Tactile Logging for Understanding Plausible Tool Use Based on Human Demonstration

Shuichi Akizuki
akizuki@elec.keio.ac.jp
Yoshimitsu Aoki
aoki@elec.keio.ac.jp

Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama,
Kanagawa, Japan

Abstract

In this paper, we propose a novel human-object interaction (HOI) representation for understanding plausible tool use; this is done by using an object-centric method. Everyday tools (such as kitchen or DIY utensils) have the ideal region and direction for being touched by hands and surrounding objects. By analyzing the human demonstrations, tactile histories that occurred on the tools are accumulated to the tool's surface as the *"Tactile Log"*, which can help to achieve the plausible tool use by robot arms. We also proposed a dataset for the evaluation of generated tactile logs, which is called the Object-Centric Interaction (OCI) dataset consists of RGB-D videos of plausible tool use by human and 3D models of the tools used in the videos. We confirm that the precision of our tactile logging is 0.77.

1 Introduction

Having robotic arms achieve human-like use of tools is one of the most important goals of intelligent robotics. One way to achieve this is by recognizing the functionality of the shape of tools, which is called affordance [4, 11, 12, 14, 17]. Humans can handle a tool even if it is their first time seeing that tool. For example, consider a hammer. We can grasp the handle anywhere, but we grasp the end part of the handle (because it makes hitting easier). As another example, consider putting objects on a tray. Any part of the tray can support an object, but we tend to place objects in the center (for stability). Thus, tools have a region to which actions can be applied and an ideal region to which actions should be applied. In this paper, the region where action can be applied is called "the candidate region," and the suitable region for a certain action is called "the ideal region" (see Fig. 1).

Robotic perception should aim to recognize the ideal region for performing specific actions. Several recent approaches [4, 12, 14] have solved the affordance recognition as a segmentation task. However, it is difficult to learn the ideal region for acting due to the lack of clear criteria of the region of each affordance label. Previous affordance recognition methods have learned the entire candidate region. For example, Fig 1(b), which was provided by the public dataset [4], labels candidate regions as the ground truth.

In this paper, we bridge the gap between the candidate region and the ideal region by directly analyzing human demonstrations. This was done by describing tactile events on the

object's surfaces with novel human-object interaction (HOI) representation by using object-centric interaction descriptions. Fig. 2 shows the differences in two types of HOI representation, the human-centric description and our object-centric description. Each representation explains the interaction between a person and a tray. In (a), the human-centric way, action (lift up) is estimated from appearance and spatial relationship between the person and the tray. In (b), the object-centric way, regions touched by hands and another object (mug) were detected on the surface of the tray. The object-centric way was suitable for learning the ideal region for plausible tool use; this was because the actual touching points can be detected on the object coordinate system. For human-centric interactions, a huge number of actions should be considered due to the variations of human poses. From the viewpoint of object-centric interactions, since there is only a difference in the tactile direction of the object surface, the classification problem can be simplified.

The proposed method tracks the 6DoF (6 degrees of freedom) pose, the 3D position and orientation, of a tool and the human pose simultaneously. This was done from RGB-D videos, which showed how a human used the tool. Tactile types (none, grasped, on, in, and under) were coursed on the tool's surface is estimated to frame by frame. Thanks to 6DoF pose tracking, our tactile log allowed the rigid transformation of tools in the temporal domain. Since many methods of describing HOI use human-centric perspectives, our method has an opposite viewpoint. This paper made the following contributions:

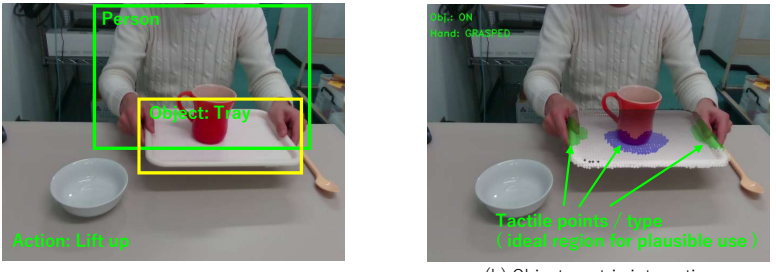
- We proposed a new method of object-centric interaction description for understanding plausible tool use. The proposed method is the first method to describe multiple tactile types on the surface of target objects while tracking the 6DoF pose of the target object.
- We introduced a new RGB-D video dataset, called the OCI dataset, for analyzing the behavior of humans and target objects in a 3D space. The OCI dataset provides 3D models and 6DoF poses of the target objects and human poses in videos. This is novel compared to other HOI datasets (such as HICO or HICO-DET [2, 3]).

2 Related Work

Affordance recognition: There are methods to estimate the functionality of the local parts of tools which are placed in indoor scene[4, 11, 12, 14, 17]. One method proposed by Schoeler et al. [17] estimates tool's function by considering the spatial relationship of each part which is divided from entire 3D shape model. Pixel-wise affordance estimation method (without explicit part division) has also been proposed by Myers et al. [11] employed hand-crafted RGB-D features (such as depth value, curvature, etc.) that were extracted from input

(a) Original image (b) Candidate regions (c) Ideal regions

Figure 1: The difference between candidate regions and ideal regions. Green and red show the regions that can be grasped and can pound, respectively.



(a) Human-centric interaction

(b) Object-centric interaction (proposed in this paper)

Figure 2: Two types of human-object interaction representation of a person and a tray: (a) An example of human-centric description and (b) the object-centric description, which was proposed in this paper. (b) is more suitable for learning of the ideal region for plausible tool use, because the specific tactile points can be detected. Green points indicate the regions that were grasped by hands. Blue points indicate the region that is touched by surrounding objects from above.

RGB-D image; they are classified by using the Structured Random Forests. Recently, the database for affordance estimation has been expanded, and CNN-based methods produced state-of-the-art result[4, 12, 14].

Since the training data of these methods are annotated parts-wise, the entire region to which action can be applied is the output. To perform robotic manipulation, it is necessary to detect the ideal region for action from the detected region by using the post processing of affordance estimation. This problem can be solved by annotating the ideal region to the dataset in advance, but this annotation is not easy to apply immediately. We believe that analyzing HOI provides direct annotations of ideal regions for specific tasks.

Human-object Interaction (HOI): There has been great progress in the field of HOI analysis. Algorithms that can recognize the behaviors of humans and objects simultaneously have been proposed [5, 9, 10, 18, 19]. By using various relationships between the detected bounding box of humans and objects, it has become possible to acquire novel image descriptions. The method introduced in [5] detects <human, verb, object> triplets from image by using appearance feature of human. Shen et al. have achieved the zero-shot learning by training an object detector and an action detector separately [18]. A method of robustly tracking hidden objects by using containment relation, inclusion relation between bounding boxes, has also been proposed [10].

HOI analysis from images and videos provides a rich description of the scene. Since all of the previous methods have described HOI based on the spatial relationship between detected bounding boxes, the spatial resolution of the description is not enough to estimate the ideal region for performing the tasks.

Action Map (AM): AM is data representation that describes the position and type of interaction for the object of interest. Methods for generating AMs on a 3D reconstructed indoor scene [7] and on a 3D object model [8] have been proposed. The AM can be used for various applications (such as estimating the function of an indoor structure; estimating the location, in a reconstructed 3D scene, where a specific action can be performed [6, 13], and generating of likely human poses and the arrangements of surrounding objects, given action

as interaction snapshots [16].

Since all of these methods assumed that the target was a static 3D structure, it was difficult to describe interactions on the surface of moving tools that were grasped by hands. The method proposed by Zhu et al.[20] tracked the 6DoF pose of tools during interactions, but, since it recorded only the presence or absence of tactile interactions, the expression capability (as an AM) was not high enough to understand tool use.

3 Object-centric Interaction

In this section, we describe OCI to generate the tactile log, which was done to accumulate the ideal usage of tools. We aim to understand how to use tools by referring to the tactile log. To do so, our method detects the following three types of information as the tactile log for the target tool M, while performing the 6DoF tracking:

- **Tactile source:** It is assumed that there are two types of objects that can touch M. One is the human hand, and the other is the surrounding objects acting on M.
- **Tactile position:** This denotes points on the M that are touched by tactile sources. Since they are points actually touched, ideal region for plausible tool use can be detected.
- **Tactile type:** We consider three types of tactile direction (on, under and in) along the gravity axis, and we consider the grasping region to be the tactile type. This information is also important to understand how to use the object.

Fig. 3 shows three types of tactile log: (a) is a video frame of tool use; (b) shows the result of 6DoF pose estimation of target tool M, which is projected on an RGB frame by green dots; (c) indicates relationship between M and contacting objects; and (d) shows the tactile log of a frame (a). Each color indicates a different tactile type.

The first row is the moment when a person hit an object with a hammer. In this situation, there is an object *"under"* the hammer, and it touches the hammer's head. A hand is also touching the end of the hammer's grip. This information is described as a tactile log of this frame, which is shown in (d). The second row shows the moment in which a small bowl is placed on a tray. Since the bowl is placed *"on"* the tray plane, tactile direction is *"on"*. The third row shows objects being poured (we used small white balls for stability of depth sensing) into a mug. The label *"in"* is accumulated on the surface of the mug.

4 Approach

4.1 Overview

Our goal is to build the temporal tactile log of 3D objects interacting with hands and surrounding objects while also performing 6DoF tracking. The input of the algorithm is a pair of a RGB-D video sequence $V = \{f_1^d, ..., f_n^d\}$ that shows plausible tool use by a person and a 3D object model M of the tool which is represented by point cloud data. $\mathbf{p} \in M$ is a point of the object model. The output is the tactile type of each point of M, denoted as $I_t(\mathbf{p})$. Our method calculated it for the all frame of the input RGB-D videos.

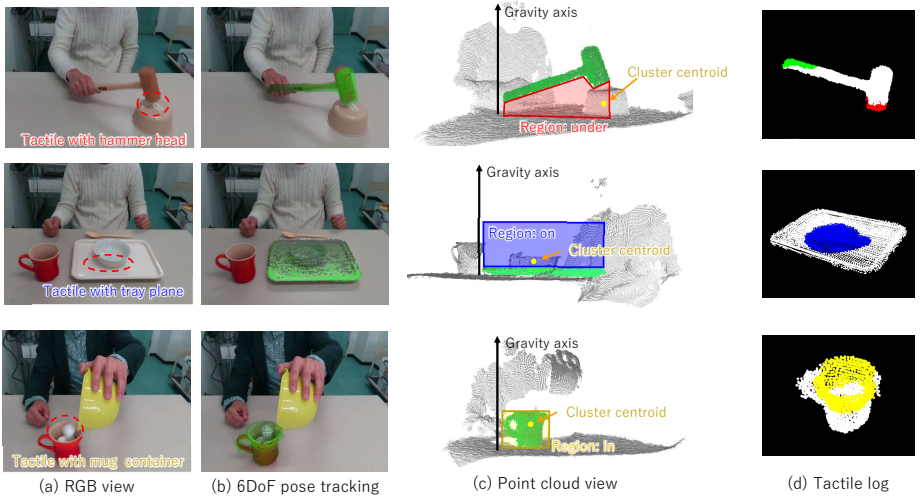


Figure 3: Tactile types, which are decided by considering the spatial relationships between the object model and the interacting objects. (a) Input video frames. (b) The 6DoF pose of object models. (c) The spatial relationship between the object model and the touching objects. (d) Estimated tactile log. Green, red, blue, yellow and white indicate *grasped*, *under*, *on*, *in*, and *none*, respectively.

Fig. 4 shows the framework of proposed tactile log generation. Our framework repeats three modules: (1) Human pose estimation, (2) Object pose estimation, and (3) Tactile type estimation. This is done for each frame. The focus of our method is the third module, which we will describe next.

4.2 Tactile type estimation

This module estimates the tactile type $I_t(\mathbf{p})$ of $\mathbf{p} \in M$ at each frame by using the 6DoF pose R_t and 3D points of human hands, which are estimated by previous modules and are described in 4.3 and 4.4, respectively.

First, we exclude the point cloud belonging to the target object from the input point cloud S_t. The remaining point cloud S_{sub} is generated by subtracting S_t and the transformed object model $M_t = R_t(M)$. We assume that the remaining point cloud S_{sub} includes only point cloud belonging to hands and surrounding objects that can touch M_t. S_{sub} is clustered into multiple segments $C = \{c_i | i = 1, ..., m\}$ by using the Euclid clustering implemented in Point Cloud Library (PCL) [15]. m represents the number of segments. Segment c_i has a point cloud and an attribute c_i^{type} that indicates the object type $\{$"*human*","*object*"$\}$. Segments that have the nearest neighbor to the 3D point of the human hand are labeled as $c_i^{type} =$ "*human*", and the others are labeled as $c_i^{type} =$ "*object*".

$I_t(\mathbf{p})$ is computed by using the nearest neighbor of $\mathbf{p}$, which is denoted as $\mathbf{q} = N(\mathbf{p}, S_{sub})$, and segment c_q, which has $\mathbf{q}$:

$$I_t(\mathbf{p}) = \delta(\mathbf{p}, N(p, S_{sub})), \tag{1}$$

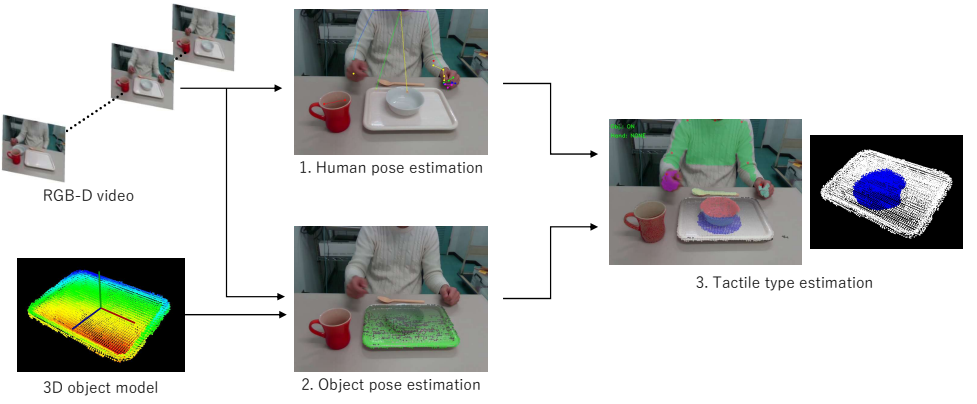


Figure 4: Overview of the proposed method. The input is a pair of RGB-D video and a 3D object model. The output is the tactile log of each frame. By using the 3D position of hands and the 6DoF pose of the object, which are estimated 1. and 2., tactile type is estimated.

$$\delta(\mathbf{p},\mathbf{q}) = \begin{cases} L(c_q, M_t) & (|\mathbf{q}-\mathbf{p}| < th) \wedge (c_q^{type} = \text{“}object\text{”}) \\ \text{“}grasped\text{”} & (|\mathbf{q}-\mathbf{p}| < th) \wedge (c_q^{type} = \text{“}human\text{”}) \\ \text{“}none\text{”} & (otherwise). \end{cases} \tag{2}$$

th is the threshold for deciding whether $\mathbf{p}$ and $\mathbf{q}$ make contact with each other or not. We used $th = 0.03[m]$. $L(c_q, M_t)$ is used to evaluate the spatial relationship between the centroid of segment c_q, which is denoted as $c_q^{centroid}$, and the silhouette of the object model; this is generated by 2D projection along the gravity axis. If $c_q^{centroid}$ is projected on under region of the projection, $L(c_q, M_t)$ returns “*under*”. If the projected point of c_q and M_t are shared, $L(c_q, M_t)$ returns “*in*”. Otherwise, it returns “*on*”.

4.3 Object tracking

This module tracks the 6DoF pose $\mathbf{R}_t$ of the target object M at frame t. Occlusions on the M frequently occurred due to interaction with hands and surrounding objects; thus, we minimize the robust cost function as follows:

$$\tilde{\mathbf{R}} = \arg\min_{\mathbf{R}} \left(\sum_{i=1}^{m} c_{in}(f_t, \mathbf{R}\mathbf{p}_i) + c_{out}(o_r) \right). \tag{3}$$

$c_{in}(f_t, \mathbf{R}\mathbf{p})$ and $c_{out}(o_r)$ evaluate the cost for the inlier and the outlier, denoted as:

$$c_{in}(f_t, \mathbf{R}\mathbf{p}) = \begin{cases} d(f_t, \mathbf{R}\mathbf{p})/th_{dist} & d(f_t, \mathbf{R}\mathbf{p}) < th_{dist} \\ 0 & othewise \end{cases} \tag{4}$$

$$c_{out}(o_r) = \frac{1}{1 + e^{-\sigma(o_r - \mu)}}. \tag{5}$$

where, $d(f,\mathbf{p}) = |f(\mathbf{p}) - p_z|$ represents the distance between the z component of $\mathbf{p}$ and the depth value $\mathbf{p}$ projected on the depth image f_t^d. o_r is the ratio of projection that exceeds the threshold th_{dist}. Equation 3 is minimized by using Particle Swam Optimization (PSO).

4.4 Human pose estimation

To estimate the interaction between the object model and hands, we detect the 3D position of hand joints throughout the video sequence. We employed an accurate human keypoint detector, which had been proposed by Cao et al. [1]. Body keypoints are detected as 2D coordinates, 3D position are reconstructed by using internal parameter of the 3D sensor.

4.5 Tactile type correction using single task assumption

The procedures described above sections are applied for all the video frames, and the tactile log for all frames are generated. After that, false tactile type labels are filtered out by using post-processing. Ideally, the tactile type between the target object and surrounding objects would be limited to one type when human action was a single task (such as hitting or putting). However obtained tactile log may have includes some tactile types due to the errors in pose tracking or the over segmentation in clustering module. We detect the dominant tactile type from generated tactile log and delete other tactile types. The effectiveness of this procedure is evaluated in 5.3.

5 Experiments

5.1 Dataset

This section explains the OCI dataset, which was used in the following experiment. The OCI dataset was composed of 95 videos with 1.5k frames which showed humans using rigid tools and 15 3D models of those tools. We used Intel Realsense SR300 for data correction. The tool categories used in the video were as follows: bowl, dish, hammer, mug, tray, and spoon.

The point-wise ground truth of the tactile types for each object is provided. To track the target object, the 6 DoF parameter of the object is prepared for the first frame.

5.2 Tactile log generation

Fig. 5 shows typical examples of generated tactile logs from our method for qualitative evaluation. For each image pairs, the left shows estimated 6DoF poses of the target objects that were projected in the RGB video frame. The right shows the generated tactile logs of the corresponding frames. Note that, we ignored the touch between the target object and the table. Each color indicates a different the tactile type (white: "*none*", yellow: "*in*", blue: "*on*", red: "*under*"). It was confirmed that the tactile log was generated on the surface, which was actually touched by hands or surrounding objects.

Fig. 6 shows two examples of the "temporal" tactile log. The first row shows an object being pounded by a hammer. In this situation, the hammer is touching the underlying object at the frame (e). Tactile type "*under*" is correctlly labeled at touching surface of the hammer's 3D model. The second row shows a mug that contains objects being lifted. The mug contained objects for all the frames, so the related surface was correctlly estimated as the

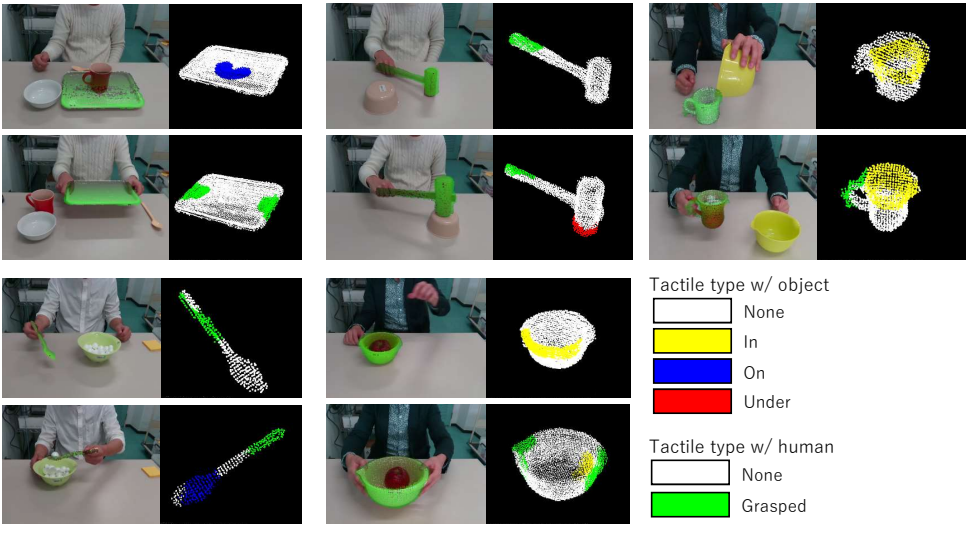


Figure 5: Qualitative results of tactile logs. Left: The recovered 6DoF pose of target objects are projected on the RGB frames. Right: The generated tactile logs. Each color indicates estimated tactile types.

tactile type "*in*". The tactile type "*grasped*" was also correctly labeled in the frames (b) - (f). By referring to the sequential tactile log, we can understand the position to be grasped; this can help in the performance of the specific tasks and how object should be touched. Furthermore, since 6DoF tracking was also performed, we can understand how to moved the grasped object. It seems that the robotic manipulation for plausible tool use can be generated by using our temporal tactile log and the trajectory of 6DoF pose.

5.3 Recognition performance

We evaluated the reliability of the generated tactile logs in each frame. It was assumed that the correct tactile type was labeled as a partial area of the candidate region when the tactile log was generated accurately. Therefore, we employed the precision rate as a reasonable metric for evaluation. To prepare the ground truth data, we manually labeled the correct tactile types in the candidate regions on the object models. The methods used for comparison are as follows.

1. Baseline: This method generates a tactile log using frame-by-frame procedure without post-processing.

2. w/ STA: This method uses tactile log correction using the single-task assumption that was introduced in 4.5.

Table 5.3 shows the precision rate for each method. In the baseline method, incorrect tactile type was generated due to the tracking errors and the over-segmentation, which reduced the precision rate. We confirmed that this error could be suppressed by using the single-task assumption.

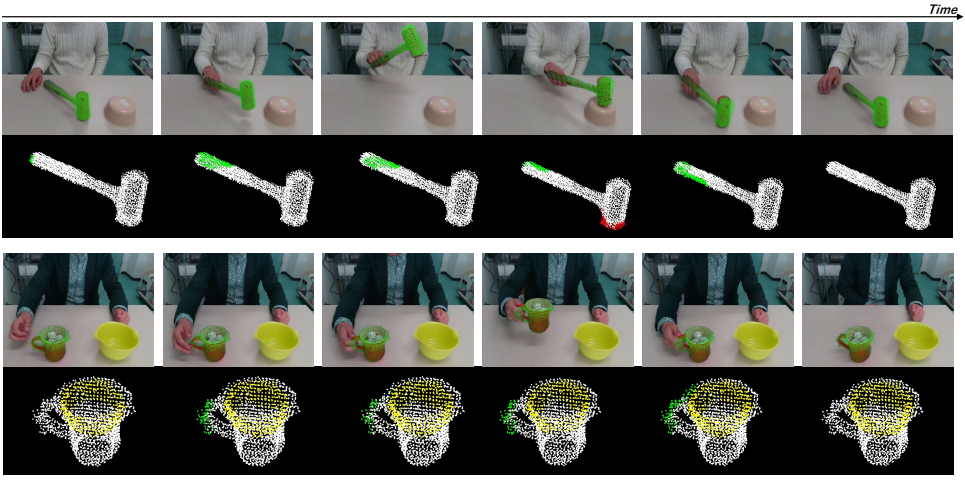


Figure 6: The qualitative results of temporal tactile logs. Our method can simultaneously recover the tactile type and the trajectory of the object.

	Grasped	In	On	Under	Average
Baseline	0.79	0.79	0.57	0.40	0.64
w/ STA	0.81	0.88	0.75	0.65	0.77

Table 1: The precision rate of the tactile log.

The dominant source for the failure tactile type estimation was errors in the 6DoF tracking for the target object. A large part of the target object was hidden when a hand or a surrounding object touched the target object. The 6DoF tracker of the proposed method has the robustness of partial occlusion. However, when the occluded regions were too large, our method did not work well.

6 Conclusion

We have proposed a novel object-centric human-object interaction (HOI) representation, which is called temporal *Tactile Log* for understanding the plausible tool use. The three types of interactions that can be described by our method are as follows: (1) tactile source (human or surrounding object); (2) tactile position, which represents touched point on the surface of the tool; and (3) tactile type, which represents the spatial relationship between the tool and the tactile source. Since these interaction could be detected while tracking the 6DoF pose of the target object, the 3D trajectory of tools were also reconstructed.

In addition, we proposed a novel dataset, called the Object-Centric Interaction (OCI) dataset, for evaluating generated tactile logs. The OCI dataset consists of RGB-D videos that show plausible tool use by humans and 3D models of the tools used in the videos. We confirmed that the precision of our tactile logging was 0.77. In our future work, we will establish a method for generating robotic motions for tool use.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 17K12761.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[2] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. pages 1017–1025, 2015.

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018.

[4] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *in IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[5] K.He G.Gkioxari, R.Girshick P.Dollar. Detecting and recognizing human-object interactions. In *Proc. IEEE International Conference on Computer Vision (CVPR)*, 2018.

[6] H. Grabner, J. Gall, and L.V. Gool. What makes a chair a chair? In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536, 2011.

[7] M. Hanrahan, A.X. Chang, P.Hanrahan, M.Fisher, and M.Nießner. Scenegrok: inferring action maps in 3d environments. *ACM Trans. Graph.*, 33:212:1–212:10, 2014.

[8] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly E. Rushmeier. Tactile mesh saliency. *ACM Trans. Graph.*, 35:52:1–52:11, 2016.

[9] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. What is where: Inferring containment relations from videos. In *25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3418–3424, 2016.

[10] Wei Liang, Yixin Zhu, and Song-Chun Zhu. Tracking occluded objects and recovering incomplete trajectories by reasoning about containment relations and human actions. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[11] Austin Myers, Ching Lik Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381, 2015.

[12] Anh Nguyen, Dimitrios Kanoulas, Darwin G. Caldwell, and Nikos Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[13] Nicholas Rhinehart and Kris M. Kitani. Learning action maps of large environments via first-person vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016.

[15] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[16] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew C Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Trans. Graph.*, 35:139:1–139:12, 2016.

[17] Markus Schoeler and Florentin Wörgötter. Bootstrapping the semantics of tools: Affordance analysis of real world objects on a per-part basis. *IEEE Transactions on Cognitive and Developmental Systems*, 8:84–98, 2016.

[18] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. 2018.

[19] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. Modeling 4d human-object interactions for event and object recognition. *Proc. IEEE International Conference on Computer Vision (CVPR)*, pages 3272–3279, 2013.

[20] Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2855–2864, 2015.