Affordance Origami: Unfolding Agent Models for Hierarchical Affordance Prediction

Viktor Seib(✉), Malte Knauf, and Dietrich Paulus

Active Vision Group (AGAS), University of Koblenz-Landau,
Universitätsstr. 1, 56070 Koblenz, Germany
{vseib,mknauf,paulus}@uni-koblenz.de
http://agas.uni-koblenz.de

Abstract. Object affordances have moved into the focus of researchers in computer vision and have been shown to augment the performance of object recognition approaches. In this work we address the problem of visual affordance detection in home environments with an explicitly defined agent model. In our case, the agent is modeled as an anthropomorphic body. We model affordances hierarchically to allow for discrimination on a fine-grained scale. The anthropomorphic agent model is unfolded into the environment and iteratively transformed according to the defined affordance hierarchy. A scoring function is computed to evaluate the quality of the predicted affordance. This approach enables us to distinguish object functionality on a finer-grained scale, thus more closely resembling the different purposes of similar objects. For instance, traditional methods suggest that a stool, chair and armchair all afford sitting. However, we additionally distinguish sitting without backrest, with backrest and with armrests. This fine-grained affordance definition closely resembles individual types of sitting and better reflects the purposes of different chairs. We report evaluation results of our approach on publicly available data as well as on real sensor data.

Keywords: Affordance · Affordance prediction · Visual affordances · Affordance hierarchies · Object recognition

1 Introduction

Since Gibson's work on affordances [1] a lot of effort was put into the theoretical investigation of affordances [2,3] and their applications in other fields. When it comes to classification in computer vision, many approaches struggle with large intraclass appearance variations. The reason is at hand: classes are defined by the functionality of objects, rather than their visual appearance. By describing action possibilities between an agent and an object, affordances allow to detect similarities on a functional level, rather than solely rely on the object's appearance. Thus, approaches exploiting affordances were shown to augment the classification process [4,5].

J. Braz et al. (Eds.): VISIGRAPP 2016, CCIS 693, pp. 555–574, 2017.
DOI: 10.1007/978-3-319-64870-5_27

This is why reasoning about an object's purpose has become an important area in today's research in robotics. While shape features are often acquired locally (i.e. around salient points) and might therefore be misleading, detecting a functionality of an object facilitates categorization. Additionally, predicting affordances of objects instead of the object classes allows objects and tools to be applied even without the precise knowledge of the class the object belongs to. Even objects of different classes can be applied according to a certain affordance required by the agent. For example, if an agent (e.g. a robot) needs to hammer, it would pick a heavy object providing enough space for grasping and a hard surface to hit on another object. This works without knowing the category *hammer* or having a hammer available by e.g. using a stone instead.

While in robotics humans play an active role in teaching affordances to robots, e.g. by interaction [6,7] or by imitation [8–10], the vision community follows other approaches. Some approaches completely omit the interacting agent and propose to derive object descriptors by physical simulation [4], by data from additional sensors (e.g. kinematic data [11]) or purely from visual sensors [12]. Other approaches create or "imagine" human models in the environment [13]. These human models are exploited to propose comfortable poses for sitting [5], to learn human-relative placement of objects [13] or to explore action possibilities in human workspaces [14]. In contrast to approaches in robotics we do not record kinematic data of an agent, neither do we detect affordances by interaction. Nowadays, it can be expected that visual perception is mostly common in robots and it is thus plausible to rely on that data. Thus, the approach proposed in this paper relies on visual data only.

In our approach we employ the *observer's* view on affordances as introduced by Şahin et al. [15]. While the environment is being observed by a robot equipped with certain sensors, the system is looking for affordances that afford actions to a predefined model. In our case this predefined model is an anthropomorphic agent representing a humanoid. In recent work [5,13] this *observer's* view is often referred to as *hallucinating interactions*.

In the proposed method we focus on the complementary nature of a humanoid agent and its environment. In our previous work we proposed detecting fine-grained or hierarchical sitting affordances with a simulated anthropomorphic agent [16]. Given an agent and an affordance model, the agent's joints are transformed from a start to a goal pose. Figuratively speaking, the agent model is unfolded from an initial pose to a functional pose that corresponds to predicted affordances (Fig. 1). The final state of the individual joints determines the predicted fine-grained affordances in the hierarchical affordance model.

We use indoor or home environments that are considered as environments specifically designed to suit the needs of humans. Therefore, the complementary agent to the investigated environment is an anthropomorphic, i.e. human, body. Thus, for the purpose of this work affordances shall be informally defined as *action possibilities that the environment offers to an anthropomorphic agent*.

In this work we extend our previous work by refining our affordance model and including more action possibilities: sitting and lying. The refined model

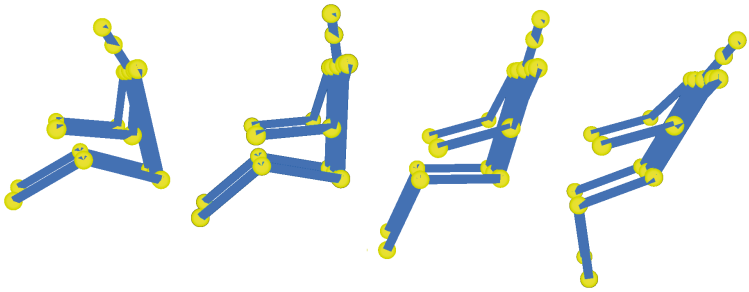

Fig. 1. Unfolding the agent model for the sitting affordance hierarchy.

allows to handle complex scenes in contrast to individual objects of our previous approach. We further provide a formal model for hierarchical affordance prediction and show its applicability on large datasets of furniture objects.

Related approaches in the literature [4,17,18] distinguish affordances on a coarse scale. The considered affordances often include sitting (chairs), support for objects (tables) and liquid containment (cups). We propose looking closely at the individual affordances and distinguishing their functional differences an a fine-grained scale. We already introduced the concept of *fine-grained* affordances in [19] to closely resemble the functional differences of related objects. Although good results could be obtained, our previous work was a proof-of-concept with several limitations.

In the presented work, we concentrate on fine-grained affordances derived from the affordance *sitting* and *lying*. We present a new algorithm for fine-grained affordance prediction that differentiates between 4 typical functionality characteristics of the *sitting* affordance. We divide the coarse affordance *sitting* into the fine-grained affordances *sitting without backrest*, *sitting with backrest*, *sitting with armrest* and *sitting with headrest*, whenever the sitting functionality is supported by additional environmental properties that can be exploited by the considered agent. Further, we give an outlook on different subaffordances of the course affordance *lying*.

A system that is able to find affordances either encounters only those objects that were specifically designed to support the affordance in question or environmental constellations that afford the desired action. Our algorithm takes point clouds e.g. from a RGB-D camera as input. The input data is directly searched for affordances (and thus functionalities) without prior object segmentation. In the core of the algorithm, the agent model is unfolded and checked for collisions with the environment. Specific goal configurations of the agent model represent different types of fine-grained affordances. The encountered affordances are segmented from the input point cloud. This segmentation can serve as an initial segmentation for a subsequent object classification step (not further explored in this work). Since the found affordances (especially on a fine-grained scale) provide many hints on the possible object class, categorization can be performed with fewer training objects or simpler object models. The presented fine-grained

Fig. 2. Example furniture objects corresponding to the different fine-grained affordances predicted by the presented approach. From left to right: stool (*sitting without backrest*), two chairs (*sitting with backrest*) and three armchairs (*sitting with backrest* and *sitting with armrest*). Additionally, the rightmost chair also supports the *sitting with headrest* affordance.

affordances correspond to objects such as a stool, chair, armchair and a chair with head support (Fig. 2). Specifically, an affordance-based categorization system can be exploited as outlined in the following. Affordances enable the detection of *sittable* objects even without knowing object classes as *stool*, *chair* or *couch*. Following the idea of fine-grained affordances, a stool standing close to a wall can even provide both affordances: sitting with and without backrest (in the former case the back is supported by the wall). This intuitively corresponds to the way a human would utilize an object to obtain different functionalities.

The remainder of this work is structured as follows. Related work on affordances in robotics is presented in Sect. 2. Section 3 introduces the model definitions applied in our algorithm and Sect. 4 explains our approach for fine-grained affordance prediction in detail. The proposed algorithm is evaluated in Sect. 5. Finally, a discussion is given in Sect. 6 and Sect. 7 concludes the paper and gives an outlook to our future work.

2 Related Work

Hierarchies in affordances have been explored mainly in design theory to reason about functional parts of objects [20,21]. Their goal is to divide objects into different functional parts that represent different affordances. This allows a designer to identify desired and undesired affordances in early stages of product design. Note however that this hierarchical affordance modeling is conceptually different from the fine-grained affordances applied in this paper. We do not separate objects in different parts with different affordances. Rather, our object independent approach separates an affordance (in this case the *sitting* and *lying* affordances) into different subaffordances on a fine-grained scale.

Other approaches like the work of Hinkle and Olson [4] use physical simulation to predict object functionality. The simulation consists of spheres falling onto an object from above. A feature vector is extracted from each object depending on where and how the spheres come to rest. The objects are classified as cup-like, table-like or sitable.

Research especially focusing on sitting affordances has been conducted over the past years. Office furniture recognition (chairs and tables) is presented by Wünstel and Moratz [12]. Affordances are used to derive the spatial arrangement of the object's components. Objects are modeled as graphs, where nodes represent the object's parts and edges the spatial distances of those parts. The 3D data is cut into three horizontal slices and within each slice 2D segmentation is performed. The segmentation results are classified as object parts and matched to the object models. Wünstel and Moratz' approach detects sitting possibilities also on objects that do not belong to the class *chair*, but intuitively would serve a human for sitting. Unlike the approach of Wünstel and Moratz, we encode the spatial information needed for affordance prediction in an anthropomorphic agent model and affordance models, rather than creating explicit object models.

Hierarchical classification of object parts has been explored in [22]. Complex object models were proposed to identify object parts and thus infer the subcategory of an object type. Affordances were not mentioned explicitly, though. In [23] a human agent model is used to classify objects. Again, affordances were not explicitly mentioned here. Contrary to our approach they needed a segmented object as input for classification.

Our algorithm takes 3D data and detects affordances inside these data. In our approach, individual objects exposing these affordances are subsequently segmented based on the detection result. We propose a hierarchical affordance model and the detection of fine-grained affordances by unfolding an anthropomorphic agent model and fitting the agent to functional object parts. By applying our agent and affordance models we do not need to create complex object models as opposed to [22] and do not have any constraints on the environment (e.g. segmented objects) as in [23]. In our case, the segmented part of the scene is a result of the detected affordances on the input data.

More recently, Grabner et al. [5] proposed a method that learns sitting poses of a human agent to detect sitting affordances in scenes to classify objects. For training, key poses of a sitting person need to be placed manually on each example training object. In detecting chairs, their approach achieves superior results over methods that use shape features only. However, as pointed out by Grabner et al. their approach has difficulties in detecting stools, since they were not present in the training data. Consequently, the approach of Grabner et al. does not allow to find affordances per se, but rather affordances of trained object class examples.

In the present paper we follow a different approach. Our goal is to directly predict sitting affordances in input data, independently of any possibly present object classes. Further, if a sitting affordance is hypothesized, it will be categorized on a fine-grained scale according to the characteristics of the input data at the position where the affordance is assumed. Our approach does not rely on examples of sitting furniture, but only on the agent and affordance models. Our fuzzy function formulation encodes expert knowledge to connect the input data with the desired functionality with respect to the given agent model. Still, our models remain simple and also work even if important parameters are changed.

We show this generality by varying the size of the applied agent model during our evaluation (Sect. 5). Additionally and similar to Grabner et al., our approach suggests a pose how the detected object can be used by the agent.

Note that our approach is ignorant of any object categories. However, our fine-grained affordance formulation allows for a more precise object categorization as a consequence of affordance prediction. Due to the fine-grained scale on which affordances are predicted, object categories can be easily linked to the prediction result (e.g. if a backrest could be detected or not). Our approach thus suggest as which kind of object the detected object exhibiting the affordance can be used. However, the detailed analysis of detected objects and their classification is left for future work.

3 Affordance Modeling

Usually, affordances are defined as relations between an agent and its environment [1,15,24]. Since these two entities are crucial for affordances, we start with their definitions. Then, a definition of fine-grained affordances is provided.

3.1 Environment and Agent

Contrary to our previous work [19], we do not need an explicit environment model. Our algorithm is designed to work on point clouds (e.g. from an RGB-D camera). Thus, a point cloud $P = \{\boldsymbol{p}_i\}, \boldsymbol{p}_i \in \mathbb{R}^3$ defines the environment in this setting. There are no further models or assumptions involved in our environment definition except the two necessary constraints when working with affordances. Firstly, the environment must correspond to the body-scale metrics of the agent. Secondly, the agent and environment must share a common coordinate frame (i.e. common ground plane and up-vector).

The affordances applied in this approach correspond to functional properties offered to humanoid agents. The agent H is modeled as a directed rooted tree $H = (V_H, E_H)$ with vertices V_H and edges $E_H \subseteq V_H \times V_H$ representing a scene graph. In this graph, nodes represent joints in a human body and edges represent parameterized spatial relations between these joints. The spatial relations correspond to average human body proportions. The nodes contain information on how the joints can be revolved while maintaining an anatomically plausible state (i.e. without harming a real human if the same state would be applied).

In the refined approach that we present here, the agent H is modeled according to average human body size and proportions as reported in a statistical investigation [25] and in [26]. Each edge $e \in E_H$ in the graph H represents parameterized body parts of the agent and is attributed with a length $l = \|e\|$ to reflect the dimensions of the human body. When fitting the agent into the environment during affordance prediction, the edges of the graph are approximated by cylinders for collision detection. Further, each vertex $v \in V_H$ represents movable joints in the broader sense and is attributed with an angle θ. This angle θ defines the current state of the joint and describes the rotation relative to the

parent joint in the graph around the lateral axis. Note that this simple model does not reflect all possible degrees of freedom of a human body. However, this simplified human model is sufficient for our purposes.

3.2 Affordances and Affordance Hierarchy

An affordance is an action opportunity offered to an agent by its environment. We suggest to consider affordances on a fine-grained scale. This means that an affordance is a generalization of similar action opportunities and thus can be divided into fine-grained affordances or subaffordances. For instance, the affordance *sitting* is a generalization of more precise relations that an agent and its environment can take. We demonstrate our ideas by distinguishing between the fine-grained affordances *sitting without backrest*, *sitting with backrest*, *sitting with armrest* and *sitting with headrest*. Further, we give an outlook on the *lying* affordance which can be seen as a generalization of lying with *elongated* or *raised body* and lying with *stretched* or *raised legs*. As is obvious from the specializations of the *lying* affordance, subaffordances can be mutually exclusive.

In the context of this paper we define an affordance A as a set $A = \{F_0 \dots F_j\}$ of fine-grained affordances F_i. For a given environment P, affordance A and initial agent configuration H_A the function Aff : $H_A \times P \times A \rightarrow \{(F, \boldsymbol{p}, H_g)_i\}$ determines a set of tuples. Each tuple contains $F \subseteq A$, a set of fine-grained affordances present at position $\boldsymbol{p}$ in the environment with a goal agent configuration H_g. The algorithm described in the next Section is an implementation of the above function Aff.

For each affordance A, an initial pose H_A of the agent needs to be defined. Thus, every fine-grained affordance F_i specializing the same affordance A has the same initial pose. In this work we use a separate initial pose for sitting and one for lying. The initial pose refers to the joint states of the simulated agent prior to any transformations and collision tests.

The fine-grained affordances F_i of each affordance A are organized in an affordance hierarchy (Fig. 3). In this work we examine two affordance hierarchies: $\mathcal{A}_{sitting}$ for sitting affordances and $\mathcal{A}_{lying}$ for lying affordances.

An affordance hierarchy is defined as a directed rooted tree $\mathcal{A}$. Each node $F \in \mathcal{A}$ corresponds to a fine-grained affordance and is a tuple $F = (J_F, E_F)$. Here, $J_F = \{(\theta_1, \theta_2, \theta_d, \mathbf{x}, v)_i\}$ defines constraints on valid angles $\theta_1 \leq \theta \leq \theta_2$ for a vertex $v \in V_H$ around axis $\mathbf{x} \in \mathbb{R}^3$. The angle θ_d defines a default pose that is used for stability checks (see Sect. 4). These angle constraints are chosen to reflect a broad range of valid poses for the corresponding affordance. Further, $E_F \subseteq E_H$ define affordance specific edges in the agent model H. The agent model edges E_F are checked for collision while the agent is unfolded and the corresponding vertices are transformed from θ_1 to θ_2.

After processing a node $F \in \mathcal{A}$ either a collision occurred or not (i.e. the fine-grained affordance F_i was found or not). The edges in the affordance hierarchy graph $\mathcal{A}$ are annotated with a constraint $c \in \{true, false\}$. The constraint c indicates whether child nodes of F_i are processed in case the affordance was found ($true$) or not ($false$). An affordance A is found if each subtree of the root

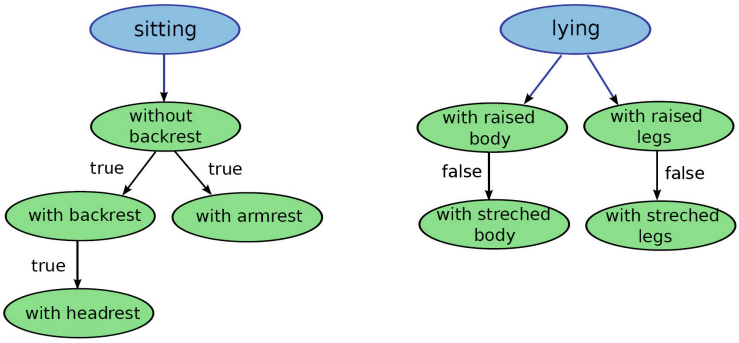

Fig. 3. The presented affordance hierarchies specialize the *sitting* and the *lying* affordance into fine-grained affordances. The arrows indicate the dependencies between the fine-grained affordances.

node in the affordance hierarchy $\mathcal{A}$ has at least one valid subaffordance F (i.e. at least one subaffordance where the agent H has contact with the environment).

Note that some of the fine-grained affordances inside an affordance hierarchy $\mathcal{A}$ depend on others. For example, if the environment affords *sitting with backrest* to the agent it must necessarily afford *sitting without backrest* as well, because the agent can choose not to use the backrest while seated. The dependencies as defined in our models are depicted in Fig. 3.

4 Predicting Fine-Grained Affordances

The algorithm for fine-grained affordance detection is essentially based on dropping an agent model in its default pose into the scene at appropriate positions. These positions need to be found beforehand. The joints of the model are then transformed to achieve maximum contact with the point cloud. Only joints relevant for a certain affordance are considered. The initial pose of the agent, as well as the joint transformations are determined by the affordance models. Further, only the agent model and the affordance models determine the current functionality of the detected object. This means that the presented approach also finds objects that might not have been designed to fulfill a certain functionality. However, based on visual information and their position in the scene they afford the desired actions. We confined the evaluation to an agent representing an average human adult and to fine-grained affordances derived from the affordance *sitting* and *lying*.

4.1 Extracting Positions of Interest

Before unfolding the agent into the scene, the search space needs to be reduced to the most promising positions. We therefore create a height map of the scene (Fig. 4(b)). The point cloud is subdivided into cells. In our experiments a size

of 0.05 m provided a good balance between precision and calculation time. The highest point per cell determines the cell height. We decided in favor of the highest point instead of the average to avoid implausible values at borders of objects, where a cell may contain parts of the object and e.g. the floor.

Subsequently, a circular template, approximating the agent's torso, is moved over the height map to test whether a cell is well suited to provide support for the agent. The diameter of this template corresponds to the width of the agent as defined in the model. The decision for each cell is based on fuzzy sets as introduced by Zadeh [27]. We define 3 membership functions: discontinuity, roughness and height (Fig. 5). Discontinuity is a measure defined in percent of invalid cells or holes within the current position of the circular template. Roughness is the standard deviation of the height of all cells within the circular template. Finally, the membership function height is used to include only cells in a certain height that allow comfortable sitting with bent knees, while the feet still touch the ground. Note that we use the same height function for lying, since in home environments positions where a person would lie down also afford sitting. However, this function can be disabled in the algorithm configuration to allow for valid positions on the ground or on higher planes like tables.

One single rule is enough to decide whether a position is a valid hypothesis for further processing. We use the intersection of these membership functions to obtain the following rule: *IF roughness is low AND discontinuity is low AND height is comfortable THEN the position is a valid hypothesis.* Of course, with more affordances, more rules will be needed. The fuzzy value obtained from these functions is defuzzyfied on the function depicted in Fig. 5(d) using the *first of maximum* rule. The test is performed for both fuzzy sets of this rule, obtaining a crisp value for *available* and *not available* and deciding in favor of the fuzzy set with the higher crisp value. For each position, a possible agent orientation is obtained by considering the height gradient descent in the height map. The orientation for *sitting* affordances is parallel to the gradient vector, while the orientation for *lying* affordances is orthogonal to this vector. The positions obtained in this manner are used as possible positions in further algorithm steps (Fig. 4(c)).

4.2 Initial Agent Fitting

In the next step, each hypothesis position is checked to provide enough space for the agent model. We test several agent model orientations in this step since the initial circular template was an approximation of the agent's torso. However, in this step also the corresponding rotation needs to be found to provide enough room for e.g. the agent's legs in case of the *sitting* affordance. To reduce the amount of tested orientations we use the hypothesis orientation obtained in the previous step (Sect. 4.1) and test a few rotations in a certain range around that orientation.

For this tests, the agent is put into a default pose (defined by the angles θ_d) and is positioned above the hypothesis position. We use the FCL library [28] for collision detection between the scene P and the agent H. FCL detects collisions between 2 objects and returns the exact position at which the collision occurred.

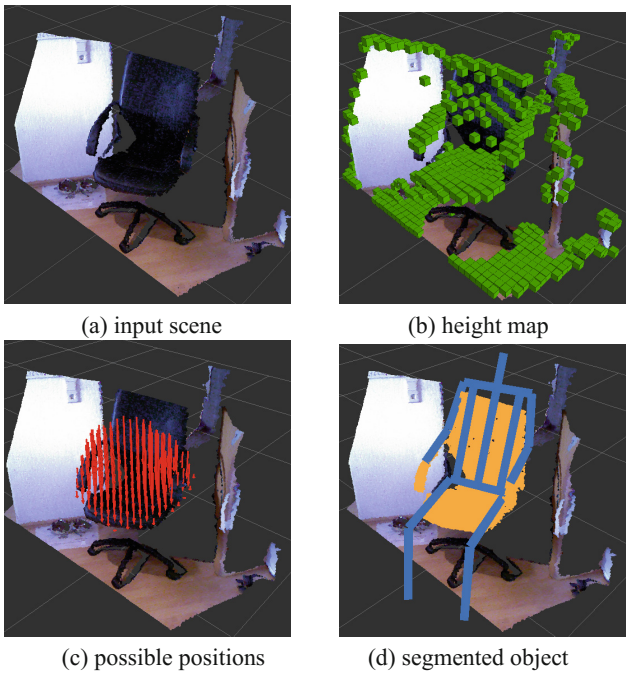

Fig. 4. Illustration of different algorithm steps. The input scene is shown in (a) and the corresponding height map in (b). Image (c) shows the possible positions for sitting affordances found by our fuzzy set formulation. The length of the red arrows corresponds to the defuzzyfied value from the availability function. The final agent pose as well as the object segmentation is shown in (d) for the fine-grained affordances *sitting with backrest* and *sitting with armrest*.

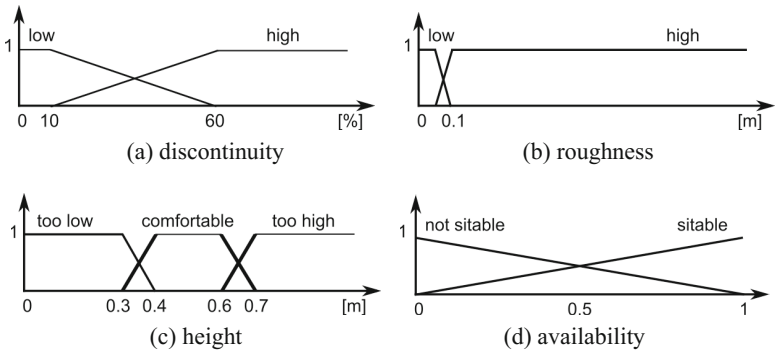

Fig. 5. Membership functions used to find valid positions for sitting and lying affordances. The functions in (a), (b) and (c) are used to evaluate the rule, while the function in (d) is used for defuzzyfication to determine the possible presence of the affordance.

As input for FCL we convert the point cloud of the scene to the OctoMap representation [29] and approximate the individual body parts of the agent by cylinders.

If a collision occurs before any lowering of the model, the current orientation at that position is discarded. If there is no collision (i.e. the scene provides enough free space at that position), the agent is lowered until a contact with the scene occurs. The edges E_F belonging to the most general affordance $F \in \mathcal{A}$ are subsampled and the distance d of each sample to the scene is determined and a stability score obtained by using an unnormalized Gaussian

$$s = \exp\left(-\frac{(\frac{1}{n}\sum_i d_i)^2}{2\sigma_d^2}\right), \tag{1}$$

where σ_d is a threshold and n the number of sampled distances d. This score ensures a stable positioning of the agent and avoids that only a small part of the agent collides with the scene. The orientation with the best stability score s at a hypothesis position is kept for further steps. All stable positions qualify for the next algorithm step.

4.3 Unfolding the Agent Model

The affordance hierarchy $\mathcal{A}$ is iterated from the root node. Each node F corresponds to a fine-grained affordances and is checked at the given position. The agent is transformed to its initial pose, defined by the θ_1 parameters in each agent node $v \in V_H$. Subsequently, all vertices in the set J_F of F are iteratively transformed from θ_1 to θ_2, while the edges $e \in E_F$ are checked for contacts with the scene P.

For instance, the fine-grained affordance *sitting with backrest* is detected during the transformation of the agent's torso, comparable to the agent's movement of leaning backwards against a backrest. The affordance is detected if a contact with the scene is encountered during the transformation. If a joint reaches its maximum goal pose without a collision the algorithms assumes that the current fine-grained affordance is not present.

The subaffordance F is detected if a collision occurs at an angle $\theta_c \in [\theta_1, \theta_2]$ during this transformation. The resulting angle θ of the vertex v is then determined as

$$\theta = \begin{cases} \theta_c, & \text{if a collision occurs,} \\ \theta_d, & \text{otherwise.} \end{cases} \tag{2}$$

In the next step, all child nodes of F in $\mathcal{A}$ are processed if the associated constraint c of the outgoing edge of F matches the collision state in F.

Note that in contrast to normal affordances, a fine-grained affordances might depend on the existence of another fine-grained affordance (Fig. 3). Taking the *sitting* affordance as an example, the *sitting without backrest* subaffordance is checked first, as other affordances depend on it. *Sitting with backrest* and *sitting*

with armrests are checked subsequently. The *sitting with headrest* affordance is checked as the last one, since it depends on the presence of a backrest. The output of this step is the final pose of the agent (position $\boldsymbol{p}$ and joint states H_g), as well as a set of predicted fine-grained affordances F. The resulting joint states represent the suggested body pose of the agent H, specifying how a hypothetical object could be used exhibiting the predicted affordance at the given position.

4.4 Combining Evidence for Affordance Presence

In our previous approach [16] the score for a detected affordance was determined by the number of detected contacts over all processed edges $e \in \bigcup_i E_{Hi}, \forall F_i \in \mathcal{A}$. However, this score did not produce meaningful results in some situations.

Our goal here is to compute the score for each affordance $F \in \mathcal{A}$ and to combine each individual score in a meaningful way. To achieve this, we follow the approach proposed in [22]. It is important to distinguish between a *local* score for all transformed edges $e \in E_F$ defined by the transformation of v in J_F and the *accumulated* score over all processed $F \in \mathcal{A}$. The local score is determined by all edges $e \in E_F$. If the processing of an edge e results in a lower score than the local score so far, the total local score should be lowered accordingly. On the other hand, when combining local scores over different affordances F_i the total accumulated score should increase, whenever there is evidence for a present fine-grained affordance F_i. The *T-norm* and *T-conorm* operators [30] have been shown to work best for the desired properties of local and accumulated scores [31].

The local score r for one transformed edge $e \in E_F$ is determined as

$$r = \begin{cases} \exp\left(-\frac{(\theta_d - \theta_c)^2}{2\sigma_\theta^2}\right), & \text{if a collision occurs,} \\ \epsilon, & \text{otherwise} \end{cases} \tag{3}$$

where σ_θ is a threshold value and $\epsilon > 0$ is a low default score. We use the *T-norm* operator

$$T(r, q) = rq \tag{4}$$

to combine local scores r and q over all $e \in E_F$ of an affordance F. At initialization, $r = 1$, however, at later steps r is the previously computed local score, whereas q is the next local score computed to be combined with r. Further, the *T-conorm* operator

$$S(r, q) = r + q - rq \tag{5}$$

is used to accumulate all local scores of all detected affordances $F \in \mathcal{A}$. For this operator, $r = 0$ at initialization, or, at later steps, the accumulated score, whereas q is the next score to be accumulated.

4.5 Object Hypothesis Segmentation

We select the pose with the best score according to Eq. 5 for each hypothesis position. All neighboring hypothesis positions around the selected pose that lie

within the agent model radius are omitted, since we do not want to obtain intersecting agent goal poses H_g.

After obtaining the affordances and highest rated poses, the partition of the scene exhibiting that affordance is segmented. We use a region growing algorithm where the position of the detected affordance serves as seed point. Each point below a certain Euclidean distance is added to the segmented scene part. A low value is well suited to close small gaps in the point cloud, but at the same time limit the segmentation result to one object. Further, points close to the floor are ignored. The segmentation result is shown in Fig. 4(d).

5 Evaluation

This section describes the different datasets used for the evaluation of our approach. Further, we present the experiments and the obtained results in this section.

5.1 Datasets

Our approach was evaluation on 3 datasets. These datasets are described in the following.

Real-world Dataset. These data was acquired in our lab. Data acquisition was performed with an RGB-D camera (Kinect version 1) that was moved around an object and roughly pointed at that object's center. In total, we acquired data from 17 different chairs and 3 stools to represent the fine-grained affordances. From these data, we extracted 248 different views of the chairs and 47 different views of the stools. Example views of these objects are shown in Fig. 2. Additionally, negative data (i.e. data without the fine-grained affordances) from 9 different furniture objects was obtained and 109 views of these objects extracted. Negative data includes objects like desks, tables, dressers and a heating element. Example views of negative data are presented in Fig. 6. The whole dataset contains 404 scene views with 295 positive and 109 negative data examples. This data is provided online[1].

Warehouse Dataset. We collected 3D models from Google Warehouse with 431 objects in total. This dataset contains 323 models with sitting affordances (stools, chairs, benches, sofas) and 108 negative examples (other furniture models without sitting affordances).

Grabner Dataset. This dataset is used for comparison with other approaches. It was used in [5] to augment chair recognition by adding affordance cues. This dataset consists of 890 objects, 110 chairs, 720 non-chairs and 60 other sittable objects.

[1] *Real-World* dataset available at http://agas.uni-koblenz.de/data/datasets/furniture_affordances/uni-koblenz_kinect_v1.tar.gz.

Fig. 6. Example scenes without sitting affordances in the evaluation dataset.

5.2 Experiments and Results

All models in the datasets were annotated with fine-grained affordances. Object points representing sitting surfaces, backrests, armrests and headrests were assigned different labels corresponding to fine-grained affordances of the sitting affordance hierarchy $\mathcal{A}_{sitting}$. We evaluate the ability of our algorithm to distinguish sittable objects and also to detect fine-grained affordances as defined in $\mathcal{A}_{sitting}$. Similar to [5] we split the evaluation with their dataset in 2 parts. In the first part only the chair and non-chair objects are used, whereas the second part uses the sittable objects to test the generalization of the approach. To further show the general validity of our approach, we perform each evaluation using 3 differently parameterized agent models corresponding to humans with the body sizes of 1.85 m, 1.75 m and 1.65 m.

Examples of affordance predictions on the *Real-World* dataset are shown in Fig. 7 and some sample poses of the agent on artificial data in Fig. 8. Additionally to the evaluated affordance hierarchy $\mathcal{A}_{sitting}$ we modeled a hierarchy $\mathcal{A}_{lying}$ for lying. The different fine-grained affordances here are lying with or without raised back and with or without raised legs. Examples for this affordances are shown in Fig. 8 (right).

Table 1 presents the results on the *Real-World* dataset. The evaluation results for the *Warehouse* dataset can be found in Table 2. Finally, the results on the *Grabner* dataset are reported in Table 3 (chairs vs. non-chairs) and in Table 4 (other sittable objects). Note that these results were achieved without any training. All the knowledge required for detection is encoded in the simple agent and affordance models.

Fig. 7. Resulting agent poses for some scenes from the *Real-World* dataset.

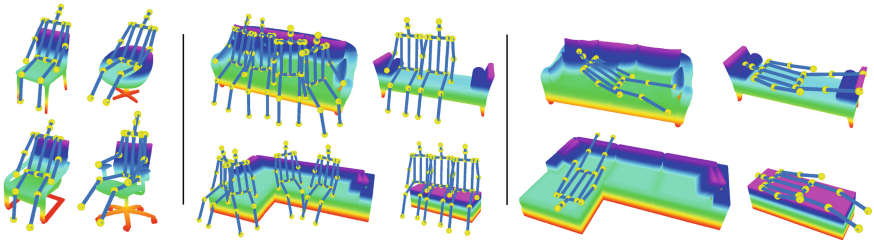

Fig. 8. Sample poses of predicted affordances for sitting on chairs (left), sitting on other furniture (center) and lying (right) on artificial data from the *Warehouse* and the *Grabner* datasets.

Table 1. Prediction results of fine-grained affordances on the *Real-World* dataset.

Agent size	1.85 m			1.75 m			1.65 m		
Metric	f-score	Precision	Recall	f-score	Precision	Recall	f-score	Precision	Recall
Sitting	0.95	1	0.90	0.95	1	0.91	0.95	1	0.91
Backrest	0.86	0.98	0.77	0.86	0.97	0.77	0.86	0.98	0.77
Armrest	0.72	0.66	0.79	0.71	0.64	0.79	0.69	0.64	0.76
Headrest	0.41	0.80	0.28	0.44	0.63	0.34	0.60	0.71	0.52

Table 2. Prediction results of fine-grained affordances on the *warehouse* dataset.

Agent size	1.85 m			1.75 m			1.65 m		
Metric	f-score	Precision	Recall	f-score	Precision	Recall	f-score	Precision	Recall
Sitting	0.95	1	0.90	0.95	1	0.91	0.95	1	0.91
Backrest	0.86	0.98	0.77	0.86	0.97	0.77	0.86	0.98	0.77
Armrest	0.72	0.66	0.79	0.71	0.64	0.79	0.69	0.64	0.76
Headrest	0.41	0.80	0.28	0.44	0.63	0.34	0.60	0.71	0.52

The results on the *Real-World* dataset are promising. Our algorithm is able to find almost all sitting possibilities, while making only little mistakes, as indicated by the results for the *sitting without backrest* and the *sitting with backrest* affordance. The ability of our algorithm to detect these two specialized affordances at the presented high rates speaks in favor of the presented approach. Further, there is almost no difference between the 3 agent model body sizes that were tested.

The results for the fine-grained affordances involving an armrest and a headrest are below the aforementioned ones. F-scores of armrests and headrests indicate that our algorithm successfully differentiates between closely related object functionalities and is able to detect the corresponding fine-grained affordances in RGB-D data. However, the low values indicate that the agent model might need more degrees of freedom during collision detection to better find differently shaped chairs.

Table 3. Prediction results of fine-grained affordances on the *Grabner* dataset (chairs vs. non-chairs).

Metric	f-score	Precision	Recall	Grabner et al. [5] (f-score)
Sitting	0.88	1	0.78	0.53
Backrest	0.75	1	0.60	-
Armrest	0.53	0.54	0.52	-

Table 4. Prediction results of fine-grained affordances on the *Grabner* dataset (other sittable objects).

Metric	f-score	Precision	Recall	Grabner et al. [5] (f-score)
Sitting	0.88	1	0.78	0.53
Backrest	0.75	1	0.60	-
Armrest	0.53	0.54	0.52	-

In contrast to the *Real-World* dataset, the *Warehouse* dataset contains full 3D models. According to the results in Table 2 our approach is again able to distinguish sitting and non-sitting objects on a fine-grained scale. This is supported by the high f-score for *sitting* and *backrest* affordances.

The *armrest* affordance has lower f-score which we attribute to the limited degrees of freedom in the agent model. This drawback will be addressed in future work. The results for most fine-grained affordances are similar along the different agent model sizes, confirming the validity of our approach. However, the *headrest* affordance has significant deviations. We attribute this to the large ambiguity of the presence of a headrest. Depending on where the agent is seated when leaning back (closer or farther away from the backrest) a normal backrest can also serve as a headrest. Additionally, the significantly higher f-score for the smallest agent model indicates that the objects in the dataset are of small size.

The results for the *Grabner* dataset were very similar across different agent sizes. We therefore report their average in Table 3 and in Table 4. Further, we omit the evaluation of the *headrest* affordance, since only very few objects with this affordance were present in the dataset. For the comparison with [5], we report their f-score corresponding to the same recall that we obtained. Considering the results reported in Table 3, our approach seems to fail. However, in [5] sitting affordances were used as a cue for object recognition, where the main goal was to tell apart chairs from other objects. For a fair evaluation, Grabner et al. scaled these other objects to the typical size of chairs, which was completely justified for their evaluation. However, since our approach is detecting sitting affordances independently of underlying object categories, many of the non-chair objects are recognized as sittable (low precision in Table 3). Indeed, e.g. a huge object with a flat and stable surface would also be considered as sittable by real humans. Still, if all objects were true to scale many of the non-chair objects would have

been identified as not providing a sitting affordance by our approach. Although only trained to recognize chairs, the approach of Grabner et al. was shown to generalize well for other sitting objects due to the additional affordance cues. Our algorithm outperforms [5] on other sittable objects, since it detects sitting affordances per se (Table 4).

6 Discussion

We have shown in this paper that our algorithm is able to differentiate affordances on a fine-grained scale without prior object or plane segmentation. Thus, the presented approach is more general and can be applied to the input data directly. To our best knowledge, no similar approaches exist in the literature that are able to differentiate affordances on a fine-grained scale. This makes it hard (if not impossible) to assess the quality of our approach and compare it to related work. The comparison made with the work of Grabner et al. [5] can only serve as an approximate comparison, since their work and ours had a different goal. We therefore want to give a discussion on certain properties of our algorithm and give a detailed outlook to our ongoing work in that field.

Apart from introducing the notion of fine-grained affordances the biggest difference to related work is that we detect affordances directly. In contrast, e.g. in [5] affordances are learned as properties of objects which allows to augment the classification ability of object recognition. However, our approach is ignorant of any object categories.

While we believe that our approach will also benefit from machine-learning techniques (e.g. by learning the membership functions for the fuzzy sets), at this point we have completely omitted the learning step. This comes at the cost of manually defining "reasonable" values for the fuzzy sets (low effort) and a deformable human model (medium effort). Additionally, this raises the question on the extensibility of the approach. An initial agent pose needs to be provided for any new affordance that is included. However, if an agent model is already available (as for *sitting* affordance) new poses can be added by simply transforming joint values in the corresponding configuration file. As a second step, the joints of interest that are involved in the new affordance description, need to be provided with a minimum and maximum angle for transformation. In total, a new affordance model can be added to the algorithm with moderate effort as we have shown with the *lying* affordance hierarchy.

A more complex extension of the algorithm would be to include a different agent, e.g. a hand for grasping. While the hand itself can be modeled again as an directed rooted tree of joints, the initial hypotheses selection step must be changed completely. Instead of finding potential sitting or lying positions in the height map, for a hand a different hypotheses selection needs to be applied (e.g. finding small salient point blobs). However, as soon as these hypotheses are found, the rest of the algorithm is the same: unfolding joints of the agent and evaluating a cost function that reflects the quality of the predicted affordance. We thus believe that the presented approach is generalizable and well suited for extension.

7 Conclusion and Outlook

In this paper we have refined the term *fine-grained affordances* to better distinguish similar object functionalities. We have presented a novel algorithm that is based on fuzzy sets to detect these affordances modeled in hierarchies of subaffordances. The algorithm has been evaluated on 4 specializations of the *sitting* affordance and examples for predicting 4 specializations of the *lying* affordance were given. We have thus shown that the presented approach is able to differentiate affordances on a fine-grained scale. Since this approach detects affordances independently of underlying object categories it can be regarded as complementary to current state of the art approaches which mostly use affordances only as an additional cue for object recognition.

We believe that an object independent affordance detector could be beneficial in existing object recognition pipelines. The segmented object that results from the affordance prediction is constrained to object classes that provide the detected affordance. Where algorithms for 3D object recognition tend to detect false positive objects, these objects could be discarded due to missing affordances that the recognized object classes should posses. If this object needs to be classified, it does not have to be matched against the whole dataset, but only against object classes exhibiting the found affordance. Thus, we are currently working on combining our 3D object recognition pipeline with the presented approach for affordance detection.

The presented algorithm is ignorant of any object classes, since our goal is to detect affordances. This is evident from the leftmost image in Fig. 7, where the agent is sitting with a backrest although the object it is sitting on does not have one. Clearly, here the environmental constellation (object and wall) provided the detected affordance. This demonstrates a strength of the concept of fine-grained affordances that we will further explore in our future work.

Further, we will investigate how an anthropomorphic agent model can be exploited to detect more fine-grained affordances from other body poses. Fine-grained affordances can also be defined for other agents, e.g. a hand. In that case, *grasping with the whole hand* and *grasping with two fingers* could be distinguished, e.g. for grasp planning for robotic arms. Additionally, fine-grained affordances for grasping actions can include drawers and doors that can be *pulled open* or *pulled open while rotating* (about the hinge). We are currently looking for more examples for fine-grained affordances for different agents, to generalize our approach of fine-grained affordances.

References

1. Gibson, J.J.: The ecological approach to visual perception. Routledge, Abingdon (1986)
2. Chemero, A.: An outline of a theory of affordances. Ecol. psychol. **15**, 181–195 (2003)
3. Turvey, M.T.: Affordances and prospective control: an outline of the ontology. Ecol. psychol. **4**, 173–187 (1992)

4. Hinkle, L., Olson, E.: Predicting object functionality using physical simulations. In: 2013 IEEE/RSJ International Conference on, Intelligent Robots and Systems (IROS), pp. 2784–2790. IEEE (2013)
5. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1529–1536 (2011)
6. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: from sensory-motor coordination to imitation. IEEE Trans. Robot. **24**, 15–26 (2008)
7. Ridge, B., Skocaj, D., Leonardis, A.: Unsupervised learning of basic object affordances from object properties. In: Computer Vision Winter Workshop, pp. 21–28 (2009)
8. Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B.: Functional object class detection based on learned affordance cues. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 435–444. Springer, Heidelberg (2008). doi:10.1007/978-3-540-79547-6_42
9. Kjellström, H., Romero, J., Kragić, D.: Visual object-action recognition: inferring object affordances from human demonstration. Comput. Vis. Image Underst. **115**, 81–90 (2011)
10. Lopes, M., Melo, F.S., Montesano, L.: Affordance-based imitation learning in robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2007, pp. 1015–1021. IEEE (2007)
11. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using object affordances to improve object recognition. IEEE Trans. Auton. Ment. Dev. **3**, 207–215 (2011)
12. Wünstel, M., Moratz, R.: Automatic object recognition within an office environment. In: CRV, vol. 4, pp. 104–109. Citeseer (2004)
13. Jiang, Y., Saxena, A.: Hallucinating humans for learning robotic placement of objects. In: Desai, J.P., Dudek, G., Khatib, O., Kumar, V. (eds.) Experimental Robotics, vol. 88, pp. 921–937. Springer, Heidelberg (2013). doi:10.1007/978-3-319-00065-7_61
14. Gupta, A., Satkin, S., Efros, A., Hebert, M., et al.: From 3D scene geometry to human workspace. In: 2011 IEEE Conference on, Computer Vision and Pattern Recognition (CVPR), pp. 1961–1968. IEEE (2011)
15. Şahin, E., Çakmak, M., Doğar, M.R., Uğur, E., Üçoluk, G.: To afford or not to afford: a new formalization of affordances toward affordance-based robot control. Adapt. Behav. **15**, 447–472 (2007)
16. Seib, V., Knauf, M., Paulus, D.: Detecting fine-grained sitting affordances with fuzzy sets. In: Magnenat-Thalmann, N., Richard, P., Linsen, L., Telea, A., Battiato, S., Imai, F., Braz (eds.) Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SciTePress (2016)
17. Sun, J., Moore, J.L., Bobick, A., Rehg, J.M.: Learning visual object categories for robot affordance prediction. Int. J. Robot. Res. **29**, 174–197 (2010)
18. Hermans, T., Rehg, J.M., Bobick, A.: Affordance prediction via learned object attributes. In: International Conference on Robotics and Automation: Workshop on Semantic Perception, Mapping, and Exploration (2011)
19. Seib, V., Wojke, N., Knauf, M., Paulus, D.: Detecting fine-grained affordances with an anthropomorphic agent model. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 413–419. Springer, Cham (2015). doi:10.1007/978-3-319-16181-5_30

20. Maier, J.R., Ezhilan, T., Fadel, G.M.: The affordance structure matrix: a concept exploration and attention directing tool for affordance based design. In: ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, pp. 277–287. American Society of Mechanical Engineers (2007)
21. Maier, J.R., Mocko, G., Fadel, G.M., et al.: Hierarchical affordance modeling. In: DS 58–5: Proceedings of ICED 09, the 17th International Conference on Engineering Design, vol. 5, Design Methods and Tools (pt. 1), Palo Alto, CA, USA, 24–27 08 2009 (2009)
22. Stark, L., Bowyer, K.: Function-based generic recognition for multiple object categories. CVGIP: Image Underst. **59**, 1–21 (1994)
23. Bar-Aviv, E., Rivlin, E.: Functional 3D object classification using simulation of embodied agent. In: BMVC, pp. 307–316 (2006)
24. Chemero, A., Turvey, M.T.: Gibsonian affordances for roboticists. Adapt. Behav. **15**, 473–480 (2007)
25. GESIS: Wie groß sind Sie? allgemeine bevölkerungsumfrage der sozialwissenschaften allbus 2014 (2015). http://de.statista.com/statistik/daten/studie/278035/umfrage/koerpergroesse-in-deutschland/. Accessed 26 Jan 2016
26. Bogin, B., Varela-Silva, M.I.: Leg length, body proportion, and health: a review with a note on beauty. Int. J. Environ. Res. Public Health **7**, 1047–1075 (2010)
27. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
28. Pan, J., Chitta, S., Manocha, D.: FCL: a general purpose library for collision and proximity queries. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 3859–3866 (2012)
29. Hornung, A., Wurm, K.M., Bennewitz, M., Stachniss, C., Burgard, W.: OctoMap: an efficient probabilistic 3D mapping framework based on Octrees. Autonomous Robots (2013). http://octomap.github.com
30. Bonissone, P.P., Decker, K.S.: Selecting uncertainty calculi and granularity: an experiment in trading-off precision and complexity. Uncertainty in Artificial Intelligence (1985)
31. Stark, L., Hall, L.O., Bowyer, K.: Investigation of methods of combining functional evidence for 3-D object recognition. In: Intelligent Robots and Computer Vision IX: Algorithms and Techniques (1991)