Physically Grounded Spatio-Temporal Object Affordances

Hema S. Koppula and Ashutosh Saxena

Department of Computer Science, Cornell University.
{hema,asaxena}@cs.cornell.edu

Abstract. Objects in human environments support various functionalities which govern how people interact with their environments in order to perform tasks. In this work, we discuss how to represent and learn a functional understanding of an environment in terms of object affordances. Such an understanding is useful for many applications such as activity detection and assistive robotics. Starting with a semantic notion of affordances, we present a generative model that takes a given environment and human intention into account, and *grounds* the affordances in the form of spatial locations on the object and temporal trajectories in the 3D environment. The probabilistic model also allows uncertainties and variations in the grounded affordances. We apply our approach on RGB-D videos from Cornell Activity Dataset, where we first show that we can successfully ground the affordances, and we then show that learning such affordances improves performance in the labeling tasks.

Keywords: Object Affordances, 3D Object Models, Functional Representation of Environment, Generative Graphical Model, Trajectory Modeling, Human Activity Detection, RGBD Videos.

1 Introduction

Functional understanding of an environment through object affordances is important for many applications in computer vision. For example, reasoning about the interactions with objects helps in activity detection [28, 43, 18], understanding the spatial and structural relationships between objects improves object detection [16] and retrieval [8], and understanding what actions are supported by the objects in an environment is essential for many robotic applications [17, 30]. Our goal is to learn a rich functional representation of the environment in terms of object affordances from RGB-D videos of people interacting with their surrounding environment.

The definition of 'affordance' had been hotly debated first in philosophy [9, 33, 32], and then in psychology (e.g., [4]). While the intuitions behind all these debates were similar, the interpretations vary from purely symbolic and abstract [9] to more physically-grounded meanings [32]. Recent works in computer vision have revisited these aspects. For example, the symbolic notion of affordances can be interpreted as an object attribute labeling problem [25, 36, 7, 3, 42], and more

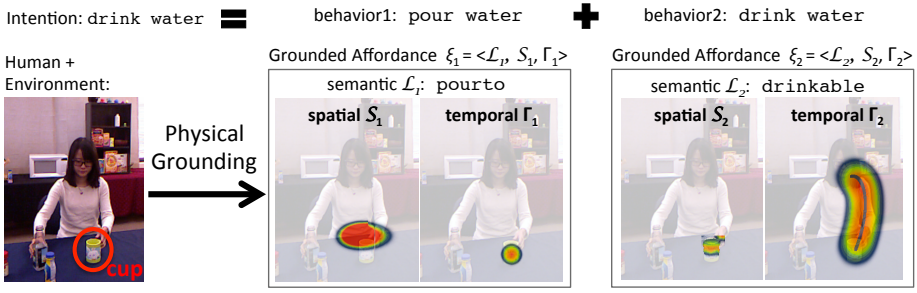


Fig. 1. Grounded affordance for an object `cup`. Given an intention I of a human H in an environment E, our approach outputs the sequence of physically-grounded affordances ξ_k for the objects in the scene. In this figure, we show the affordance of a `cup` for the intention of `drinking water`. The grounded affordances comprise semantic affordances $\mathcal{L}_k$, 3D spatial affordances $\mathcal{S}_k$ and 3D temporal trajectories Γ_k. Due to noise and uncertainty in the agent's behavior, several groundings are valid, and therefore our approach outputs a *belief* over the possible groundings. In this figure, the belief is represented by heatmaps for the spatial and temporal affordances.

recently, physical aspects have been explored in [10, 12, 16], where they model the functionality-based spatial interactions of humans with their environments. For example, Grabner et al. [10] uses the interactions between a sitting human pose and the environment to identify *sittable* regions, Delaitre et al. [5] observed people to extract semantic and geometric affordances for large furniture-like objects, and Jiang, Koppula and Saxena [16] uses the spatial affordances of objects with respect to possible human poses for the task of labeling objects. These works only consider the spatial aspect of static affordances.

In contrast, affordances are often dynamic, have a temporal motion aspect, and they vary depending on the environment and the intention of the human. Consider the cup in Fig. 1, where in order to `drink water` the affordance is `pour-to` for it to receive water, and then it is `drinkable` for transferring water into the human's mouth. The actual 3D coordinates of the interactions with the object and the object's 3D trajectory would vary depending on the geometry of the environment. Furthermore, if the intent of the human was to hurt someone, the cup could also be used as a projectile to throw at someone! Capturing these dynamic and temporal aspects of affordances is necessary in many applications. For example, assistive robots need to reason about 'what can be done with objects?' as well as 'how?' for planning their actions [23, 22].

In this work, we take a unified view where we focus on grounding the affordances into a given environment for a human intention. As illustrated in Fig. 1, by grounding we mean outputting the semantic affordances, the 3D location of interaction on the object ('spatial affordances'), as well as object's motion trajectory ('temporal affordances'). Multiple groundings are valid because of noise and uncertainty in the agent's behavior. We therefore model this uncertainty using a generative probabilistic model for the semantic, spatial and temporal groundings of the affordances. Our generative model is based on a conditional mixture of density functions, where the density functions are discrete (for seman-

tic affordances), product of Gaussians and von Mises (for spatial affordances), and parameterized Gaussian Processes (for the temporal affordances). We train the parameters of our model from the training data comprising RGB-D videos, and test on hold-out test data.

We present extensive evaluation of our proposed affordance learning framework on RGB-D videos form the Cornell Activity Dataset – where we introduce a new affordance dataset consisting of semantic activities along with spatio-temporal motions for several objects. We show that our generative model can reconstruct these trajectories well. We also show that our approach can improve the affordance and activity detection performance on the CAD-120 dataset [23]. The contributions of this paper are as follows:

- We present a representation for affordances that consist of semantic, 3D spatial and *temporal trajectory* components. Our work thus extends previous works that considered only semantic or spatial affordances.
- Our grounding of affordances into spatial and temporal belief maps is context-dependent on the environment and the intention of the agent.
- Our generative probabilistic approach models the uncertainty and variations in the grounded affordances.
- We contribute a new affordance dataset, on which we show that we can predict grounded affordances well. We also show improvement in the labeling performance on an existing RGB-D activity dataset.

2 Related Work

J.J. Gibson [9] described the concept of affordance as the "Action possibilities in the environment in relation to the action capabilities of an actor". The term *affordances* was later appropriated by D. Norman [33] as the *"perceived action possibilities"*. This makes the concept also depend on the actor's goals, plans, values, beliefs, and past experiences. There are other definitions which narrow down the meaning of affordances, for example, *physical affordances* [32] which are perceived only from the physical structure of objects.

Symbolic affordances. There have been many attempts in the computer vision and robotics literature to reason about object functionality (e.g., sittable, drinkable, etc.) instead of object identities (e.g., chairs, mugs, etc.). Most works take a recognition based approach where they first estimate physical attributes/parts and then jointly reasoned about them to come up with an object hypothesis [38]. Some works predict affordance-based or function-based object attributes. For example, [19] consider newspapers and books as *readable* and books and hammers as *hammerable*. Such interpretation was also used in several other works [25, 36, 7]. These works are the first step for a functional understanding of the scene. Our work, in contrast, is focussed on grounding these symbolic affordances.

Scene Understanding: Geometry, Humans and Objects. Physical aspects of affordances have been recently explored in [10, 12, 5, 16]. For example, interactions between a sitting human pose and the environment are used to identify *sittable* regions [10], semantic and geometric affordances of large objects such

as furniture are extracted by observing people [5, 12], and spatial affordances of objects with respect to possible human poses are used for placing and labeling objects [17, 16]. Another notable work is [11], where they looked at how humans manipulate objects for the purpose of recognizing them. These works use particular interpretations of affordances suited to the specific application. In particular, they consider the spatial aspect of static affordances only. In contrast, we consider temporal affordances and infer a belief over the physically grounded affordances. Koppula and Saxena [22] proposed generation of possible future object trajectories for anticipating future activities, where they represent object trajectories as Bézier curves and estimate the parameters from data. However, the Bézier curves can only model limited types of object trajectories. We build upon these ideas and propose a generative probabilistic model which provides a generic framework for modeling various types of affordances and also show that it performs better than [22] for predicting future object trajectories.

Robotics Planning: Navigation and Manipulation. Most of the work in robotics community has focused on predicting opportunities for interaction with an object either by using visual cues [39, 14, 2] or through observation of the effects of exploratory behaviors [31, 35, 13]. For instance, Sun et al. [39] proposed a probabilistic graphical model that leverages visual object categorization for learning affordances and Hermans et al. [14] proposed the use of physical and visual attributes as a mid-level representation for affordance prediction. Aldoma et al. [2] proposed a method to find affordances which depends solely on the objects of interest and their position and orientation in the scene. There is some recent work in interpreting human actions and interaction with objects [26, 1, 20] in context of learning to perform actions from demonstrations. Lopes et al. [26] use context from objects in terms of possible grasp affordances to focus the attention of their recognition system. This work is specific to robotic grasping task. Affordances (i.e., prediction of the object's reaction to robot's touch) have also been used in planning (e.g., [27, 41]). Jain et al. [15] used object affordances for planning user-preferred motion trajectories for mobile manipulators. Misra et al. [29] learned the relation between language and robotic actions. Pandey et al. [34] proposed mightability maps and taskability graphs that capture affordances such as reachability and visibility. However, they manually define affordances in terms of kinematic and dynamic constraints. Recently, Koppula et al. [23, 21] show that human-actor based affordances are essential for robots working in human spaces in order for them to interact with objects in a human-desirable way. They applied it to look-ahead reactive planning for robots. These works in robotics planning are complementary to ours.

3 Affordance Representation and Grounding

Previous formalizations of affordance in literature (e.g., [37]) include defining relation instances of the form $\mathcal{A} = \langle \text{effect}, (\text{object}, \text{behavior}) \rangle$, which state that there exists a potential to generate a certain *effect* by applying the *behavior* on an *object*. Here, the *object* refers to the state of the environment as perceived

by an agent. For example, the lift-ability affordance implies that a *lift behavior* applied to an *object*, say a stone, results in the *lifted effect*, i.e., the stone will be perceived as elevated compared to its previous position. Here, one needs to provide a physical-grounding to each of these elements $\langle \text{effect}, (\text{object}, \text{behavior}) \rangle$. We define one such physical-grounding of these elements for a given agent, intention and the environment.

For physically grounding an affordance we consider the following context: 1) the agent H, which takes into account the physical capability of agent to perform a behavior, for example, a *sittable* object might be too small for the person to perform the *sit* behavior, 2) the intention I of the agent, which determines which affordance of the object is of relevance, for example, the agent wants to *sit* in a chair vs *move* a chair, 3) the environment E, which takes into account the physical constraints to perform a behavior in a particular situation, for example, an object might not be *liftable* if there is another object blocking it from above. This gives us a generic grounded representation ξ of the affordances $\mathcal{A}$ as:

$$\begin{aligned} \mathcal{G}(\mathcal{A}|H, E, I) &= \xi \\ \mathcal{G}(\langle \text{effect}, (\text{object}, \text{behavior}) \rangle | H, E, I) &= \langle \mathcal{L}, \mathcal{S}, \Gamma \rangle \end{aligned} \tag{1}$$

where the symbols denote the following:

$\mathcal{L}$ semantic affordance label, e.g., *pourable*, etc.

$\mathcal{S}$ spatial distribution of the affordance

Γ motion trajectory, 6-dof location/orientation over time

For example, when an object has the *liftable* affordance, the physical grounding of the behavior and its effect are specified by the spatial distribution $\mathcal{S}$ on the object indicating where to hold the object and a vertical motion trajectory Γ for lifting the object, and for the *sittable* affordance, the behavior and its effect are specified by the spatial distribution over the objects indicating where a person can sit on it and a stationary trajectory for the object.

Note that for more complex intentions such as *drinking coffee*, an object can have a sequence of affordances, for example, a cup is first *reachable*, then *movable* and *drinkable*, and finally *placeable*. We denote the sequence with the corresponding symbols in bold $\boldsymbol{\mathcal{A}}$, and denote the k^{th} element in the sequence with a subscripted symbol $\mathcal{A}_k$.

4 Probabilistic Model for Physically Grounding the Affordances

Our goal is to infer the grounding $\xi = \langle \mathcal{L}, \mathcal{S}, \Gamma \rangle$, given the context (H, E, I). In order to model the variations in the grounding, we formulate the grounding inference problem as a probabilistic model $P(\xi|H, E, I)$, where the probability indicates how likely a particular grounding ξ is. During inference time (e.g., for use in some application), one can use the full belief or compute the most likely grounding as: $\xi^* = \arg\max_\xi P(\xi|H, E, I)$.

In detail, we assume the following: (i) the environment, not including the object of interest and the human, is static; and (ii) there is only one active affordance at a given instant of time for a given object. Each intention can

have multiple sequential sub-goals and hence the object can have multiple active affordances in the given sequence of frames.

The relationship between the components of the grounded affordance ξ can be viewed as a graphical model shown in Fig. 2. The k^{th} semantic affordance $\mathcal{L}_k$ depends on human pose H, the environment E and parameters Θ_L, for a given intention I. The spatial affordance $\mathcal{S}_k$ depends on the human pose H_k, the environment E_k, the active semantic affordance $\mathcal{L}_k$ and parameters Θ_S. The parameters for the affordance motion trajectory Γ_k are denoted by θ_Γ and depend on the semantic affordance $\mathcal{L}_k$ as well as the human pose H_k and the environment E_k, as shown by the directed edges. Following the independencies in the graphical model, the joint distribution of all the variables can be written as:

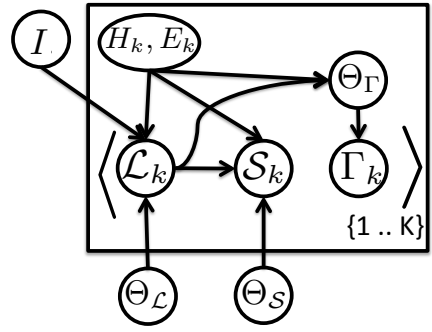


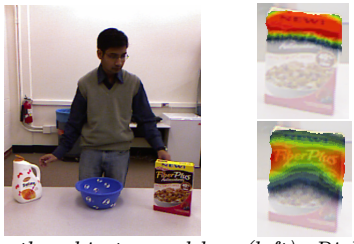
Fig. 2. Our graphical model: For a given intention I, human H and environment E, our model generates the grounded affordance $\langle \mathcal{L}, \mathcal{S}, \Gamma \rangle$. Θ's are the parameters and $k = 1, \ldots, K$ indicates the affordance sequence.

$$P(\langle \mathcal{L}, \mathcal{S}, \Gamma \rangle, \Theta_\Gamma | I, H, E, \Theta_\mathcal{L}, \Theta_\mathcal{S}) = \prod_{k=1}^{K} \underbrace{P(\mathcal{L}_k | I, H_k, E_k, \Theta_\mathcal{L})}_{\text{Semantic Affordance}}$$
$$\underbrace{P(\mathcal{S}_k | \mathcal{L}_k, H_k, E_k, \Theta_\mathcal{S})}_{\text{Spatial Affordance}} \underbrace{P(\Gamma_k | \Theta_\Gamma) P(\Theta_\Gamma | \mathcal{L}_k, H_k, E_k)}_{\text{Temporal Affordance}} \quad (2)$$

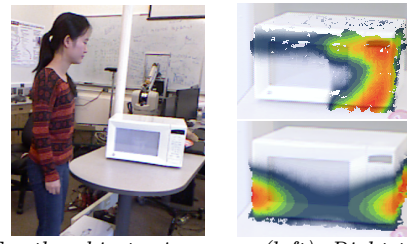
We discretize the time to align with frames of the video and consider the set of time-steps (video frames) corresponding to one intention as one instance of data. Therefore, at each time-step we have the 3D coordinates corresponding to the human, objects and the environment from the video frames. We now describe the conditional distributions and their parameters in more detail.

Semantic Affordance. Each semantic affordance variable $\mathcal{L}_k$ can take a value from $\{1..M\}$, where M is the total number of class labels. We model the probability that the object has an active affordance $l \in \{1..M\}$ given the observations, environment E_k, human poses H_k and the intention I, as a discrete distribution generated from the training data based on the object's current position with respect to the human in the scene (e.g., in contact with hand). For example, if the human is holding the object and intention is to drink water, then the affordances *drinkable* and *pour-to* have equal probability, with all others being 0.

Spatial Affordance. We model the spatial affordance as a potential function which gives high scores to the contact points for the given semantic affordance label. The relative location and orientation of these contact points with respect to the object depends on the activity as well as the human pose. For example, these contact points are usually on the top of the object, say a box, for opening it compared to the sides of the box when moving it (see Fig. 3). Also, which sides of the box would be held depends on the relative orientation of the box with respect to the human. Therefore, for the scoring function to capture these properties we consider the following potentials – distance potentials for modeling the distance



For the object cereal-box (left), Right-top shows the spatial affordance for openable *and Right-bottom for* movable.



For the object microwave (left), Right-top shows the spatial affordance for openable *and Right-bottom for* movable.

Fig. 3. Learned Spatial Affordance Distributions. For the objects in an environment, we show the spatial distribution heat map (red indicates high probability).

between the contact points and the skeleton joints, normalized distance and height of contact points with respect to the object, and angular potentials for modeling the orientation of the contact points with respect to the object and human. The general form of this distribution for a semantic affordance label l given the observations is $P(S_k|\cdots) = \prod_i \psi_{dist_i} \prod_j \psi_{ori_j}$, where ψ_{dist_i} is the i^{th} distance potential and ψ_{ori_j} is the j^{th} relative angular potential. We model each distance potential with a Gaussian distribution and each relative angular potential with a von Mises distribution.

We find the parameters of the affordance potential functions from the training data using maximum likelihood estimation. Since the potential function is a product of the various components, the parameters of each distribution can be estimated separately. In detail, the mean and variance of the Gaussian distribution have closed form solutions, and we numerically estimate the mean and concentration parameter of the von Mises distribution.

Temporal Affordance. We model the object motion trajectory Γ_k as a Gaussian Process with mean trajectory $\mu(\cdot)$ and covariance function $\Sigma(\cdot,\cdot)$ as shown in Eq. (3). The mean trajectory $\mu(\cdot)$, defines the general shape of the trajectory, for example, a circular trajectory for the *stirrer* affordance. The deviation from the mean trajectory is modeled by the covariance function $\Sigma(\cdot,\cdot)$. The mean and covariance functions are parametrized by Θ_Γ.

$$P(\Gamma_k|\Theta_\Gamma) \sim \mathcal{GP}\big(\mu(\cdot;\Theta_\Gamma), \Sigma(\cdot,\cdot;\Theta_\Gamma)\big) \tag{3}$$

The form of the parametrization and the trajectory generation process is explained in more detail in Section 4.1. These parameters depend on the semantic affordance $\mathcal{L}_k$, human poses H_k and the environment E_k, i.e., certain trajectories are more likely for a given semantic affordance, human pose and environment than others. We model this probability $P(\Theta_\Gamma|\mathcal{L}_k, H_k, E_k)$ as:

$$P(\Theta_\Gamma|\mathcal{L}_k, H_k, E_k) \propto \exp\Big(-w^T \phi(\mathcal{L}_k, H_k, E_k)\Big) \tag{4}$$

where w is the weight vector. The features ϕ we consider are the human pose features described in [40] and the relative features of the object w.r.t. the skeleton joints and other objects in the environment, e.g., distance between the object centroid and human joints, distance to the closest object and average distance

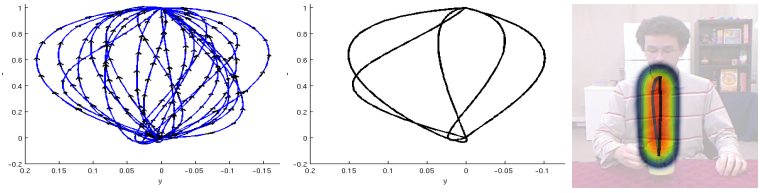


Fig. 4. Illustration of our Trajectory Representation and Generation. Left: The blue lines show the bezier curves fitted to the normalized drinking trajectories from the dataset (the black arrows indicate the direction of motion). Middle: The black lines represent the set of trajectories obtained by clustering the parameters of the drinking trajectories. Right: Drinking activity with predicted trajectory shown in gray and the corresponding probability distribution is shown as the heat map (red corresponds to high probability). The black line denotes the actual ground-truth trajectory.

to all objects in proximity. In the next section, we describe how we represent the mean trajectory function $\mu(\cdot)$, and use it for generating trajectories.

4.1 Trajectory Generation

Objects can follow various types of motion trajectories depending on the active affordance. In this section, we describe the different types of motion trajectories we consider and how we parametrize them to obtain the mean and covariance functions of the Gaussian process. The trajectory types we consider are:
1) Goal-location based trajectories: These trajectories depend on the object's goal location. For example, a cup is moved to the mouth for drinking and moved to a shelf for storing, etc. These trajectories are usually smooth 3D curves.
2) Periodic motion trajectories: These trajectories are characterized by the repetition of certain motions, e.g., a knife is moved up and down multiple times for chopping and a spoon undergoes a periodic circular motion when used to stir.
3) Random motion trajectories: Random trajectories with zig-zag motion are often observed in activities where the goal is not directly related to a particular physical location, for example, cleaning, scrubbing, etc. Even though there might be some repetition/periodic motion in these trajectories owing to their random nature, e.g., scrubbing the same spot over and over again, repetition is not a characteristic property seen in this type of trajectories.

The trajectory types described above cover the majority of the object motions we come across in our daily life. Our framework is generic and other types of trajectories can be included similarly. We now describe how we model each of these trajectory types below.

Goal Based Trajectories. For these trajectories we need to model both the goal location of the object as well as the path taken by the object to reach the goal location. Similar to [22], we model these as parametrized cubic equations, in particular Bézier curves, which are often used to generate human hand like motions [6]. Such a cubic Bézier curve (Eq. 5), is parameterized by a set of four points: the start and end point of the trajectory (L_0 and L_3 respectively), and two control points (L_1 and L_2) which define the shape of the curve.

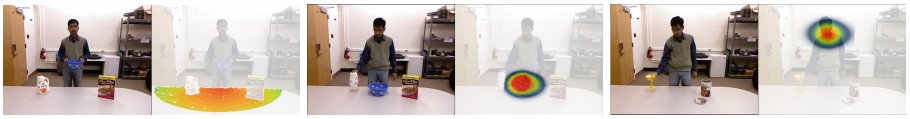


Fig. 5. Distribution over trajectory goal locations. The heatmaps show the distribution over goal locations for *placeability* (left), *pourability* (middle) and *drinkability* (right). The red signifies the most likely goal location for a given affordance.

$$B(x) = (1-x)^3 L_0 + 3(1-x)^2 x L_1 + 3(1-x)x^2 L_2 + x^3 L_3\,, \quad x \in [0,1] \quad (5)$$

We know the current position L_0 of the object, therefore the remaining three control points (L_1, L_2 and L_3) form the trajectory parameters Θ_Γ. Therefore, for the goal based trajectories, the mean trajectory $\mu(\cdot)$ of the Gaussian Process in Eq. 3 is given by Eq. 5 using the estimated parameters.

During learning phase, we first transform and normalize the trajectories in the training data so that all of them have the same start and end points. We estimate the two control points, L_1 and L_2, for each trajectory in the training data. We then cluster these trajectories and obtain a representative set of control points C_l for the affordance class l. Figure 4 shows the trajectories from the training data and those corresponding to the cluster centroids for the *drinking* sub-activity. For a test scenario, we sample the end point L_3 from the distribution over the goal location (as described below) and pair them with the representative set of control points C_l after applying the appropriate inverse transform.

Distribution over goal locations. In order to obtain the probability distribution over the possible goal locations of the object, we define a potential function similar to the one for spatial affordance based on how the object is being interacted with when a particular semantic affordance label is active. We use distance potentials for modeling the distance of the object to skeleton joints and to the other objects in environment and use angular potentials for modeling the orientation with respect to the human pose and other objects in environment, i.e., $P(L_3|\cdots) = \prod_i \psi_{dist_i} \prod_j \psi_{ori_j}$,The parameters are learnt from the training data by maximizing the log likelihood of the goal location for a given active affordance. Figure 5 shows the heatmaps generated for the goal location of the object when its semantic label is *placeable*, *pourable* and *drinkable*.

Periodic Motion Trajectories. For modeling these trajectories, we define two parameterized periodic motion templates – 1) circular and 2) to-and-fro motion. The circular one is used for affordances such as stirring or cycling, and it has four parameters for specifying radius and orientation. The to-and-fro trajectories model the rest of the repetitive motions such as shaking or cutting. These trajectories are parameterized by the curvature, the arc length and the orientation. Similar to the case of the goal-based trajectories, during learning, we compute a representative set of parameters from the training data by clustering and use this basis to represent the trajectories during inference time.

Random Motion Trajectories. These trajectories are the hardest to reconstruct or predict exactly due to their random nature. However, it is easy to

generate trajectories which have similar semantic and statistical properties using information from the environment and human poses. For example, for a cleaning action, the goal is to move the cleaner over the object being cleaned. We therefore, generate a trajectory by randomly selecting 3D locations from the target object which are reachable given the human pose and do not result in collisions in the environment. Note that considering the human pose and environment information is important to obtain semantically meaningful activities as shown later in Section 5.1.

4.2 Inference

We focus on the task of inferring the set of physically grounded affordance $\boldsymbol{\xi}$ for the given human intention and environment, i.e., computing the most likely grounding $\xi^* = \arg\max_\xi P(\xi|H, E, I)$. We take a sampling approach for this, where we generate many samples from the learnt conditional distributions and use the sample with the highest likelihood under the joint distribution in Eq. (2) as the final predicted physical grounding of the affordances.

Given (I, H, E), we first sample the semantic affordance labels $\mathcal{L}_k$ from the discrete distribution. We then sample the contact points from the spatial distribution $P(\mathcal{S}_k|\mathcal{L}_k, H_k, E_k, \Theta_\mathcal{S})$. For grounding the motion trajectory, we first sample the trajectory parameters Θ_Γ depending on the type of the motion trajectory associated with the semantic class $\mathcal{L}_k$ as described in Section 4.1. We then construct the mean and covariance functions with these parameters and sample the motion trajectory from the Gaussian process in Eq. (3).

5 Experiments

We first evaluate our approach on the task of generating physically-grounded affordances, and then we apply our approach to the task of labeling activities.

5.1 Generating Physically-Grounded Affordances

We collected a new physically-grounded affordance dataset. It consists of 130 RGBD videos of humans interacting with various objects. There are a total of 4 subjects, **35 object-types** and **17 affordance types**. The activities include {*moving, stirring, pouring, drinking, cutting, eating, cleaning, reading, answering phone, wearing, exercising, hammering, measuring*} and the corresponding affordances are {*movable, stirrable, pourable, pourto, drinkable, cuttable, edible, cleanable, cleaner, readable, hearable, wearable, exercisable, hammer, hammerable, measurer, measurable*}. We use the OpenNI skeleton tracker, to obtain the skeleton poses in these RGBD videos. We obtained the ground-truth object bounding box labels via labeling and SIFT-based tracking using the code given in [23]. The dataset is publicly available (along with open-source code): http://pr.cs.cornell.edu/humanactivities/data.php. We combined our affordance RGB-D videos with those from the CAD-120 dataset [23] and obtain a total of 815 instances for our experiments.

On this combined dataset, we evaluate the affordance prediction task. Here we are given the human intention (or the activity) that is being performed, the initial human pose and environment at the beginning of the activity. We predict the grounded affordances $\boldsymbol{\xi}$, i.e., the sequence of semantic labels, spatial distribution and the object motion trajectories.

Evaluation Metric. We perform four-fold cross-validation by dividing the data into four folds, with each fold having data belonging to one subject. We train the parameters of our model on three folds and test it on the fourth fold. Specifically, we always test on a *new subject*. We use the following metrics for evaluation:
1) Spatial Likelihood. For evaluating the spatial affordances, we compute the likelihood of the observed contact regions under the predicted distribution.
2) Trajectory Metric. For evaluating the quality of the predicted temporal affordance, we compute the modified Hausdorff distance (MHD) as a physical measure of the distance between the predicted object motion trajectories and the true object trajectory from the test data.[1]

Baseline Algorithms. We compare our method against the following baselines:
1) Chance: It selects a random training instance for the given human intention and uses its affordances as the predictions.
2) Nearest Neighbor Exemplar: It first finds an example from the training data which is the most similar to the test sample and uses the affordances from that training sample as the predicted affordances. To find the exemplar, we perform a nearest neighbor search in the feature space for the first frame, using the features described in Section 3.
3) Koppula et al. [22]: This method models the goal-based trajectories with Bézier curves (Eq. 5). The L_1 and L_2 parameters are learnt from the trajectories in the training data and the object's target location is modeled using the spatial distribution over goal locations as described in Section 4.1. This method does not model the uncertainty in the trajectories.
4) Data-driven uniform sampling: We first compute the set of possible trajectory parameters, Θ_Γ, from the data and then uniformly sample parameters for trajectory generation.
5) Our model - Estimated Goal: Our model where we estimate the goal location and sample the rest of the trajectory parameters from a uniform distribution.
6) Our model - Estimated Goal + Straight Line Trajectory: Our model where we predict straight line trajectories to the estimated goal location.
7) Our method - Estimated Goal + Near Neighbor Trajectory: Our model where we estimate the goal location and use the remaining trajectory parameters of the Nearest Neighbor Exemplar described above.
Results. Table 1 shows the results for predicting the temporal affordance. It shows that the baseline methods give quite high error. However, using our es-

[1] The MHD metric allows for local time warping by finding the best local point correspondence over a small temporal window. When the temporal window is zero, MHD is same as Euclidean distance between the trajectories. We normalize the distance by the length of the trajectory in order to compare performance across trajectories of different lengths.

Table 1. Temporal Affordance Evaluation. Over 4-fold cross-validation, testing on a new subject in each time, we report the error (in centimeters) in predicting the temporal affordances. In addition to MHD metric (see text), we also report the error in predicting the end-point for goal-based trajectories.

Model	End Point Dist. for Goal-based traj.	Error per Trajectory Type (MHD in cm)			Average Error (MHD in cm)
		Goal-based	Periodic	Random	
Chance	56.4	53.3	67.0	75.4	65.2
Nearest Neighbor (NN)	32.7	32.0	40.3	36.4	36.2
Data-driven uniform sampling	47.7	20.6	34.6	50.8	35.3
Koppula et al. [22]	19.4	11.8	-	29.4	20.6
Our Method - Est. Goal + Str. Line Traj.	19.4	10.8	15.7	19.5	15.3
Our Method - Est. Goal + NN	19.4	9.3	13.5	19.5	14.1
Our Full Method	19.4	**8.9**	**10.3**	**19.5**	**12.9**

timation method gives significant improvements. Examples of the observed and estimated trajectories for the various affordances can be seen in Fig. 6. We discuss the following aspect of our approach in more detail.

1) Going beyond target locations for modeling spatial affordances. Our model allows learning affordance-dependent interactions with objects. For example, Fig. 3 shows the learnt spatial interaction heatmap for the *openable* and *movable* affordances. This goes beyond the target location prediction proposed by Koppula et al. [22], where they predict the reachable regions on the objects by predicting the target location of the hand joints. In their work, it only helped in anticipating the most likely *reaching trajectory* for a short duration. However, for the purpose of object manipulation, it is desirable that the reached locations on the objects support the intention of the reach, for example, if the intention is to open the object, the reached location should allow for opening the object, and this is captured by our spatial affordance model as can be seen in Fig. 3. We evaluate the spatial affordances by computing the likelihood of the observed contact points under the learnt distributions. We obtain an average likelihood of 0.6 for the observed contact points compared to a likelihood of 3.1×10^{-5} for randomly chosen contact points on the object.

2) How important is modeling the probability over possible trajectories? As compared to the baselines which select a trajectory (or the corresponding parameters) from the training data, our method achieves a significant reduction in trajectory error metric. This shows the importance of modeling the uncertainties and variations in the temporal affordances that vary with the human intentions as well as with the surrounding environments.

3) Choice of mean trajectory function. Our model reduces the trajectory error metric significantly compared to the baseline methods, even when we approximate the trajectories as straight lines (Table 1-row 5). However, by incorporating the shape of the trajectories into the mean functions (Bézier curves as mean functions for goal-based trajectories and the curvature parameter for periodic trajectories), we can achieve further reduction in the error metric. Also, as we can see from Fig. 6, goal-based trajectories are easier to estimate as they are more deterministic in nature, but the rest have a large variation in the way the objects are moved, for example in the cleaning or shaking activities. Our approach allows us to cope with these variations in a principled way by using appropriate mean functions to modeling the different trajectory types.

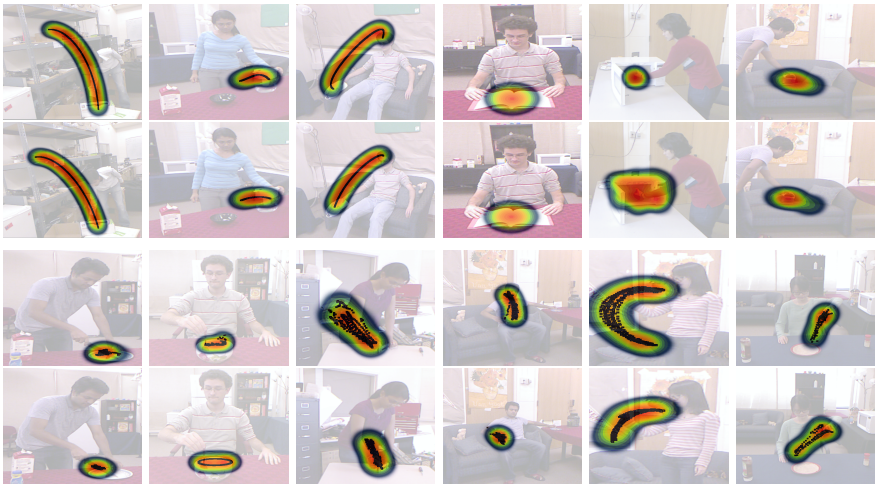
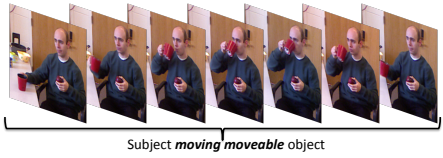


Fig. 6. Learned Temporal Affordance Distributions. The images show the observed trajectories (rows 1 and 3) and the corresponding predicted trajectories (rows 2 and 4) as sampled trajectory points and the distribution as a heatmap. Red signifies higher probability. The affordances in top two rows are (left to right): *placeable*, *pourable*, *wearable*, *readable*, *cleaner* and *cleaner*, and the bottom two rows are: *cutter*, *stirrer*, *hammer*, *shakable*, *exercisable* and *salter*. The trajectory points of the *cleaner* affordance are not shown for clarity.

5.2 Activity and Semantic Affordance Labeling

Previous activity labeling approaches [23, 22] heavily rely on human poses for temporal segmentation and labeling, which sometimes miss boundaries between sub-activities and result in labeling errors. Koppula et al. [23] show that good temporal segmentation is very important for the labeling task. We show that using additional information in the form of grounded affordances can provide an important cue for temporal segmentations. As can be seen in Fig. 7-right, we identify better transitions between the activities using our method, resulting in better labeling. We do this by finding the active affordance for each object in sampled video frames, and identify where changes in active affordances occur.

The intuition behind this is that a change in the active affordance of an object usually happens with change in the current activity. Therefore, our grounded affordances can be used to identify temporal boundaries with high probability. Note that here, *we do not know the human intention* as the video is not labeled. To find the active affordance, we compute the likelihood of the observations under our learned affordance model and take the one which has the highest value. This gives us temporal boundaries in the video based on the active affordances. We evaluated our approach on the CAD-120 dataset [23], which has 4 subjects performing 120 high-level activities and each high-level activity is a sequence of sub-activities. We take the labeling output of [24] and modify it by including the temporal boundaries computed as above. This gives us a new segmentation hypothesis, which we label using the full energy function described in [24].



When physically-grounded affordances are not considered, the drinkability affordance is missed, resulting in erroneous labels.



Using our spatial and temporal grounding of the drinkability affordance results in detecting the correct labels.

Fig. 7. Physically-grounded affordances for activity and semantic affordance labeling. The labeling results generated using the labeling algorithm of [24] for the *having meal* activity from the CAD-120 dataset is shown on the left. We identify the active affordances of objects using our approach, and use this additional information to improve labeling performance. The image sequence on the right marks the frames where the active affordance of the cup is detected as *drinkable*.

Table 2 compares the labeling metrics for the various segmentation methods which use uniform length segmentations, heuristic segmentation hypotheses [23], energy function based segmentation [24], and our method of using additional affordance based segments. Our approach improves the f1-scores for semantic affordance labeling as well as activity detection. We observe that our grounded affordance model mainly helps in improving precision and recall values of infrequent classes.

Table 2. Activity Detection Results. 4-fold cross validation results on CAD-120 dataset (tested on a new subject).

model	Sub-activity Detection		Object Affordance Detection	
	Accuracy	f1-score	*Accuracy*	f1-score
chance	10.0	10.0	8.3	8.3
Uniform+Heuristic[23]	68.2	66.3	83.9	69.6
Koppula et al. [24]	70.3	70.2	**85.4**	71.9
Our Method	**70.5**	**71.2**	84.6	**72.6**

6 Conclusion

Our work extended the affordance-based understanding of objects, where we considered grounding the affordances into a given environment as: the semantic affordances, the spatial locations and the temporal trajectories in the 3D space. We presented a generative probabilistic graphical model for modeling the uncertainty in the grounded affordances. Our model used Gaussian Processes for representing the uncertainty in the trajectories. We then evaluated our approach on predicting the grounded affordances and showed that our approach improves performance on labeling activities.

There are several directions for future work: 1) The space of objects and affordances is significantly richer than what our work have considered—scaling to a larger and richer object and affordance set would be useful; 2) There are many possible applications of our grounded object affordances approach. While we have considered RGB-D activity detection, this approach could be useful in the area of human-robot interaction, as well as in other applications such as 3D scene understanding, robot planning, function-based object retrieval, and so on.

Acknowledgements. We thank Hakim S. and Xingyu X. for their help with the data collection. This work was supported by ARO award W911NF- 12-1-0267, Google PhD Fellowship to Koppula, and NSF Career Award to Saxena.

References

1. Aksoy, E.E., Abramov, A., Dörr, J., Ning, K., Dellen, B., Wörgötter, F.: Learning the semantics of object-action relations by observation. IJRR 30(10), 1229–1249 (2011)
2. Aldoma, A., Tombari, F., Vincze, M.: Supervised learning of hidden and non-hidden 0-order affordances and detection in real scenes. In: ICRA (2012)
3. Anand, A., Koppula, H., Joachims, T., Saxena, A.: Contextually guided semantic labeling and search for 3d point clouds. IJRR (2012)
4. Borghi, A.: Object concepts and action: Extracting affordances from objects parts. In: Acta Psyhologica (2004)
5. Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: ECCV (2012)
6. Faraway, J., Reed, M., Wang, J.: Modeling three-dimensional trajectories by using bezier curves with application to hand motion. J Royal Stats Soc Series C-Applied Statistics 56 (2007)
7. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
8. Fisher, M., Savva, M., Hanrahan, P.: Characterizing structural relationships in scenes using graph kernels. In: SIGGRAPH (2011)
9. Gibson, J.J.: The ecological approach to visual perception. Houghton Mifflin (1979)
10. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR (2011)
11. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR (2007)
12. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR (2011)
13. Hermans, T., Rehg, J.M., Bobick, A.: Decoupling behavior, perception, and control for autonomous learning of affordances. In: ICRA (2013)
14. Hermans, T., Rehg, J.M., Bobick, A.: Affordance prediction via learned object attributes. In: ICRA: Workshop on Semantic Perception, Mapping, and Exploration (2011)
15. Jain, A., Wojcik, B., Joachims, T., Saxena, A.: Learning trajectory preferences for manipulators via iterative improvement. In: Neural Information Processing Systems (NIPS) (2013)
16. Jiang, Y., Koppula, H., Saxena, A.: Hallucinated humans as the hidden context for labeling 3d scenes. In: CVPR (2013)
17. Jiang, Y., Lim, M., Saxena, A.: Learning object arrangements in 3d scenes using human context. In: ICML (2012)
18. Jiang, Y., Saxena, A.: Modeling high-dimensional humans for activity anticipation using gaussian process latent crfs. In: Robotics: Science and Systems (RSS) (2014)
19. Kjellstrom, H., Romero, J., Kragic, D.: Visual object-action recognition: Inferring object affordances from human demonstration. In: CVIU (2011)
20. Konidaris, G., Kuindersma, S., Grupen, R., Barto, A.: Robot learning from demonstration by constructing skill trees. IJRR 31 (2012)
21. Koppula, H., Jain, A., Saxena, A.: Anticipatory planning for human-robot teams. ISER (2014)
22. Koppula, H., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. In: RSS (2013)
23. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. IJRR 32(8) (2013)

24. Koppula, H.S., Saxena, A.: Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In: ICML (2013)
25. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
26. Lopes, M., Santos-Victor, J.: Visual learning by imitation with motor representations. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans 35(3), 438 –449 (june 2005)
27. Lorken, C., Hertzberg, J.: Grounding planning operators by affordances. In: Int'l conf Cog Sys (2008)
28. McCandless, T., Grauman, K.: Object-centric spatio-temporal pyramids for egocentric activity recognition. In: British Machine Vision Conference (BMVC) (2013)
29. Misra, D.K., Sung, J., Lee, K., Saxena, A.: Tell me dave: Context-sensitive grounding of natural language to mobile manipulation instructions. In: Robotics: Science and Systems (RSS) (2014)
30. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: from sensory–motor coordination to imitation. Robotics, IEEE Trans 24(1), 15–26 (2008)
31. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: From sensory–motor coordination to imitation. IEEE Trans Robotics 24(1), 15–26 (Feb 2008)
32. Neisser, U.: Cognition and Reality : Principles and Implications of Cognitive Psychology. W. H. Freeman (1976)
33. Norman: The Psychology of Everyday Things. Basic Books (1988)
34. Pandey, A.K., Alami, R.: Mightability maps: A perceptual level decisional framework for co-operative and competitive human-robot interaction. In: IROS (2010)
35. Ridge, B., Skočaj, D., Leonardis, A.: Unsupervised learning of basic object affordances from object properties. In: Proc 14th Comp.Vision Winter Work. (CVWW) (2009)
36. Russakovsky, O., Fei-Fei, L.: Attribute learning in large-scale datasets. In: ECCV Int'l Work. Parts & Attributes (2010)
37. Sahin, E., Cakmak, M., Dogar, M.R., Ugur, E., Ucoluk, G.: To afford or not to afford: A new formalization of affordances toward affordance-based robot control. Adaptive Behavior 15(4) (2007)
38. Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B.: Functional object class detection based on learned affordance cues. In: ICVS (2008)
39. Sun, J., Moore, J.L., Bobick, A., Rehg, J.M.: Learning visual object categories for robot affordance prediction. IJRR (2009)
40. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: ICRA (2012)
41. Ugur, E., Sachin, E., Oztop, E.: Affordance learning from range data for multi-step planning. In: Epirob (2009)
42. Wu, C., Lenz, I., Saxena, A.: Hierarchical semantic labeling for task-relevant rgb-d perception. In: RSS (2014)
43. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)