Scene Affordance: Inferring Actions from Household Scenes

Timo Lüddecke
timo.lueddecke@phys.uni-goettingen.de
Florentin Wörgötter
worgott@gwdg.de

Computational Neuroscience Group,
Third Institute of Physics,
Georg-August-University Göttingen

Abstract

In order to be helpful some day, robots do not only need to perceive their environment accurately but also have to be capable of carrying out appropriate tasks automatically. In a household setting, it is desirable to have a robot that does not require explicit instructions but rather identifies relevant tasks automatically.

We suggest to approach this challange by collecting a large dataset which describes actions between physical objects in household scenes. Various gravity-based, structured representations of the scene layout are proposed, which enable training models that reason about constituents of the scene and its associated actions. In order to foster a large number of annotations we created a web-based annotation tool which employs scenes from the SUNRGB-D dataset [8]. This way, the collection of annotations can be crowdsourced.

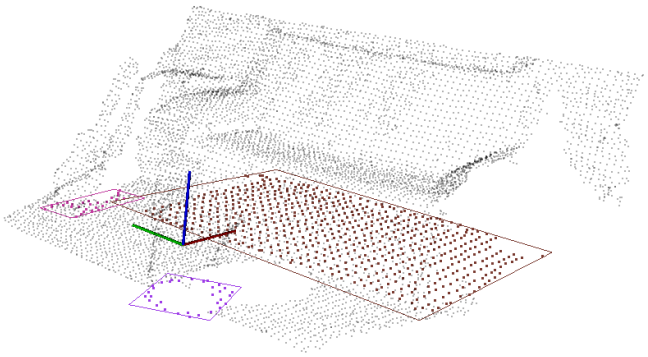

Figure 1: Visualization of the scene representation (here involving a sofa, guitar and a trash bin). Objects (indicated by colors) are identified based on their annotated segmentation and projected to the fitted ground plane. Object oriented bounding boxes are computed for the projected objects (rectangles).

1 Introduction

Driven by the success of deep learning based techniques, performance on many recongition tasks has recently improved by a large fraction. However, these task primarily deal with an accurate description of what is there. This work goes a step beyond description and asks what can be done, particularly with robotic applications in mind. We argue that even if you have a robot that is perfectly capable of correctly observing and understanding its environment it will not be of great value for humans until it automatically can figure out useful tasks on its own, depending on the situation. As a household robot owner, you want your robot to have some notion of common sense that enables it to understand required tasks like: clear the dishwasher or push the chair under the table, without explicit request. Analoguously to shape and texture of objects hinting possible actions to humans, which is known as affordances, some scene arrangements suggest actions. We dub the described kind of possible actions "scene affordances" to emphasize the difference to "object affordance", although the original affordance term by Gibson [1] encompasses both. Having a large-scale dataset of scenes with annotated actions involving objects and their states (empty, dirty, ...) could bring forward data-driven scene understanding and would allow machine learning methods to be applied on this particular problem.

Since 3d pointclouds by default are fairly large, low-level representations, a method to extract multiple smaller, gravity-based object matrices is suggested, where distances are differentiated as being orthogonal or parallel to the gravity axis. This way, details like appearance and shape are waived while the abstract spatial layout of the scene is maintained.

Using the described toolkit of scene representation and associated actions, rules like *if dirty dishes are close to each other stack them* or *if room is empty use remote control to turn off tv* could be learned from data.

2 Related Work

Currently we are not aware of any approach that explicitly deals with the problem of inferring actions from observed scenes.

The anticipation of human activities that is addressed in Koppula and Saxena [5] can be considered a closely related task. They model human pose, object affordances, object locations and sub-activities in a graph which changes over time through a temporal conditional random field. By sampling from this model, prospective activities can be predicted. These possible futures could involve actions we are interested in. But while their dataset only comprises 120 scenes, we prefer a high number of scenes rather than detail within scenes.

Zhu et al. [10] predict object affordances from images using a Markov Logic Network that expresses uncertainty with respect to the presence or abscence of certain attributes. Jiang et al. [4] model the relationship between human poses and scene objects using a Dirichlet process mixture model. In contrast to their work, we operate on the scene instead of object level, hence we call it scene affordance. Also, both approaches employ datasets comprising a smaller number of examples whereas we aim at a larger scale, although a direct comparison is not feasible because of the strong difference in settings.

The SUNRGB-D dataset already provides a good deal of the data that we need to solve the problem and unites multiple previous scene labelling approaches [3, 7, 9]. It comprises 10,000 scenes (involving rgb as well as depth) with associated pixel-wise object annotations and 3D object annotations, which however are not linked with each other. Often, there are more objects being annotated in 2d than in 3d (147K vs. 65K) with 3D objects often being chairs or tables. In this case, we favor the richness of the scene (many 2d objects) to the details of individual annotations (few 3d objects). Therefore we use 2D annotations and reconstruct objects by projecting the segmentation into the reconstructed pointcloud. Although SUNRGB-D provides annotations of the room layout, the ground plane or desk plane is automatically detected. This has the key advantage of a more general algorithm being applicable in more settings (and to prospective further datasets).

[2] also aim at extracting object arrangements from RGB-D scenes by matching objects with consideration of pose from a database to in-scene objects. Unlike them, we work on a more abstract level by employing available segmentation and concentrating on structured representations which are useful for detrmining actions.

3 Scene Representation

To facilitate reasoning we propose multiple representation schemes for the scenes that provide a more structured view onto the object arrangement of the scene than the raw RGBD data. These representation might serve as an input for algorithms addressing scene affordance. A key observation is that gravity forces the room layout to assume some structure: Objects must stand on something. We explicity consider gravity in our distance based representation by seperating distances along the gravity axis and distances orthogonal to it.

The goal of the representations is to be as abstract as possible and allow algorithms to find possibly relevent semantic relations like *next to* or *above* without explicitly defining threshold distances for them. We capture both, relative representations which express pairwise distances

between objects and absolute locations of objects. This allows the dataset to be applied in multiple different learing and reasoning settings, e.g. the pairwise distances spawn a fully connected graph which could be employed in an energy-based model whose potentials could be learned. The separation between distances in height and in plane is due to our assumption that gravity is essential for object arrangements. Specificly, the representations are listed below:

Relative (pairwise)	**Absolute (for each object)**
• In-plane euclidean distance	• Coordinates of corners of each object-oriented bounding box
• Objects' top distance (along gravity axis)	• Height over ground (or desk)
• Objects' bottom distance (along gravity axis)	• Object's height

The emphasis on heights is intentional because an alignment with respect to two object's bottoms might serve as a clue to infer their functions, e.g. a plate and a mug are usually placed on the table hence they share their bottom's height.

Note, the absolute scene representation grows linear with the number of objects while pairwise distances grow quadratically. Since the number of objects per scene is limited, even pairwise matrices remain at a manageable size.

Technically, obtaining the scene representation is realized by fitting a ground-plane into the scenes. A large planar set of points is searched within a certain range around a common camera tilt. Since this does not neccessarily guarantee the presence of a ground plane, we assume that the variance of inlier normals to be lower than a certain threshold. Sometimes neither the ground nor a reliable table is captured by the camera, in these cases the extraction fails. By projecting objects to the groundplane and the gravity axis (normal of the ground plane) we obtain our representation space where distances and locations are calculated.

Sometimes the same word refers to different objects. We address this issue by matching WordNet [6] concepts to the labels provided by SUNRGB-D through an semi-automatic procedure. Expressing labels in terms of WordNet synsets also allows for incorporaing further data, e.g. the synset definition text, hypernyms or other datasets relying on WordNet synsets.

4 Web-based Tool

To allow the collection of annotations on a large scale, we employ a web-based annotation tool. Currently we are not sure whether we will actually crowdsource the data or ask member of our lab. While the former might lead to a larger dataset the latter would possibly produce higher quality data since we can instruct the annotators more thoroughly. In either case, the web tool will be of great help. It supports the annotator by various means: Objects can be selected directly from the segmentation of the scene, or typed into text field with suggestions being made. The definition of WordNet synset is shown to make shure the right concept is chosen.

5 Action Annotation

The actions we collect involve either one or two objects, which should cover a large fraction of actions as even complex action plans can often be disintegrated into small actions. Additonally, each object is assigned one or multiple of the states *broken, full, dirty open, on* or *default* if none of the former states applies.

We believe it would cause too much confusion to ask for actions a robot should carry out thus we ask for actions that seem to be useful for a human. Also, in the long term one can expect that the overlap of activities conductable by robots and humans increases. Hence, a human is implicitly being assumed to be present in the scene, e.g. "sit on chair" actually means "human sits on chair" and "water pour into cup" means "human pours water into cup".

Having obtained the annotations, different types of reasoning are possible. Most obvious - and possibly most applicable - is determining a couple of useful actions for an unknown (but already segmented) scene. Depending on the underlying model it might also be possible to infer scene representations for actions or reason on the scene representation itself.

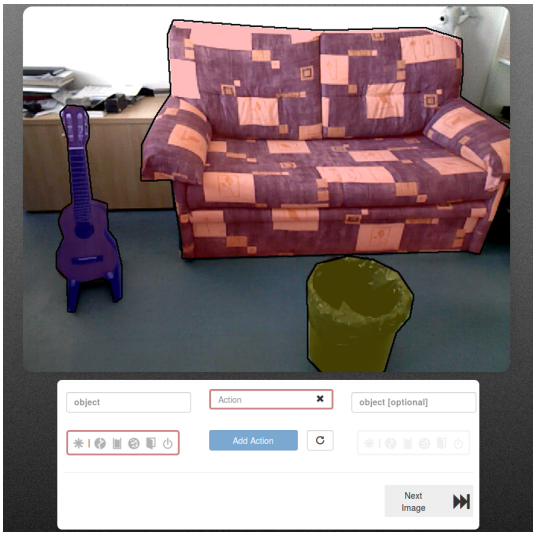

Figure 2: Illustration of the web based annotation tool showing the same scene as Figure 1. Red borders indicate fields that need to be filled by the user.

The latter could refer to questions like, what is the most likely additonal object in a given scene or how likely is a provided scene layout.

6 Conclusion

This work contemplates affordances of multiple objects, on a scene level, which is, to the best of our knowledge a new approach. We propose to collect a dataset of action annotations involving up to two objects in household scenes. As a vehicle for reasoning, a diverse set of object-based representations of these scenes is suggested, allowing flexibility for various model types operating on these representations. We believe these contributions are a good starter-kit for machine learning based methods that address the task of scene affordance.

References

[1] James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[2] Saurabh Gupta, Pablo Arbelaez, Ross Girshick, and Jitendra Malik. Aligning 3d models to rgb-d images of cluttered scenes. In *CVPR*, June 2015.

[3] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T. Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *IEEE Workshop on Consumer Depth Cameras for Computer Vision*, 2011.

[4] Yun Jiang, Marcus Lim, and Ashutosh Saxena. Learning object arrangements in 3d scenes using human context. *ICML*, 2012.

[5] Hema Swetha Koppula and Ashutosh Saxena. Anticipating Human Activities using Object Affordances for Reactive Robotic Response. In *Robotics: Science and Systems*, 2013.

[6] George A Miller. WordNet: a lexical database for English. 38(11): 39–41, 1995.

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. 2012.

[8] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[9] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *CVPR*, pages 1625–1632, 2013.

[10] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about Object Affordances in a Knowledge Base Representation. In *ECCV*, pages 408–424. Springer, 2014.