Learning Affordance Maps by Observing Interactions

Manolis Savva, Angel X. Chang, Matthew Fisher, Matthias Nießner, Pat Hanrahan
Computer Science Department
Stanford University
{msavva,angelx,mdfisher,niessner,hanrahan}@cs.stanford.edu

Abstract

We address the problem of predicting affordances for dense 3D geometry scans of real-world scenes. Using an RGBD camera setup we observe people interacting with objects and learn the correlation between body part positions and interacted geometry. We encode this information as affordance maps *over 3D geometry and predict affordances for novel scenes where no observations are available.*

1. Introduction

We present a work in progress report for a project aiming to learn affordances directly from observations of people in indoor scenes. The concept of affordances has been shown to be useful for describing potential human actions [3]. Prior work has leveraged affordances for a variety of applications: improving human robot interaction [12], planning for object placement with robots [6], and detecting and anticipating human activity in robotics [8, 9]. Other recent work has focused on hallucinating human pose presence using the concept of affordances [5]. In contrast to all this work, we focus on predicting affordances over dense 3D meshes with no annotations. The closest prior work to our project predicts the presence of sittable objects by sampling virtual scenes with a posed 3D model of a person in the seated position [4]. We focus on expanding this latter work by learning directly from observations of people as they perform a variety of actions in real scenes.

Our key idea is that by observing these human interactions with objects in everyday scenes we can empirically estimate the correlation of body part positioning during actions to properties of the scene and its objects. From such observations, we extract likelihoods of interactions and encode them as *affordance maps*: probability functions over over the 3D geometry of a scene describing the likelihood of human actions taking place. For example, the seat of a chair should have a high "sittability" affordance expressed as a high likelihood of hips being in contact with the top surface of the seat. We use this encoded knowledge to predict interactions in scenes without activity observations. We show that we can transfer affordance maps from observed scenes to novel environments.

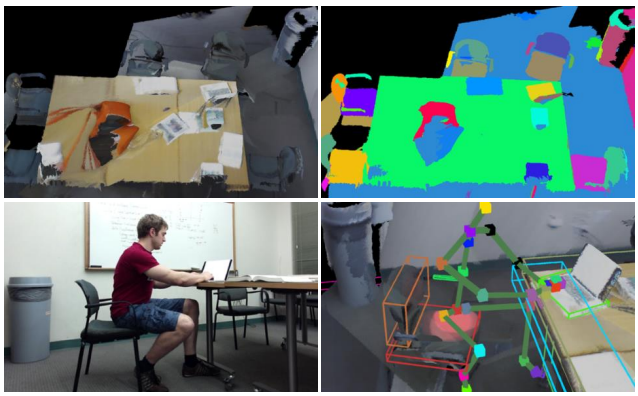

Figure 1. **Top left**: example input reconstructed 3D geometry. **top right:** segmented mesh. **Bottom left**: observed interaction with objects in scanned scene. **Bottom right:** tracked body part positions with activated 3D mesh segments. In this example the person's hips activate the seat surface, table, and laptop.

2. Method

We first obtain 3D reconstructions of scenes by scanning them with the Kinect sensor and using a voxel fusion framework [11]. Once the 3D mesh has been obtained, we oversegment the geometry using a graph segmentation method [2] using surface normals, as described by recent work [7]. See Figure 1 (top row) for an example.

We then place a Kinect One sensor in a static position observing the scene and record people as they carry out common daily activities. The positions of body parts within the 3D scene are tracked using the Kinect sensor and stored in addition to the color and depth frames. We scanned in four scenes: an office, a conference room, and two hallway seating areas. In these scenes, we recorded two subjects (6 sessions per subject) as they carry out common actions including *sitting in chairs, using laptops, reading books and writing on whiteboards*. The average observation time was about two minutes, and the total recording time was approximately 30 minutes.

The recordings were annotated by a volunteer who indicated time ranges over which actions were performed (*e.g.* "reading book", "typing on laptop"). Figure 1 (bottom row) shows an example annotated as "typing at laptop" and "sitting on chair". Using these annotations, we extract from each annotated time range all body part positions and the closest 3D segment within 30cm of that body part. We fit oriented bounding boxes to each segment and compute simple geometric properties such as centroid height over ground, size along each dimension, and aspect ratio between dimensions. We used this set of simple geometric features in our preliminary experiments and for the presented results, but more advanced geometric features can better discriminate between mesh segments.

To predict affordance maps, we train a binary predictor for each body part and each action using the segment features of segments close to the body part during that action. We use a random forest classifier trained with 10 trees using 6 randomly chosen segment features. We predict interactions in two scenes where no observations have taken place: a living room and a bathroom. The scenes are scanned and segmented, and we predict the likelihood of body part contact for each segment (see Figure 2). Despite the simplistic approach, we see plausible predictions for body part and object interactions in novel scenes. The success of this simple method indicates that the input data is highly informative.

3. Future Work

We plan to continue with our data acquisition by collecting a broader variety of scenes and action recordings. We will investigate more advanced learning methods that better account for the geometrical context of actions by jointly considering object segments and features of the human pose. Our longer term goal is to work on grounding action terms such as "reading a book" to the concrete physical properties of the object and body part involved. We aim to construct a semantic representation for actions using these interactions. For example, we would like to learn that "sitting" constitutes resting of the hips on flat surfaces at roughly knee height. This form of attributed representation [10] can be highly beneficial for a variety of applications such as zero-shot learning of human actions [1], and for scene understanding through affordances.

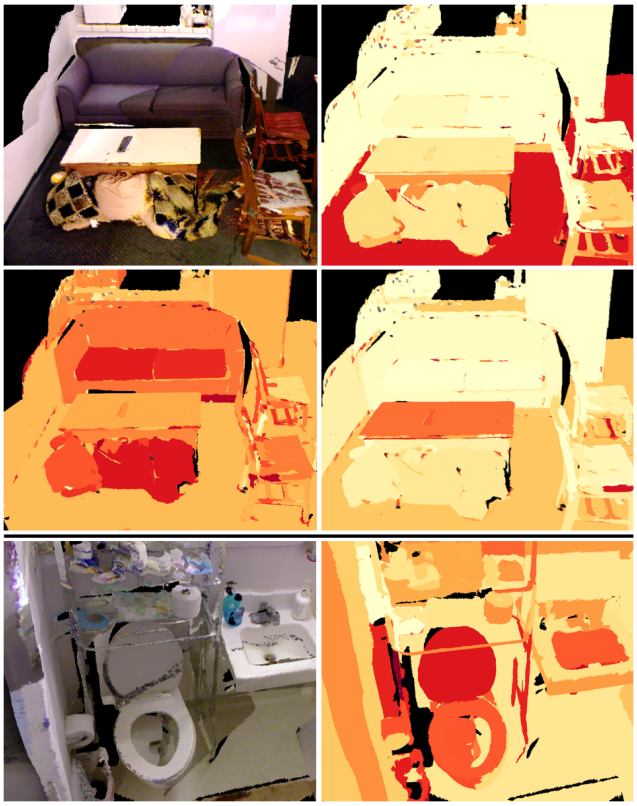

Figure 2. **Top left:** living room 3D scan, **top right:** predicted probability of foot contact during "walking", indicated as red saturation. **Mid left:** probability of hip contact during "sitting", **mid right:** hand contact during "reading". **Bottom left:** input bathroom 3D scene, **bottom right:** probability of hip contact during "reading". Note that there were no observations in these scenes.

References

[1] H.-T. Cheng, M. Griss, P. Davis, J. Li, and D. You. Towards zero-shot learning for human activity recognition using semantic attribute sequence model. In *ACM Pervasive and Ubiquitous Computing*, 2013. 2

[2] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 1

[3] J. Gibson. The concept of affordances. *Perceiving, acting, and knowing*, 1977. 1

[4] H. Grabner, J. Gall, and L. Van Gool. What makes a chair a chair? In *CVPR*, 2011. 1

[5] Y. Jiang, H. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3D scenes. In *CVPR*, 2013. 1

[6] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. In *RSS*, 2013. 1

[7] A. Karpathy, S. Miller, and L. Fei-Fei. Object discovery in 3d scenes via shape analysis. In *ICRA*, 2013. 1

[8] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from RGB-D videos. *IJRR*, 2013. 1

[9] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *RSS*, 2013. 1

[10] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 2

[11] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 2013. 1

[12] A. K. Pandey and R. Alami. Taskability graph: Towards analyzing effort based agent-agent affordances. In *RO-MAN*, 2012. 1