Audio-Tactile Annotation and Registration of 3D Point Clouds for Robotic Manipulation

Jivko Sinapov, Darren Earl, Derek Mitchell, Heiko Hoffmann
HRL Laboratories, LLC
Malibu, CA 94025
jsinapov@iastate.edu, {djearl, dwmitchell, hhoffmann}@hrl.com

***Abstract*—The recent advent of low-cost 3D sensing technologies has greatly increased the use of 3D point cloud-based representations in robotics. Such representations have a variety of applications, including object recognition, pose estimation and grasp point selection. A major limitation of 3D point clouds, however, is that they fail to capture an object's functional features – for example, a 3D model of a stapler does not in itself encode where the stapler should be pressed. To bridge the gap between an object's shape and an object's function, this paper proposes an approach for robotic annotation of 3D point clouds that captures the object's functional features. In our experiments, the robot explored objects by performing a variety of behaviors on them at different locations and observed the auditory and tactile sensations as a result of its actions. Using 3D registration algorithms, the resulting auditory and tactile point clouds were matched against the 3D object point cloud, resulting in a multi-modal point cloud representation encoding the spatial locations on the 3D model that afford the robot to actualize the object's function. The results show that the approach successfully discovers the functional locations in a 3D point cloud for three different types of objects: a stapler, a drill and a flashlight.**

I. Introduction

Autonomous robot manipulation of tools is a long standing goal of the robotics community [1]. For many tools (e.g., a hand-held drill, a flashlight, a stapler, etc.) successful manipulation requires that the robot is capable of detecting the specific location on the object that needs to be actuated (e.g., a trigger or a button) in order to turn on the tool. Yet, to date, virtually all methods for robotic tool use require that the human programmer specify what part of the object needs to be actuated and therefore, such methods cannot generalize to novel objects that have not been previously seen by the human programmer. Furthermore, most object representations used by robots consists entirely of vision-based 2D and 3D object models, which, by themselves, do not encode the location on the object that needs to be actuated.

Towards addressing these challenges, this paper proposes a method that enables a robot to detect the functional parts of hand-held tools using tactile and auditory sensory feedback coupled with exploratory behaviors that the robot applies on the objects. The proposed approach consists of extracting auditory and tactile events detected while the robot manipulates an object and mapping those events onto the object's 3D point cloud model. The result of this process is a novel visual-audio-tactile point cloud object representation which not only captures the object's visual appearance, but also encodes the functional components of the object (e.g., buttons) that afford tool actuation.

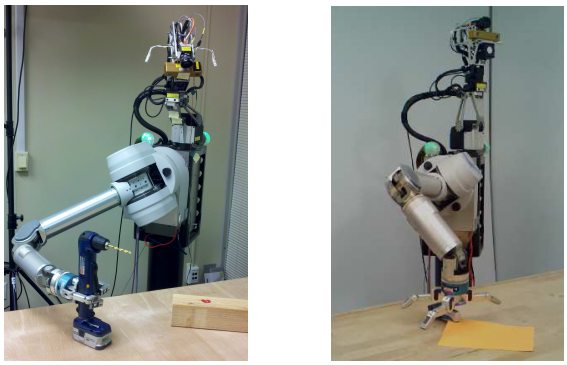


Fig. 1. The ARM-S robotic platform, shown here performing two of the manipulation challenges: drilling (left) and stapling (right).

The proposed method was tested on an upper-torso humanoid robot, shown in Figure 1, and three different tools: a hand-held drill, a flashlight, and a stapler. By exploring the objects using two different behaviors, squeeze and press, the robot was able to detect the functionally important locations of the objects that afford actuation. The results make a strong case that for robots to operate tools autonomously, they need to be able to explore the tools and, in addition, use multi-modal 3D object representations instead of ones based on vision alone.

II. Related Work

Experiments in psychology have demonstrated that both tactile and auditory feedback are important sources of information for establishing multi-modal object representations during object exploration [2], [3]. One way in which humans leverage information from different sensory modalities is through the use of what psychologists call *exploratory behaviors* (see [4]) or *exploratory procedures* (see [5]). In "Play and Exploration in Children and Animals", Power writes:

Jivko Sinapov is currently at the Developmental Robotics Lab, Iowa State University, Ames, IA.

This work was supported in part by the DARPA ARM program under contract W91CRB-10-C-0126. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Distribution Statement 'A': Approved for Public Release, Distribution Unlimited.

"[. . .] exploratory behavior in infancy and childhood appears to serve an information-gathering function. Using a variety of methods, researchers have demonstrated that during exploration infants and young children extract at least short-term information about the characteristic of objects, including information about texture, hardness, weight, shape, size, and sound potential." [4]

Infants' use of exploratory behaviors when learning about objects is tightly connected to their ability to detect sensory events that occur over the course of object manipulation. Gibson [6] concludes that our basic knowledge about how objects behave in the natural world is gathered through constant observation of how objects are affected by our own actions. In other words, when exploring an object, both infants and adults observe perceptual outcomes (e.g., sounds, tactile sensations, and movement patterns) that are subsequently used to form expectations about how an object behaves when a specific action is applied on in (see [6]).

In contrast, the vast majority of representations used in robotics today are solely vision based (see [7], [8], [9], [10], [11] for a representative sample of systems that use such approaches). While such 2D and 3D representations capture how an object looks like, they do not encode multi-modal information (e.g., how an object feels like or sounds like) that may be necessary for successful manipulation. In other words, a 3D model of a hand-held drill cannot on its own provide a robot with the functionally relevant location of the drill's trigger. Because of this limitation, when robots are tasked with manipulating objects (e.g., pressing a button), they're typically pre-programmed by the human user to apply the behavior at a hard-coded location.

To address these limitations, several lines of research have pursued methods and approaches that enable robots to utilize non-visual as well as multi-modal cues when learning about objects. For example, several experiments have demonstrated that robots can use auditory [12] [13], proprioceptive [14], as well as multi-modal sensory feedback [15] for object recognition. The drawbacks of those methods is that they fail to take the object's geometry into account and can only handle simple objects that have no degrees of freedom (e.g., a cup, a box, etc., but not a stapler, a drill with a button, etc.).

In addition to object recognition, experiments have also demonstrated that non-visual cues can be used to improve robot manipulation of everyday objects. For example, Jain *et al.* [16] describe an experiment in which a robot was able to characterize doors and drawers using proprioceptive feedback detected over the course of opening them. In addition, experiments by Sukhoy *et al.* [17] [18] show that auditory and proprioceptive feedback can be used by a robot to estimate the location of a button, which if pressed, produced a sound. The main limitations of that work, however, are that the robot's perception of the object was only in 2D and that the method was applied only on one type of object, a button, and using only one type of behavior, pressing.

Fig. 2. The three objects used in our experiments: a flashlight, a hand-held drill, and a stapler.

III. Experimental Methodology

A. Robot

The experiments reported in this paper were performed with the upper-torso humanoid robot shown in Fig. 1. The robot was equipped with one 7-DOF Barrett Whole Arm Manipulator with the three fingered Barrett Hand as its end effector. The hand has four 4×6 tactile arrays, one for each finger and one for the palm. The robot was also equipped with two Audio-Technica U853AW cardioid microphone mounted in the head, which were used to capture auditory feedback at the standard 16-bit/44.1 kHz resolution and rate over a single channel.

B. Objects

Three different objects, shown in Figure 2, were used in our experiments: a drill, a flashlight, and a stapler. The objects were part of the DARPA ARM-S manipulation challenge, in which the robot was tasked with autonomous drilling, turning on the flashlight, and stapling a piece of paper. In addition to the physical objects, the DARPA ARM-S program provided high resolution 3D models which we used in our experiments.

C. Behaviors

The robot explored objects using two different behaviors, *squeeze* and *press*. Both behaviors were designed such that they produce both tactile as well as auditory feedback when applied on objects. Each behavior had two stages. In the *tactile localization* stage, the robot performed an action designed to capture tactile feedback that can be used to build a tactile point cloud of the object. In the second, *tool actuation* stage, a subsequent action was performed during which auditory feedback was used to detect whether the tool was actuated or not. The behaviors are described in more details below.

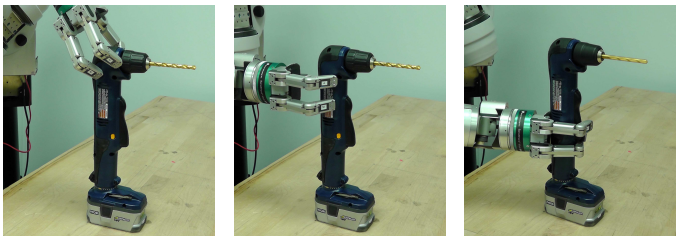
1) Squeeze: The squeeze behavior was applied on the flashlight and the drill objects and consisted of the following steps. First, the end-effector was positioned near the object so that closing the fingers of the hand would result in the object being grasped or touched. Next, the three fingers were closed using direct torque control. During the execution of the closing action, the tactile feedback from the fingers and the palm was monitored and once a tactile sensor was triggered (see Section III.D), the corresponding finger stopped closing. This constituted the *tactile localization* stage of the behavior. Finally, during the *tool actuation* stage, the robot attempted a squeezing action with the third finger (F3), during which audio feedback was monitored (see Section III.D) in order to detect whether the squeezing action actuated the tool. Figure 3.a) shows three example executions of the behavior on the drill at different locations on the object.

2) Press: The press behavior, shown in Figure 3.b), was applied on the stapler and consisted of the following steps: first, the hand was positioned above the stapler such that the hand's palm is horizontal relative to the table plane. Next, during the *tactile localization* stage of the behavior, the hand was lowered using the robot's Cartesian controller such that the orientation of the palm remained constant. Once tactile feedback was detected on the palm, the robot executed a pressing action using a Cartesian torque controller for a period of 10 seconds, which constituted the *tool actuation* stage of the *press* behavior. During this execution, the auditory feedback was monitored so that successful actuation of the tool can be detected. The next subsection describes the method used to detect auditory and tactile events from the robot's audio and tactile sensory streams.

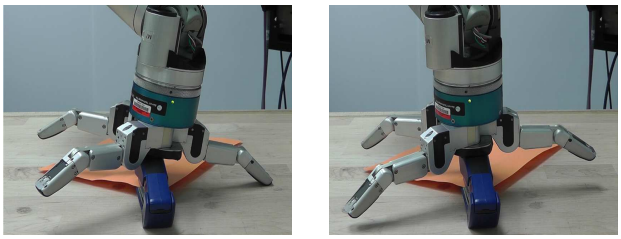
D. Event Extraction

The general approach for detecting sensory events used in this paper consisted of two steps: 1) for each sensory channel, learn a background model that encodes the expected sensor readings when no action is performed (i.e., no contact for the tactile sensor and no tool actuation for the audio sensor), and 2) during behavior execution, use the background model to detect events whose sensor signatures deviate from what is expected if no contact or no tool actuation is present. Following, the application of this approach is described in detail for both the tactile and auditory sensory channels.

1) Auditory Event Detection: Auditory events corresponding to tool actuation were extracted using the Discrete Fourier Transform (DFT) computed over the waveform captured by the robot's left microphone. The DFT was computed using 256 frequency bins with a window of 26.6 milliseconds computed every 13.3 milliseconds, and thus, each auditory sample can be represented by $x_t \in \mathbf{R}^{256}$. The background model was learned by recording a set of DFT samples $\mathcal{X}_{bg} = \{x_1, x_2, \ldots\}$ over a period of 0.5 seconds right before the *tool actuation* stage of each behavior was performed. From the set of samples $\mathcal{X}_{bg}$, for each of the 256 frequency bins, a tuple of the form (μ_i, σ_i) was computed so that it encodes the mean and the standard deviation for the i^{th} bin computed over the 0.5 seconds used to learn the background model. Therefore, the background model corresponds to the set $\{(\mu_i, \sigma_i)\}_{i=1}^{i=256}$ and encodes the expected values and their expected variance for each of the 256 frequency bins.

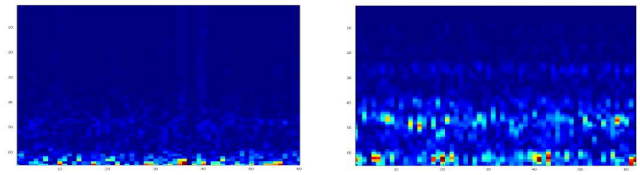


a) *Squeeze* behavior applied on the drill



b) *Press* behavior applied on the stapler

Fig. 3. Examples of the exploratory behaviors performed on the objects. a) *Squeeze*: The squeeze behavior, shown here performed on the hand-held drill, consisted of closing the robot's fingers and subsequently squeezing finger 3 in order to actuate the drill. In this case, successful actuation was achieved in the third example. b) *Press*: The press behavior, performed on the stapler, consisted of pressing down on the stapler at various locations. The stapler was successfully actuated in the second example.



a) Background noise b) Drill actuation

Fig. 4. a) Discrete Fourier Transforms (DFT) of the background noise in the lab; b) DFT of the sound produced when the drill is actuated.

During the execution of the *tool actuation* behavior, the background model was used to detect auditory events corresponding to successful actuation of the tool. Given a DFT sample $x_t \in \mathbf{R}^{256}$, the sample was classified as deviating from the background if k of the DFT bins have values deviating from the expected means by at least 2.5 standard deviations. If 5 consecutive samples were classified as deviating form the background, an auditory event was detected. The parameter k was set $\lceil 256/3 \rceil = 86$. Figure 4 shows sample waveforms captured by the robot's microphones and the detected tool actuation while manipulating the hand-held drill.

2) Tactile Event Detection: During the *tactile localization* stage of each behavior execution, the tactile activation values were used to extract events denoting instances in which the robot's fingers or the palm touched the object. Since each tactile array contains $4 \times 6 = 24$ cells, the learned background model was represented by the set $\{(\mu_i, \sigma_i\}_{i=1}^{i=24}$,

where each tuple (μ_i, σ_i) represented the expected mean and standard deviation for cell i when no contact was present. To compensate for sensor drift, the estimates were re-computed each time before the *tactile localization* stage of the behavior was performed using samples collected for 1.0 second.

As before, a novel sample was classified as deviating from the background if $\lceil 24/3 \rceil = 8$ of the tactile channels deviated from their expected background values by at least 2.5 standard deviations. If 5 samples in a row were classified as deviating after executing the *localization* behavior, a tactile touch event was detected. In the case of the drill and the flashlight, this procedure was used used for all 3 finger tactile arrays and the palm array. In the case of the stapler, only the palm array was used since the behavior used to explore the stapler did not result in the fingers touching the object. The next subsection describes how the behaviors described so far, coupled with the event detection routine, were used to explore three different objects: a drill, a flashlight, and a stapler.

E. Object Exploration and Data Collection

For each object, data was collected as described in the following steps:

1) Position the arm in a random configuration around the object so that the appropriate behavior can be executed.
2) Perform the *tactile localization* stage of the behavior. For each activated tactile sensor, record the 3D position of the sensor by performing forward kinematics given the current arm and finger joint values.
3) Perform the *tool actuation* stage of the behavior. Record the 3D position of the tactile sensor in contact with the tool (finger 3 for the drill and flashlight and the palm for the stapler) along with a corresponding label denoting whether the tool was actualized or not.
4) Release the tool, and go back to Step 1.

The location and orientation of each tool were kept constant during the data collection process. Forward kinematics was used to compute the 3D positions of tactile events. The procedure resulted in the collection of several point clouds:

- A point cloud corresponding to all tactile events detected by the robot's tactile sensors, $\mathcal{T} \in \mathbf{R}^{3 \times n_t}$.
- A point cloud corresponding to all auditory events that indicate successful actuation, $\mathcal{A}_+ \in \mathbf{R}^{3 \times n_{a_+}}$.
- A point cloud corresponding to all instances in which an auditory event was not observed when performing the second stage of each behavior, i.e., $\mathcal{A}_- \in \mathbf{R}^{3 \times n_{a_-}}$.

IV. Multi-Modal Point Cloud Registration

A. Problem Formulation

The overall task of the robot is to detect the functional features of each object and map them onto the object's 3D model. More specifically, given an object O, let the point cloud $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ denote the object's 3D model, where each point $p_i \in \mathbf{R}^3$. Given the 3D model point cloud $\mathcal{P}$, the tactile point cloud $\mathcal{T}$, and the auditory point clouds, $\mathcal{A}_+$ and $\mathcal{A}_-$, the task of the robot is to estimate a density function $\mathcal{F}$ over the points in the set $\mathcal{P}$, such that for each point $p_i \in \mathcal{P}$, $\mathcal{F}(p_i) \in [0, 1)$ encodes the probability of successful manipulation when the *tool actuation* behavior is applied at point p_i.

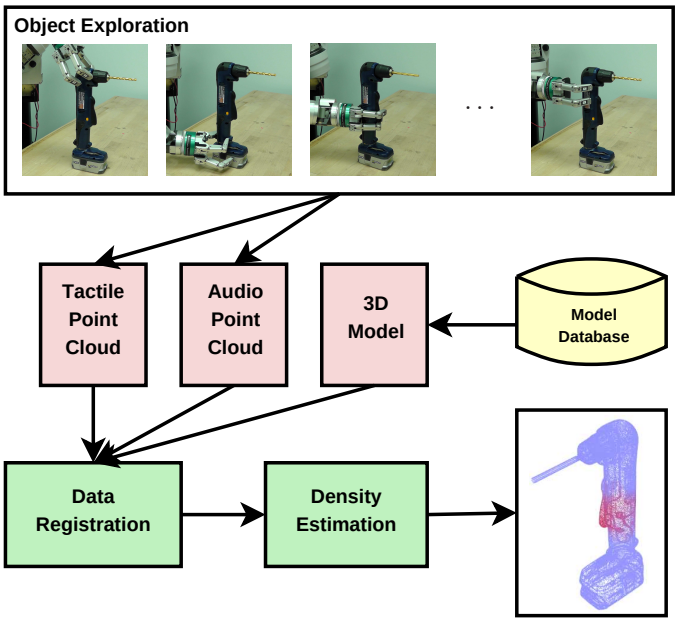


Fig. 5. An overview of the method used for audio-tactile point cloud registration and annotation. The method consists of three stages: 1) object exploration, during which the robot explores the object and records tactile and auditory point clouds encoding the locations at which the object was touched and actuated; 2) data registration, during which the recorded audio and tactile point clouds are transformed into the reference frame of the object's 3D model point cloud; and 3) density estimation, during which the transformed audio point cloud is used to fit a density function over the 3D model that encodes the likelihood of successful tool actuation.

Figure 5 shows the overall approach used to solve the problem, which uses three main stages:

- *Object exploration:* During the first stage, the robot explores each tool using a *tactile localization* and *tool actuation* behaviors performed at different locations on the object (see Section III.E).
- *Data registration:* Once the robot has extracted tactile and auditory point clouds, 3D data registration methods are applied to transform those point clouds into the 3D object model's reference frame.
- *Density Estimation:* Finally, the auditory point clouds are used to fit a density function over the 3D model that encodes the likelihood of successful tool actuation for different locations of the object.

Following, the next two sub-sections describe in details the algorithms used during the *data registration* and the *density estimation* stages of the proposed method.

B. Audio-Tactile Point Cloud Registration

Let $\mathcal{T} \in \mathbf{R}^{3 \times n_t}$ be the set of points corresponding to the tactile point cloud of the object, $\mathcal{A}_+ \in \mathbf{R}^{3 \times n_{a_+}}$ be the point cloud corresponding to all auditory events that indicate successful actuation, and $\mathcal{A}_- \in \mathbf{R}^{3 \times n_{a_-}}$ be the locations in which

the actuating behavior did not produce successful actuation. Finally, let $\mathcal{P} \in \mathbf{R}^{3\times n}$ be a point cloud corresponding to the object's 3D model.

The goal of this stage is to estimate a transformation matrix, $\mathbf{T} \in \mathbb{R}^{4\times4}$, that encodes the rotation and translation requires to transform the point clouds $\mathcal{T}$, $\mathcal{A}_+$, and $\mathcal{A}_-$ into the same reference frame as that of the 3D model point cloud, $\mathcal{P}$. To do that, two different registration algorithms, SAmple Consensus Initial Alignment (SAC-IA) [19] and the Iterative Closest Point (ICP) algorithm [20] were applied using the following steps:

1) Let $\mathcal{C} = \mathcal{T} \cup \mathcal{A}_+ \cup \mathcal{A}_-$, i.e., $\mathcal{C} \in \mathbf{R}^{3\times(n_t+n_{a_+}+n_{a_-})}$ is the union of the tactile and the two auditory point clouds.
2) Let $\mathbf{T}_{\text{sac-ia}} = \text{SAC-IA}(\mathcal{C}, \mathcal{P})$ where $\mathbf{T}_{\text{sac-ia}} \in \mathbb{R}^{4\times4}$ is the transformation matrix obtained after aligning cloud $\mathcal{C}$ to cloud $\mathcal{P}$ using the SAC-IA algorithm.
3) Let $\mathcal{C}' = \text{transform}(\mathcal{C}, \mathbf{T}_{\text{sac-ia}})$, i.e., $\mathcal{C}'$ is the result of transforming $\mathcal{C}$ according to the matrix $\mathbf{T}_{\text{sac-ia}}$.
4) Let $\mathbf{T}_{\text{icp}} = \text{ICP}(\mathcal{C}', \mathcal{P})$, i.e., $\mathbf{T}_{\text{icp}} \in \mathbb{R}^{4\times4}$ is the transformation matrix obtained after aligning $\mathcal{C}'$ to $\mathcal{P}$ using the ICP registration algorithm.
5) Let $\mathbf{T} = \mathbf{T}_{\text{icp}}\mathbf{T}_{\text{sac-ia}}$, i.e., $\mathbf{T}$ is the transformation matrix obtained by first applying transformation $\mathbf{T}_{\text{sac-ia}}$, followed by $\mathbf{T}_{\text{icp}}$.
6) Finally, let the clouds $\mathcal{A}''_+ = \text{transform}(\mathcal{A}_+, \mathbf{T})$ and $\mathcal{A}''_- = \text{transform}(\mathcal{A}_-, \mathbf{T})$.

In summary, the extracted tactile and auditory clouds are used to compute a transformation from the robot's frame of reference to that of the object model. This was done by sequentially applying the SAC-IA and the ICP registration algorithms, as implemented in the Point Cloud Library [21]. The end result consists of the two point clouds, $\mathcal{A}''_+$ and $\mathcal{A}''_-$ corresponding to the locations of successful and unsuccessful applications of the tool actuation behavior in the same reference frame as that of the 3D object model point cloud, $\mathcal{P}$. The next subsection describes how the transformed auditory point clouds were used to fit a density onto the object model that estimates the likelihood of successful actuation at different object locations.

C. Density Estimation

The last stage of the proposed method consists of fitting a density function over the 3D model, $\mathcal{P}$, that encodes the likelihood of successful tool actuation for different locations of the object. Let $p_i \in \mathbb{R}^3$ be the i^{th} point in the cloud $\mathcal{P}$. For each point p_i, the likelihood of successful manipulation $\mathcal{F}(p_i)$ is estimated using the following procedure:

1) Compute the point set $\mathcal{N}^+_{p_i}$ such that it contains all points from $\mathcal{A}''_+$ that are within d centimeters of the point p_i.
2) Compute the point set $\mathcal{N}^-_{p_i}$ such that it contains all points from $\mathcal{A}''_-$ that are within d centimeters of the point p_i.
3) Compute $\mathcal{F}(p_i)$ as:

$$\mathcal{F}(p_i) = \frac{|\mathcal{N}^+_{p_i}|}{|\mathcal{N}^+_{p_i}| + |\mathcal{N}^-_{p_i}|}$$

In summary, each point on the 3D model point cloud is annotated with the estimated probability that performing the tool actuation behavior in the neighborhood of the point will result in successful tool actuation. In our experiments, the value for the parameter d was set to 6.0 cm. The next section describes the results of applying the data registration and density estimation procedure on data gathered by exploring a hand-held drill, a flashlight and a stapler.

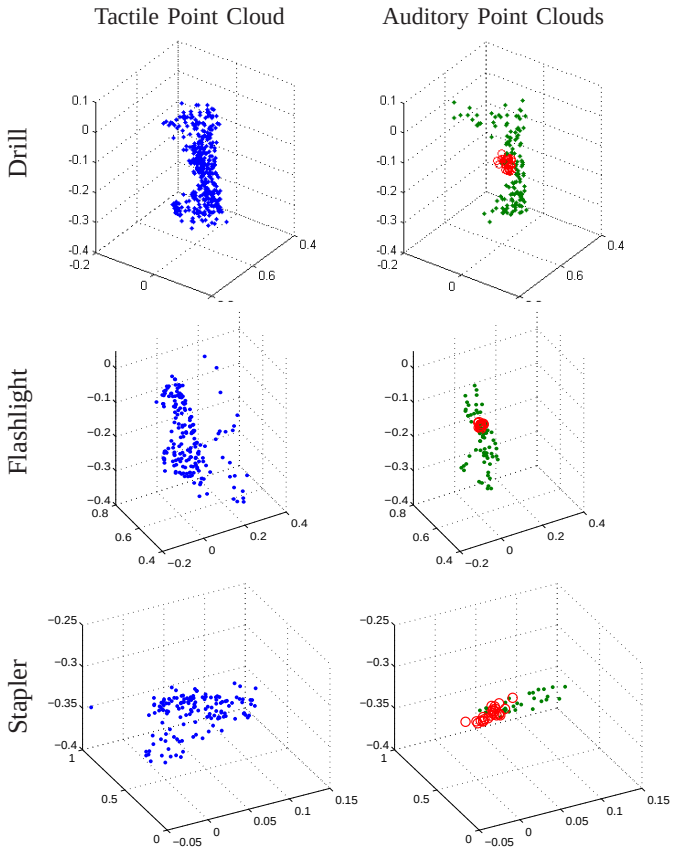


Fig. 6. *Left*: The tactile cloud, $\mathcal{T} \in \mathbf{R}^{3\times n_t}$, collected over the course of repeatedly applying the *squeeze* behavior at different locations on the object; *Right*: The auditory clouds, $\mathcal{A}_+$ (in green) and $\mathcal{A}_-$ (shown in red) recorded as a result of applying the *tool actuation* stage of the behavior used to manipulate the drill.

V. Results

Figure 6 shows the collected tactile cloud, $\mathcal{T}$, and the auditory clouds, $\mathcal{A}_+$ and $\mathcal{A}_-$, for the three objects explored by the robot. As expected, the collected data contains a lot of noise which further complicates the task of mapping the functional components of the objects onto their corresponding 3D models. Some of that noise is due to errors in the forward kinematics estimate of the fingertips' positions. In addition, as can be seen in the tactile cloud for the flashlight object, some of the registered tactile events were not caused by contact with the object, but are instead a result of noisy sensor readings.

Figure 7 shows the resulting annotated point clouds after applying the data registration and density estimation method described in Section IV. The results clearly show that the proposed method can detect the functionally important feature on each of the three objects, despite the presence of noise in

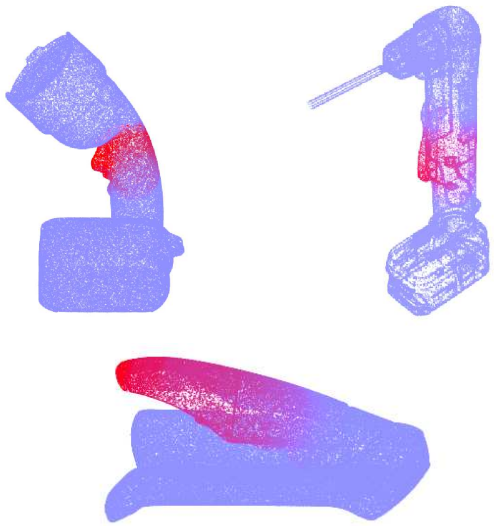


Fig. 7. The resulting annotated object model point clouds. The color of each point encodes the likelihood of successful actuation at different locations on the object, with red/purple indicating high probability of actuation, while light blue indicates low probability of actuation.

the data. As expected, for the drill, the density peaks at the location of the button. For the flashlight object, on the other hand, the density is high all around the top of the handle. This is because even the finger tip is not directly in contact with the button, the object was still often actuated by the inner link of the robot's finger. Finally, for the stapler, the magnitude of the estimated density increases gradually along the long axis of the top of the stapler.

VI. Conclusion and Future Work

Detecting the functionally important features of objects is a pre-requisite skill for autonomous tool use in unstructured environments. Towards solving this problem, this paper proposed a behavior-grounded method for audio-tactile registration and annotation of 3D point clouds. The experimental results showed that by applying exploratory behaviors on tools and observing auditory and tactile outcomes, the robot was able to annotate the object's 3D model with the probability that applying a behavior at a given location results in successful tool actuation. Unlike 3D representations that are based on visual sensors alone, the resulting object representation encoded aspects of the object's function as well as its shape.

A direct line for future work that may further advance the state of the art in autonomous tool use involves scaling up the proposed object representation to a much larger number of objects. Such a dataset would enable the use of data mining methods that can learn classifiers for annotating new 3D model point clouds that have not been explored by the robot. Scaling up to a large number of objects will also require that the robot explores the object in an intelligent, rather than random, manner, which can be achieved by using active learning methods for behavior selection. Overall, the results in this paper highlight the importance of integrating non-visual sensory percepts with 3D object representations, an approach that has the potential to greatly bridge the gap between human and robotic perception of objects.

References

[1] C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 20–29, 2007.

[2] S. Lederman and R. Klatzky, "Hand movements: A window into haptic object recognition," *Cognitive psychology*, vol. 19, no. 3, pp. 342–368, 1987.

[3] S. Lederman, "Chapter 4. the perception of texture by touch," in *Tactual perception: A sourcebook*, W. Schiff and E. Foulke, Eds. Cambridge Univ Press, 1982, pp. 130–168.

[4] T. Power, *Play and Exploration in Children and Animals*. Mahwah, NJ: Layrence Erlbaum Associates, Publishers, 2000.

[5] S. Lederman and R. Klatzky, "Haptic classification of common objects: knowledge-driven exploration," *Cognitive Psychology*, vol. 22, pp. 421–459, 1990.

[6] E. J. Gibson, "Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge," *Annual Review of Psychology*, vol. 39, pp. 1–41, 1988.

[7] M. Quigley, E. Berger, and A. Ng, "STAIR: Hardware and software architecture," *Presented at AAAI 2007 Robotics Workshop*, 2007.

[8] S. Srinivasa, C. Ferguson, D. Helfrich, D. Berenson, A. Collet, R. Diankov, G. Gallagher, G. Hollinger, J. Kuffner, and M. VandeWeghe, "Herb: A Home Exploring Robotic Butler," *Autonomous Robots*, vol. 28, no. 1, pp. 5–20, 2009.

[9] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927 – 941, 2008.

[10] K. H. B. Rasolzadeh, M. Bjorkman and D. Kragic, "An active vision system for detecting, xating and manipulating objects," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 133–154, 2010.

[11] Z. Marton, F. Seidel, F. Balint-Benczedi, and M. Beetz, "Ensembles of strong learners for multi-cue classification," *Pattern Recognition Letters*, 2012.

[12] J. Sinapov, M. Weimer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 2518–2524.

[13] A. Rebguns, D. Ford, and I. Fasel, "Infomax control for acoustic exploration of objects by a mobile robot," in *Lifelong Learning: Papers from the 2011 AAAI Workshop*, 2011.

[14] L. Natale, G. Metta, and G. Sandini, "Learning haptic representation of objects," in *Proceedings of the International Conference on Intelligent Manipulation and Grasping*, 2004.

[15] J. Sinapov, T. Bergquist, C. Schenck, U. Ohiri, S. Griffith, and A. Stoytchev, "Interactive object recognition using proprioceptive and auditory feedback," *The International Journal of Robotics Research*, vol. 30, no. 10, pp. 1250–1262, 2011.

[16] A. Jain, H. Nguyen, M. Rath, J. Okerman, and C. Kemp, "The complex structure of simple devices: A survey of trajectories and forces that open doors and drawers," in *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*. IEEE, 2010, pp. 184–190.

[17] V. Sukhoy, J. Sinapov, L. Wu, and A. Stoytchev, "Learning to press doorbell buttons," in *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*. IEEE, 2010, pp. 132–139.

[18] V. Sukhoy and A. Stoytchev, "Learning to detect the functional components of doorbell buttons using active exploration and multimodal correlation," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*. IEEE, 2010, pp. 572–579.

[19] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *The IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, 05/2009 2009.

[20] P. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[21] R. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–4.