TAPS: Tactile-Acoustic Perception for Vision-denied Robot Operation

Siyoon Sung[1], Chaerin Lee[2], Jiseok Jung[2], Yoonjoo Nam[2], Wonseok Choi[1], Sungjae Lee[1]

Abstract— **Robotic perception in vision-denied environments remains a fundamental challenge, as conventional vision-based sensors such as RGB cameras and LiDAR degrade severely under smoke, darkness, or occlusion. In this work, we present TAPS (Tactile–Acoustic Perception System), a vision-independent object recognition framework that relies solely on a robot's proprioceptive state and audio signals. TAPS actively explores an object placed in front of the robot for which no prior information is available at runtime, and estimates its spatial location and geometric properties through contact-based tactile exploration, without external sensors. The robot then generates deliberate impacts on the object and analyzes the resulting acoustic responses to infer material properties using a deep learning-based audio classifier. Finally, the estimated geometric and material information is fused to identify the object from a predefined candidate set. Experimental results in a tabletop setting demonstrate that TAPS can reliably recognize objects made of multiple materials under vision-denied conditions. Our code is available at `https://github.com/ubless607/TAPS`.**

I. INTRODUCTION

Robotic systems are increasingly deployed in hazardous and unstructured environments, such as those encountered in disaster response, deep-sea exploration, and planetary missions. In these settings, reliable perception becomes especially critical for successful task execution.

Most robotic systems typically rely on vision-based sensors, including RGB cameras and LiDAR, to perceive the environment. While these sensors perform well in structured and well-lit settings, their reliability degrades significantly under vision-denied conditions, such as smoke, occlusion, or complete darkness [1]–[3]. As a result, the robot may not perceive its environment, and this can lead to task failure.

To overcome these limitations, recent research in robotics has focused on alternative sensing modalities that do not rely solely on vision [4]–[6]. In particular, tactile and acoustic sensing have emerged as promising candidates, as they are based on physical interaction and remain effective even when visual information is unreliable [7]–[9].

Humans often infer an object's shape and material properties through *active touch*, also known as haptic exploration, which involves deliberate physical interactions such as tapping or pressing when visual information is unavailable [10]. Motivated by this behavior, we propose **TAPS (Tactile–Acoustic Perception System)**. This vision-independent object recognition framework integrates tactile interaction and acoustic sensing to probe object properties actively. Specifically, TAPS explores an object placed in front of the robot, for which no prior information is available at runtime, and estimates its spatial location and geometric properties using the robot's proprioceptive information without external sensors. It then applies an impact to the object and analyzes the impact sound to estimate the object's material. Our findings suggest that tactile-acoustic perception provides a viable alternative for object recognition in vision-denied environments.

[1]Graduate school of AI, POSTECH
[2]Department of Electrical Engineering, POSTECH

II. METHODOLOGY

A. Problem statement

This study addresses the problem of single-object recognition using a robotic manipulator in a vision-denied tabletop environment. The objective is to classify an unknown object within the workspace through physical interaction alone, without access to visual sensing.

The proposed recognition framework consists of three sequential stages. First, the robot actively explores the workspace using proprioceptive feedback from joint encoders to localize the object and estimate its size (Secs. B and C). Second, the robot infers the object's material by analyzing impact sound signals generated during physical contact, which are captured by a microphone mounted on the end-effector (Sec. D). Finally, the object is classified by combining the estimated size and material information from a predefined set of candidate objects (Sec. E).

Due to the limited motor control capabilities of the tested robotic arm (the Koch low-cost robot arm[1]), a strong assumption is made in this study: the object remains fixed at its initial location during interactions, i.e., it does not move when struck by the robot.

B. Horizontal Plane Localization

The robot's first task, without prior knowledge of the object is to determine its location within the workspace and estimate its planar size. To achieve this, we employ a bidirectional sweeping strategy based on physical interaction, as illustrated in fig. 1.

The exploration procedure begins by lifting the robot arm sufficiently upward to avoid unintended collisions and rotating it toward the right boundary of the workspace (fig. 1a). It then descends parallel to the tabletop surface and moves slowly to the left (fig. 1b).

When the robot arm makes contact with the object during this horizontal motion, the motor's movement becomes constrained. As a result, a discrepancy arises between the commanded joint target and its actual physical state. In this work, such state discrepancies are leveraged to detect

[1]`https://github.com/AlexanderKoch-Koch/low_cost_robot`

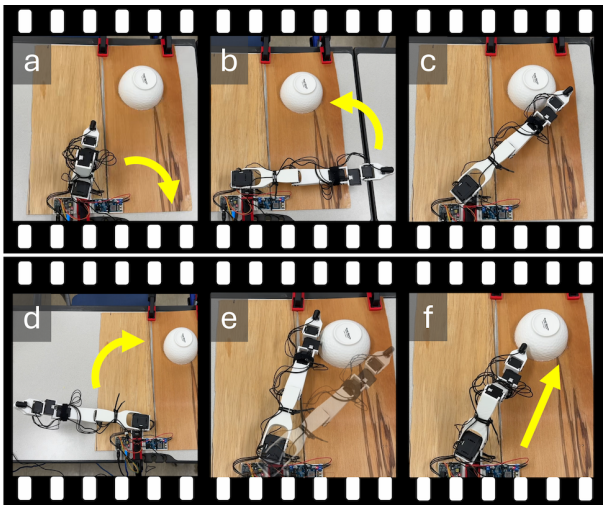


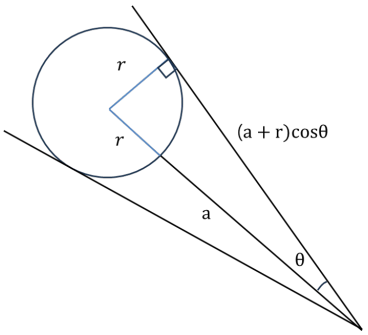
Fig. 1. Active contact-based localization and sizing (xy-plane)

physical contact, and the actual end-effector position in the physical system is recorded as a contact point (fig. 1c).

After completing the sweep in one direction, the robot arm is lifted again and moved to the left boundary of the workspace (fig. 1d). The same procedure is then repeated while sweeping toward the right to acquire a second contact point (fig. 1e). Through this bidirectional sweeping process, both contact points corresponding to the left and right boundaries of the object are obtained.

Based on the two recorded contact points, their midpoint is computed. The end-effector is then slowly moved along the corresponding direction toward the object to measure the distance a from the robot base to the object surface (fig. 1f).

To estimate the object's planar size, we model it as a circle. Under this assumption, as illustrated in fig. 2, the robot base, the object center, and a contact point form a right-angled triangle. Let θ denote the angle at the robot base of the right-angled triangle, and let a be the distance from the robot base to the object surface. The object radius r can then be derived from the corresponding trigonometric relationship.



$$\sin\theta = \frac{r}{a+r}$$

$$\therefore r = \frac{a\sin\theta}{1-\sin\theta}$$

Fig. 2. Geometric relationship between the robot base, the contact point, and the object center.

The object's planar shape is modeled as a circle for two reasons. First, a circle is fully defined by a single parameter (radius r), allowing the robot to estimate the object's approximate planar size from limited contact information without visual sensing. Second, the robotic manipulator used in our experiments has a limited reachable workspace, making complex multi-view exploration infeasible. Under these constraints, the circular model enables the robot to extract meaningful geometric information through simple bidirectional horizontal motion, making it an effective choice.

C. Height Measurement

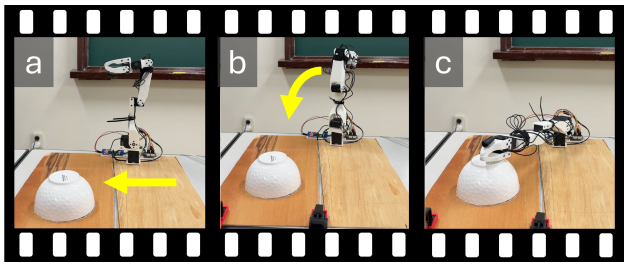


Fig. 3. Active contact-based localization and sizing (z-plane)

Upon completion of localization and radius estimation in the horizontal plane, the robot proceeds to measure the object's height h. First, the end-effector is moved to a position above the previously estimated object center (x, y) (fig. 3a). Subsequently, it slowly descends along the z-axis (fig. 3b). The descent speed is maintained sufficiently low to prevent potential damage to the robot or the object upon contact. As in the *horizontal plane localization* phase, physical contact with the object's top surface is detected by monitoring state discrepancies (fig. 3c). Once contact is detected, the robot stops the descent motion and records the angle of the second joint.

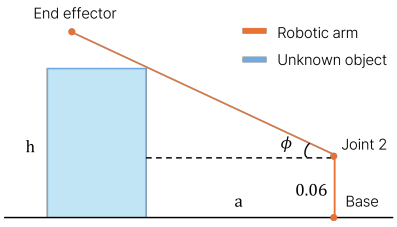


Fig. 4. Height estimation from contact geometry

As illustrated in fig. 4, the geometric relationship between the second joint and the contact point on the object's top surface forms a right-angled triangle. Let ϕ denote the angle of the second joint and let a denote the horizontal distance to the object estimated in the previous phase. The following trigonometric relationship holds:

$$\tan\phi = \frac{h - h_0}{a}, \tag{1}$$

where h_0 is a constant representing the reference height of the second joint from the tabletop (set to $0.06\,\mathrm{m}$ in this study). Rearranging this equation, the object's height h is derived as:

$$h = a\tan\phi + h_0. \tag{2}$$

Using this approach, the object's height can be robustly estimated using only the robot's proprioceptive information and simple geometric relationships, without the need for additional external sensors.

D. Material Classification with Audio

Subsequent to the height measurement phase, the robot performs an additional vertical interaction to infer the object material. While the height measurement phase applies a slow and cautious vertical descent to detect contact, the material classification phase executes a similar vertical motion with a higher force to generate a clear impact sound.

The collected audio signals are transformed into log-Mel spectrograms and used as inputs to the deep learning model for material classification. This representation encodes temporal information along the x-axis and spectral information along the y-axis in an image-like format.

For material classification, we adopt PANNs (Large-Scale Pretrained Audio Neural Networks) [11], a CNN-based architecture pretrained on over 5,000 hours of audio data, supporting sound event recognition across 527 classes. PANNs are designed to process image-like audio representations (e.g., log-Mel spectrograms), allowing the model to leverage the spatial inductive biases of CNN. In addition, compared to transformer-based architectures that typically require large amounts of training data, PANNs have a relatively simple structure, enabling effective fine-tuning even with limited dataset sizes. To this end, we collected a dataset of 500–600 impact sound samples spanning five material classes: Paper, Plastic, Ceramic, Wood, and Aluminum.

E. Object Classification

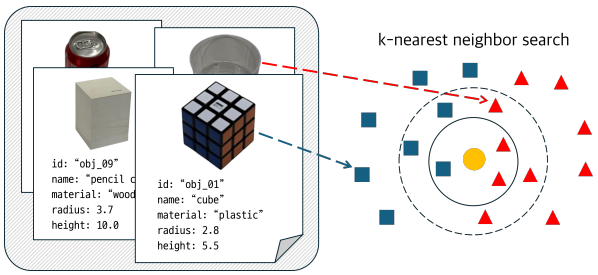


Fig. 5. Object classification via hierarchical k-nearest neighbor matching

In the final stage, object classification is performed by fusing multimodal sensory information obtained in the preceding stages, including estimated geometric parameters (radius r and height h) and material information.

For this purpose, the system references a pre-constructed object database. As illustrated in fig. 5 (left), each object o_i in the database is represented by an attribute set $\mathbf{f}_i$:

$$\mathbf{f}_i = \{\text{id}, \text{name}, \text{material}, \text{radius}, \text{height}\}. \tag{3}$$

Final object classification is performed using a k-nearest neighbor algorithm with a hierarchical search strategy to improve efficiency and reduce misclassification. First, the database is filtered based on material information, retaining only objects with matching material attributes. Next, geometric similarity is evaluated within this filtered candidate set by computing the Euclidean distance between the geometric feature vector of the query object, $\mathbf{x}_{\text{query}} = (r, h)$, and that of each database object, $\mathbf{x}_i = (r_i, h_i)$:

$$d(\mathbf{x}_{\text{query}}, \mathbf{x}_i) = \sqrt{\left(\ln \frac{r}{r_i}\right)^2 + \left(\ln \frac{h}{h_i}\right)^2}. \tag{4}$$

Finally, as illustrated in fig. 5, the Top-1 object with the smallest Euclidean distance, which corresponds to the highest geometric similarity, is selected as the final result.

III. EXPERIMENT

A. Evaluation Protocol

Experiments are conducted in a tabletop setup with a single object. For each object, five independent trials are performed, with the object's location randomly spawned at the start of each trial. In each trial, TAPS estimates the object's material category and geometric parameters, including radius and height. Material classification performance is reported as accuracy over five trials, while geometric estimation performance is evaluated using the root mean square error (RMSE), with normalized errors reported as percentages of the ground-truth values.

The initial object dataset consists of two objects for each of the five material classes. Due to time constraints, experiments are conducted on only two objects: a plastic cup and a Hot6 beverage can.

B. Results

1) Plastic Cup: The test object is a plastic cup with a ground-truth radius of 4.0 cm and height of 13.5 cm. TAPS correctly identified the material in 4 out of 5 trials (80% accuracy). Radius estimates had an RMSE of 0.80 cm (20%), while height estimates showed greater variation (RMSE = 3.88 cm, 28.7%). Table I summarizes the trial results.

TABLE I
PREDICTED GEOMETRY AND MATERIAL FOR THE PLASTIC CUP

Trial	Radius	Height	Material	Correct
1	3.2	9.5	Plastic	✓
2	3.3	9.8	Wood	×
3	3.3	9.8	Plastic	✓
4	3.1	9.5	Plastic	✓
5	3.1	9.5	Plastic	✓
Average	3.2	9.62	Plastic	Plastic cup

2) Hot6 Beverage Can: The aluminum can has a ground-truth radius of 2.5 cm and a height of 15.0 cm. Across all five trials, TAPS correctly identified the material as aluminum, achieving perfect classification accuracy (100%). The radius estimates showed a consistent tendency to be slightly larger than the true value, resulting in an RMSE of 0.84 cm (33.6%), while the height predictions remained generally close to the actual measurement, with an RMSE of 1.02 cm (6.8%). Table II presents the detailed predictions for each trial.

TABLE II

PREDICTED GEOMETRY AND MATERIAL FOR THE HOT6 CAN

Trial	Radius	Height	Material	Correct
1	2.6	15.9	Aluminum	✓
2	2.8	16.4	Aluminum	✓
3	2.7	15.9	Aluminum	✓
4	3.1	16.7	Aluminum	✓
5	3.1	18.7	Aluminum	✓
Average	2.86	16.7	Aluminum	Hot6 can

C. Results Using YCB-Impact Dataset

The YCB-impact dataset [9] is a publicly available benchmark for impact-sound-based material classification, collected by manually tapping everyday objects from the YCB object set using a handheld gripper. In our initial experiments, the PANNs model fine-tuned on this dataset achieved high classification accuracy on the test split, indicating strong performance under the dataset's original conditions.

However, when the same model was deployed in our experimental environment, its performance degraded substantially. This degradation can be attributed to the impact-sound generation mechanism. While the YCB-impact dataset was collected through human-operated tapping, impacts in TAPS are autonomously generated via motor-controlled robotic motions. As a result, motor actuation noise and joint vibrations are introduced into the recorded audio signals, leading to noticeable differences in amplitude and spectral characteristics compared to the human-collected data.

IV. CONCLUSION

We introduced **TAPS**, a tactile–acoustic perception framework for object recognition in vision-denied environments. By leveraging active physical interaction, TAPS enables a robotic manipulator to estimate object geometry through contact-based exploration and to infer material properties from impact-induced acoustic signals, without relying on vision or external tactile sensors. Experiments show reliable recognition of objects of different sizes and materials, demonstrating that tactile–acoustic perception is a viable alternative for robust operation in visually challenging settings, such as search-and-rescue or planetary exploration.

A. Potential applications

The proposed TAPS can serve as an alternative sensing system when vision sensors become unreliable. We present two application scenarios that highlight its potential as a complementary sensing modality.

Search-and-Rescue. Dense smoke and airborne debris in fire environments severely reduce visibility and contaminate camera lenses. Although thermal cameras and LiDAR can mitigate some of these limitations, they remain susceptible to scattering and attenuation from particulate matter. TAPS can complement these sensors for object identification in low-visibility conditions.

Planetary Exploration. Exploration rovers on distant planets may experience extended periods of darkness during which camera-based perception becomes infeasible. While artificial lighting could address this challenge, energy constraints make continuous illumination impractical. TAPS enables light-independent sensing, supporting nocturnal exploration and object recognition without additional lighting.

B. Limitations and Future Direction

*1) **Assumption of a Single Stationary Object**:* The system operates under the assumption that a single object exists in the workspace and remains stationary during the exploration process. This assumption is a deliberate design choice that enables contact detection and geometric estimation using only the robot's proprioceptive information, without relying on external sensors. Consequently, scenarios in which the object moves or rotates upon contact are not considered.

*2) **Simplified Geometric Modeling**:* The geometric properties of the object are modeled as a circle in the horizontal plane and a cylinder in the vertical direction. This simplification enables stable size estimation from limited contact information; however, it introduces constraints when dealing with objects that have asymmetric structures or complex geometries. While this design choice reduces computational complexity, it limits the range of object shapes for which the approach is effective and makes it unsuitable for applications requiring fine-grained geometric measurements.

*3) **Limitations of Acoustic-Based Material Recognition**:* Material recognition relies on acoustic signals generated by deliberate impacts. As a result, the extracted acoustic characteristics may vary due to environmental noise, subtle variations in impact force, and differences in contact location. Furthermore, the current approach assumes that the object exhibits sufficiently stiff mechanical properties to produce distinctive impact sounds. This assumption limits performance when applied to soft or deformable objects, whose impact responses are attenuated or highly variable. Although experiments in this study were conducted in a controlled environment, additional validation is required to assess robustness in real-world settings with higher noise levels, diverse materials, and varying object compliance.

*4) **Closed-Set Object Recognition Assumption**:* Object recognition is performed based on a predefined database of candidate objects. Consequently, recognition performance is limited for novel objects not included in the database, and the proposed system does not directly address the open-set object recognition problem.

*5) **Future Research Directions**:* Future work will focus on integrating tactile sensors or force/torque sensing to improve contact stability and extend the framework to non-stationary objects and multi-object environments. In addition, adopting more general geometric representations will enable improved recognition of objects with complex shapes, while developing material perception methods that are robust to environmental variations and object compliance will further expand the applicability of TAPS in real-world scenarios.

REFERENCES

[1] J. Park, H. Lee, I. Kang, and H. Shim, "No thing, nothing: Highlighting safety-critical classes for robust LiDAR semantic segmentation in adverse weather," in *CVPR*, pp. 6690–6699, 2025.

[2] M. R. U. Saputra, P. P. B. de Gusmao, C. X. Lu, Y. Almalioglu, S. Rosa, C. Chen, J. Wahlström, W. Wang, A. Markham, and N. Trigoni, "DeepTIO: A deep thermal-inertial odometry with visual hallucination," *IEEE RA-L*, pp. 1672–1679, 2020.

[3] S. Son, W. Lee, H. Jung, J. Lee, C. Kim, H. Lee, H. Park, H. Lee, J. Jang, S. Cho, *et al.*, "Evaluation of camera recognition performance under blockage using virtual test drive toolchain," *Sensors*, p. 8027, 2023.

[4] L. Macesanu, B. Folefack, S. Singh, R. Ray, B. Abbatematteo, and R. Martín-Martín, "CAVER: Curious audiovisual exploring robot," *arXiv:2511.07619*, 2025.

[5] L. Mack, F. Grüninger, B. A. Richardson, R. Lendway, K. J. Kuchenbecker, and J. Stueckler, "Visuo-tactile object pose estimation for a multi-finger robot hand with low-resolution in-hand tactile sensing," *arXiv:2503.19893*, 2025.

[6] Z. Liu, C. Chi, E. Cousineau, N. Kuppuswamy, B. Burchfiel, and S. Song, "ManiWav: Learning robot manipulation from in-the-wild audio-visual data," *arXiv:2406.19464*, 2024.

[7] J. Liu and B. Chen, "Sonicsense: Object perception from in-hand acoustic vibration," *arXiv:2406.17932*, 2024.

[8] M. Lee, U. Yoo, J. Oh, J. Ichnowski, G. Kantor, and O. Kroemer, "SonicBoom: Contact localization using array of microphones," *IEEE RA-L*, 2025.

[9] M. Dimiccoli, S. Patni, M. Hoffmann, and F. Moreno-Noguer, "Recognizing object surface material from impact sounds for robot manipulation," in *IROS*, pp. 9280–9287, 2022.

[10] S. J. Lederman and R. L. Klatzky, "Hand movements: A window into haptic object recognition," *Cognitive Psychology*, pp. 342–368, 1987.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2880–2894, 2020.