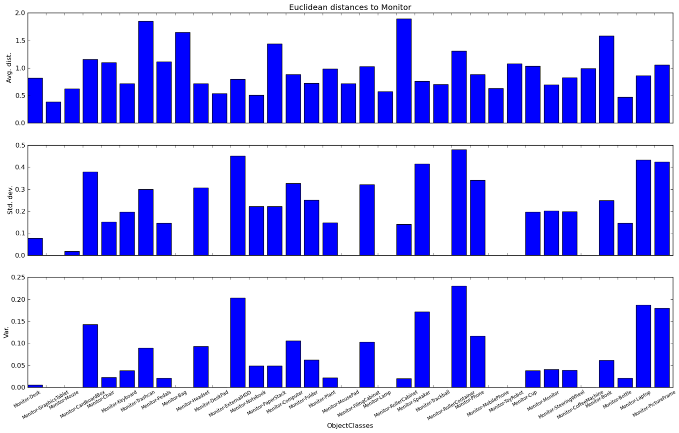

Fig. 5. Euclidean distances to class "Monitor". From left to right: Desk, GraphicsTablet, Mouse, CardboardBox, Chair, Keyboard, Trashcan, Pedals, Bag, Headset, DeskPad, ExternalHDD, Notebook, PaperStack, Computer, Folder, Plant, MousePad, FilingCabinet, Lamp, RollerCabinet, Speaker, Trackball, RollerContainer, Phone, MobilePhone, ToyRobot, Cup, Monitor, SteeringWheel, CoffeeMachine, Book, Bottle, Laptop, PictureFrame

that this information could be a candidate to characterize a certain scene type.

B. Object sizes

An interesting measurement for applications like object detection and localization is the average size of an object class as this could be used to reduce the search space or for plausibility checks. In fig. 4, the dimensions of all object classes in local X, Y and Z direction is visualized using box plots. Variance for most object classes is low in all dimensions, though some classes show interesting results. For example the class "Desk" shows a big variance in Y direction, which is somewhat unexpected. This can be explained by the fact that the annotated scenes contained two types of desks, one of which being L-shaped while the other type had a rectangular shape. The bounding box of the L-shaped desks is then naturally much bigger in one direction, which can be seen in the data. This indicates that this object class might be too general and would benefit from further specialization.

C. Euclidean distance

The next metric we were interested in is the Euclidean distance between object classes. The average distance, the standard deviation and the variance between the object class "Monitor" and all other classes can be seen in fig. 5. It can be seen that for many classes the variance is small, which could be an indication for a relationship that is observed regularly in this type of scene. Since some classes only appear in one scene the variance information for these classes is not representative and would need more data to have informational value. With the current amount of data, the Euclidean distance metrics value is rather limited.

D. Relative positions

Finally, the relative positions of object classes to each other are evaluated. Figure 6 shows the relative positions of the object classes "Book", "Chair", "Computer", "Cup", "Keyboard", "Monitor" and "Phone" to the object class "Monitor". The positions are projected into the x-y-plane and converted to polar coordinates where 0° is the direction of the query class. The plot shows that certain object classes are grouped together in specific spots, like the class "Keyboard" and "Chair" being almost exclusively found inside a 45° segment around the 0° direction and also in similar distances to the center. This is an empirical basis for the intuitive notion that a keyboard will almost always be placed in front of the monitor, and the chair for this workspace will also be located in front of the monitor, but further away than the keyboard. Looking at relative object positions thus allows the definition of regions where an object is most likely to be encountered relative to a reference object.

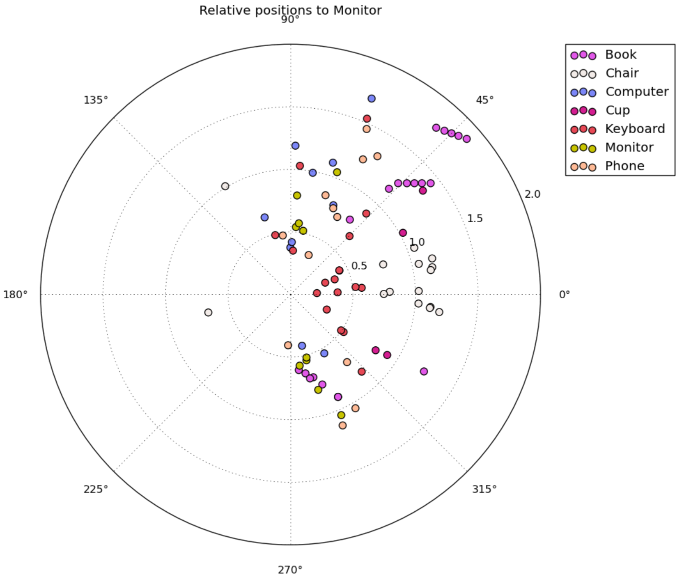

Fig. 6. Relative positions of selected object classes to class "Monitor".

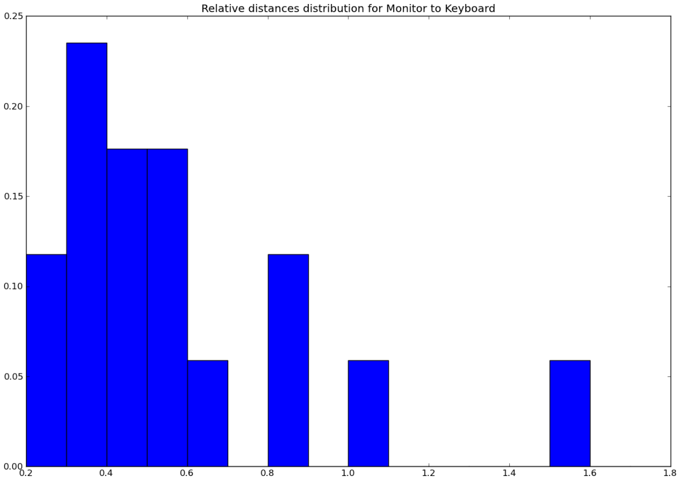

Fig. 7. Distribution for Euclidean distance values in x-y-plane for classes 'Monitor' and 'Keyboard'.

VIII. USING THE DATA

Having obtained the data about the relations, we want to demonstrate a short example on how this could be used in a typical scenario for service robotics — object recognition. Using the relation information as training data we can compute probability distributions about the values of the relations and the involved object classes. With these distributions one can answer the following question: Given an object whose class is known and the values for the relations to another, unknown, object, what is the most likely (known) class this object belongs to?

For a single relation R (like Euclidean distance), a known object of class A and an unknown object of class U, this can be formulated using the Bayes theorem:

$$P(U|A,R) = \frac{P(R|U,A) \cdot P(U|A)}{P(R|A)} \tag{4}$$

Using the training data, $P(R|U,A)$ can be calculated for all known object classes given A, the same is true for $P(U|A)$. To calculate the distribution of relation values for a given pair of object classes, the relation values have to be discretized because of the low amount of data. For the distance values, a discretization size of 10 cm has been chosen, based on the quality of the sensor data. Due to the amount of data and the resolution of the sensor, direction values have been discretized by a value of 10°.

Figure 7 shows $P(R|'Monitor','Keyboard')$ for the Euclidean distance in the x-y-plane. It can easily be seen that there is a peak around 0.4 m distance, which coincides with everyday experience. $P(U|A)$ can be calculated by counting how often a pair of object classes appeared in a given scene. This term represents how often this type of relation has been observed (we calculate relations for all pairs of objects in a scene) and thus is an important information about the relation. Depending on the correlation and weighting of the various relations, a most likely class can be derived. If there is more than one known object, the results from the above calculations for all the known objects can be combined to make a more robust prediction. Furthermore, if an approximate size for the unknown object is given, the information about average object sizes can be used to reduce the amount of possible candidates.

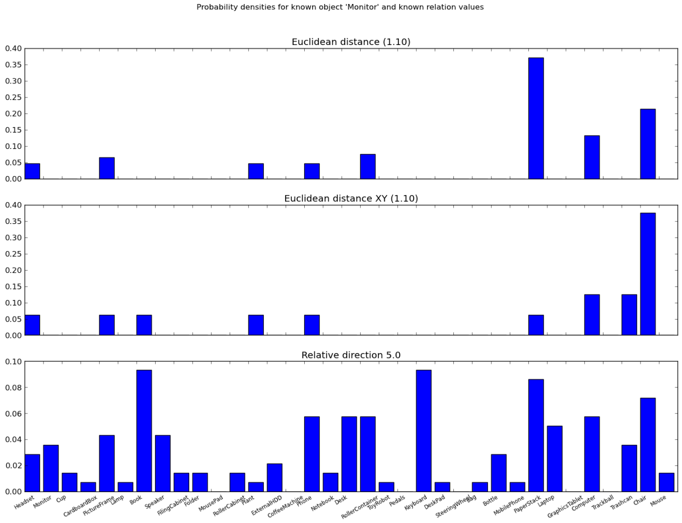

Fig. 8. Probability distributions for a known object class and given relation values. From left to right: Headset, Monitor, Cup, CardboardBox, PictureFrame, Lamp, Book, Speaker, FilingCabinet, Folder, MousePad, RollerCabinet, Plant, ExternalHDD, CoffeeMachine, Phone, Notebook, Desk, RollerContainer, ToyRobot, Pedals, Keyboard, DeskPad, SteeringWheel, Bag, Bottle, MobilePhone, PaperStack, Laptop, GraphicsTablet, Computer, Trackball, Trashcan, Chair, Mouse

Figure 8 and fig. 9 show sample results for such a query. The known object class in this case was set to 'Monitor' and the relations used are the Euclidean distance between object centers, the Euclidean distance between object centers in x-y-plane and the direction from 'Monitor' to the unknown object. Values for the Euclidean distances for fig. 8 are 1.1 m and 0.4 m in fig. 9, whereas the value for the direction was 5° in both cases. The results show that the distributions vary with changing relation values, but also that there's a difference between the Euclidean distance relations and an even greater difference to the direction relation. For example if we look at the object classes with the highest probabilities for the Euclidean distance, the class "Chair" is not the one with the highest value, whereas this is the case for the Euclidean distance in the x-y-plane (see fig. 8). The chart for the relative direction also rates the class "Chair" high, but there's others like "Book" or "Keyboard" that have even higher probabilities. This indicates that further investigation on the importance of a relation type and possible correlations between relations is needed, because different relations tend to "favour" different solution classes in many cases. At the same time the results strengthen the hypothesis that objects in an everyday scene are not placed randomly but have an inherent structure observable in the spatial relations between them.

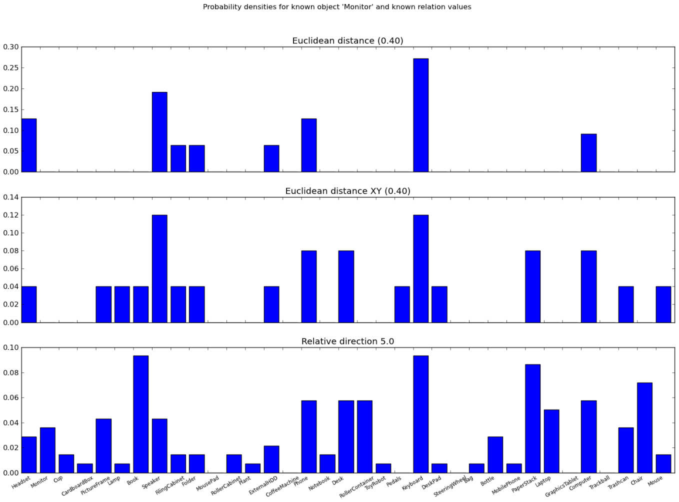

Fig. 9. Probability distributions for a known object class and given relation values. From left to right: Headset, Monitor, Cup, CardboardBox, PictureFrame, Lamp, Book, Speaker, FilingCabinet, Folder, MousePad, RollerCabinet, Plant, ExternalHDD, CoffeeMachine, Phone, Notebook, Desk, RollerContainer, ToyRobot, Pedals, Keyboard, DeskPad, SteeringWheel, Bag, Bottle, MobilePhone, PaperStack, Laptop, GraphicsTablet, Computer, Trackball, Trashcan, Chair, Mouse

IX. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

We have presented a system to gather spatial information about everyday scenes using a highly portable sensor setup that is capable of generating data, which is usable to build a manually annotated data set. Furthermore, spatial relations between objects have been introduced as a way to learn the structure of such a scene. We see this work as a first step towards an empirical base for scene understanding and the generation of machine readable background knowledge about objects in scenes. We believe that this knowledge will be of great benefit for service robots in multiple applications, e.g. object recognition, localization and manipulation. A theoretical example has been shown, which illustrates how the structural knowledge can be used to infer the class an object belongs to, solely by the relations it shares with the other objects in the scene. Results indicate that this approach can generate reasonable results that are coherent with the intuitive perception of a human, even with respect to the relatively small amount of data used.

B. Future Works

Future work will include the acquisition of more data, to make the training data more robust and statistical features more representative. Furthermore, other spatial relations shall be investigated as well as correlations between the appearance of objects in a scene and between spatial relations. Finally, other types of scenes need to be annotated as well and be compared to the existing set of data. Intuitively it should be possible to separate scene types by the objects typically appearing therein. It would also be interesting to apply the scene knowledge within a state of the art perception system. Finally, the annotated data sets and probability distributions will be made available, so other algorithms can be tested and developed.

X. ACKNOWLEDGMENTS

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) in the course of the Collaborative Research Center 588 (SFB 588 “Humanoid Robots – Learning and Cooperating Multimodal Robots”).

REFERENCES

[1] ARToolKit Home Page. http://www.hitl.washington.edu/artoolkit/. [Online]. Available: http://www.hitl.washington.edu/artoolkit/

[2] A. Aydemir, K. Sjöö, and P. Jensfelt, “Object search on a mobile robot using relational spatial information,” in *Proc. of the 11th Int Conference on Intelligent Autonomous Systems (IAS-11)*, Aug. 2010.

[3] S. K. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, “An empirical study of context in object detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[4] M. Fisher and P. Hanrahan, “Context-based search for 3d models,” in *ACM SIGGRAPH Asia 2010 papers*, ser. SIGGRAPH ASIA ’10. New York, NY, USA: ACM, 2010, pp. 182:1–182:10. [Online]. Available: http://doi.acm.org/10.1145/1866158.1866204

[5] J. M. Keller and X. Wang, “A fuzzy rule-based approach to scene description involving spatial relationships,” *Computer Vision and Image Understanding*, vol. 80, no. 1, pp. 21 – 41, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCX-45FBSGJ-H/2/6106f8a8da68eee6ce72220b3630d55d

[6] T. Malisiewicz and A. A. Efros, “Beyond categories: The visual memex model for reasoning about object relationships,” in *NIPS*, December 2009.

[7] B. Russell, A. Torralba, K. Murphy, and W. Freeman, “Labelme: A database and web-based tool for image annotation,” *International Journal of Computer Vision*, vol. 77, pp. 157–173, 2008, 10.1007/s11263-007-0090-8. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0090-8

[8] R. B. Rusu and S. Cousins, “3D is here: Point Cloud Library (PCL),” in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[9] R. B. Rusu, Z. C. Marton, N. Blodow, A. Holzbach, and M. Beetz, “Model-based and learned semantic object labeling in 3d point cloud maps of kitchen environments,” in *Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems*, ser. IROS’09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 3601–3608. [Online]. Available: http://portal.acm.org/citation.cfm?id=1733023.1733323