Recognizing object surface material from impact sounds for robot manipulation

Mariella Dimiccoli[1], Shubhan Patni[2], Matej Hoffmann[2], Francesc Moreno-Noguer[1]

***Abstract*— We investigated the use of impact sounds generated during exploratory behaviors in a robotic manipulation setup as cues for predicting object surface material and for recognizing individual objects. We collected and make available the YCB-impact sounds dataset which includes over 3,000 impact sounds for the YCB set of everyday objects lying on a table. Impact sounds were generated in three modes: (i) human holding a gripper and hitting, scratching, or dropping the object; (ii) gripper attached to a teleoperated robot hitting the object from the top; (iii) autonomously operated robot hitting the objects from the side with two different speeds. A convolutional neural network is trained from scratch to recognize the object material (steel, aluminium, hard plastic, soft plastic, other plastic, ceramic, wood, paper/cardboard, foam, glass, rubber) from a single impact sound. On the manually collected dataset with more variability in the speed of the action, nearly 60% accuracy for the test set (not presented objects) was achieved. On a robot setup and a stereotypical poking action from top, accuracy of 85% was achieved. This performance drops to 79% if multiple exploratory actions are combined. Individual objects from the set of 75 objects can be recognized with a 79% accuracy. This work demonstrates promising results regarding the possibility of using impact sound for recognition in tasks like single-stream recycling where objects have to be sorted based on their material composition.**

I. Introduction

Despite rapid progress in visual-based object recognition, physical object properties like their material composition are challenging to extract through distal sensing. In material recognition, image-based approaches have showed some intrinsic limitations due to the diversity in material appearances [1]. Haptic or tactile exploration can be also employed for material recognition (see [2] for a review). The sensory modality that has been overlooked so far but that bears great potential regarding material recognition is sound. Auditory response from impacts on the object surface reveals important characteristics about its material composition. Like touch and unlike vision, the sensory data induced by this interaction mode is independent of lighting conditions or object properties that are not relevant (like color). Contact-based object material recognition can be relevant in a number of application areas, like for example single-stream recycling [3], [4], where sound could aid recognition based on vision or touch.

Human experience of the world is inherently multimodal and sounds allow us to infer events in the world that are often not perceptible through vision [5]. Inspired by this, sound is finding its way into robotics in object-material segmentation [6], attribute shape prediction [7] and material recognition [8]. However, in these works sound is used only as a complementary modality for vision.

[1] Institut de Robòtica i Informàtica Industrial, CSIC-UPC in Barcelona
[2] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague.

Fig. 1: Experimental setup. Robot manipulator pokes objects with a closed gripper, generating impact sounds used for material classification.

In this paper, we propose to leverage impact sounds, that is the sounds an object emits as reaction to exploratory behavior, to learn object surface material. To this goal we introduce the first dataset of impact sounds with material annotations for the YCB objects and model set dataset, which includes over 3,000 impact sounds for 75 objects and propose a simple approach to learn a model from our dataset that can easily be transferred in a Robotic interactive learning pipeline (see Figure 1). Our contributions can be summarized as follows:

- We introduce and make publicly available the YCB-impact-sound dataset containing on average forty impact sounds and associated videos for each object in the YCB dataset.
- We propose a complete pipeline to predict material from impact sounds.
- We show that 75 individual objects can be recognized with a 79% accuracy.
- We demonstrated that the model learned with our dataset can be easily transferred to a robot to recognize material from impact sounds generated during exploratory behaviors.

The remainder of this paper is organized as follows: we first review related work in Section II, then we introduce the YCB-impact sounds dataset in Section III, and detail our approach and the experimental results in Section IV, respectively. We conclude the paper with final remarks in Section V.

II. Related work

A. Vision and Sound in Robot Perception

Sound is increasingly becoming an important modality for several tasks such as scene classification [9], object reconstruction [10], object tracking [11], robot perception [12]. In particular, impact sounds have proved to be useful for a variety of tasks including action recognition [13], object-material segmentation [6], self-supervised learning of visual models where sounds act as supervisory signals [14], attribute shape prediction [7], joint material and geometry classification [15], collision detection and localization [16], material recognition [8]. However, in these works, sound is typically used in conjunction with vision, serving a supplementary role.

B. Vision-based Material Estimation

Vision alone has been largely used for material recognition, even if the large variability in rich surface texture, geometry, sensitiveness to lighting conditions, and clutter make the problem particularly challenging. To deal with such extreme variability, the most successful approaches employ Convolutional Neural Networks (CNN) over larger and larger datasets [1], [17], [18]. However, state of the art methods still suffer from domain adaptation when the model is tested on a target dataset whose distribution differs from the source dataset [1].

C. Vision and Sound for 3D shape and Material Estimation

Since surface material provides useful hints for improving the performance of 3D scene understanding, recently there has been an increased interested in predicting material and shape jointly from videos and impact sounds. In [8], material is recognized by combining sound and visual geometry of the object, including global and local geometry around the contact point generating the sound. The model has been validated on a new dataset of synthetic objects encoded in a voxelized representation, where impact sounds were generated by modal sound synthesis. However, such information is difficult to acquire when testing on real objects.

D. Haptic Material Estimation

For material recognition by tactile sensing, Luo et al. [2] distinguish surface texture and object stiffness based recognition. The former typically involves sliding along the object surface and perceiving the vibrations using acoustic or tactile sensors. The latter, object stiffness based tactile material recognition, may involve tapping on the object, pressing it against a surface, or squeezing it between the gripper jaws. Tapping is mainly estimating the object's *hardness* (e.g. [19]), while pressing or squeezing probes its stiffness or material elasticity.

E. Vibration-based Material Estimation

Objects typically vibrate in a set of preferred modes that are closely related to their geometry and material properties, with high frequencies and small scales often imperceptible to humans. In the context of non-destructive testing [20], object material is often inferred through the measurement of its vibrations using contact sensors or laser vibrometers, which are very expensive. Recently, a method to predict material from videos by relying on visual vibrometry theory has been proposed in [21]. However, this approach only works with objects having fixed or known geometry.

F. X-ray based Material Estimation

X-ray diffraction (XRD) imaging has been used widely for material identification based on the intensity distribution of X-rays that directly interact with a material's molecular structure [22]. It produces high accuracy predictions transmission imaging but requires access to standard reference data that must meet homogeneity requirements, hence preventing use in an interactive setting. This limitation is also shared by X-ray Transmission Imaging technology, which is instead used to estimate material properties [23]. Recently, the combination of X-ray diffraction and X-ray Transmission Imaging has also been considered [24].

G. Laser-induced breakdown spectroscopy (LIBS)

LIBS is an atomic emission spectroscopic technique for analyzing the elemental composition of various solids, liquids and gases. It works by analyzing the spectral signature emitted by the optical emission from a high temperature plasma of the element under consideration, which is generated by removing a very tiny part from it by a high-power pulsed laser. However, handheld LIBS are very expensive and the deployment of LIBS in teleoperable and autonomous robotic platforms is still subject of investigation.

In this paper we propose a simple and portable approach to estimate material from impact sounds that does not require any previous knowledge about the object, nor the use of expensive technology.

H. Material datasets

The interest in material recognition from images led to the publication of several datasets in the past two decades. Since the first one, the CUReT dataset [25] published at the end of the nineties, the size of *material from images* datasets scaled up from a few thousands to millions of samples [1], [17], [18], [26], [27]. However, due to the intrinsic limitations of visual data for material recognition, more recent datasets combine visual information with sound [6]–[8], [14], [15] or are based on haptic features [28] such as force, temperature, and vibration.

To the best of our knowledge, our is the first dataset for material recognition from impact sounds that can be used as an independent source of information in a robotic manipulation setup.

III. YCB-impact sounds dataset

We used the YCB set of everyday objects [29] https://www.ycbbenchmarks.com/. Our dataset of impact sounds generated with these objects is available at https://osf.io/YCB-impact-sounds/. An illustration of the process of data generation is in the accompanying video (https://youtu.be/YC-impact-sounds-video).

Fig. 2: Capture setup used for our manually collected YCB-impact sounds dataset. A metallic gripper (hold by one hand) is used to interact with the object. Impact sounds generated by the interaction are captured by a shotgun microphone. A static camera is used to capture a close view of the object.

A. Manual dataset collection

We used 75 out of 77 objects of the YCB set (timer and rubik's cube were left out). The dataset was captured in a room with closed windows. We placed a shotgun microphone Rode Videomic Pro[1] connected to a digital audio recorder Zoom H1N[2] close to the source of sound and we used the PAL Robotics gripper made of anodised aluminium 7075-T6[3] to interact with the objects. We also placed a static GoPro camera Hero 4[4] on the side to capture a close top-down perspective of the interaction (48fps) similar to those that would have a camera placed on a robotic arm (see Figure 2). The same person captured all the dataset in the same place by performing three different actions to generate impact sounds: hitting, scratching and dropping. The hitting action was performed on average forty times by applying a different force at different locations for each object to capture richer data. The scratching and dropping actions were performed on average five times each per object at different locations and from different heights, respectively.

B. Dataset annotation

Surface material annotation was manually performed by using visual and tactile inspection. The mass and dimensions of the object provided by the authors of the YCB set [29] could not be used for determining the material since often the surface material differs from the internal material (e.g. drill). We distinguished eleven different surface materials: plastic, wood, ceramic, fiber, felt, foam, glass, paper, metal, rubber, leather. We also considered a more fine grained classification that distinguished two types of metals (steel, aluminum) and three types of plastics (soft plastic, hard plastic, other plastic). We also found 10 objects out of 75 to be composed of different materials (chips can, Master Chef can, Skillet lid, fork, spoon, knife, scissors, two screwdrivers and the hammer). For these objects, we labeled the impact sounds with the material of the object part that generated the sound. As can be observed in Fig. 3, the material distribution is very unbalanced. A visual illustration of YCB objects grouped by material is provided in Figure 4.

C. Dataset from gripper mounted on robot arm

For data collection using a robot arm, we used the Kinova Gen 3 manipulator with a Robotiq 2F-85 gripper – see Fig. 1. A subset of 49 YCB objects was used—the materials that do not produce sufficiently loud impact sounds (like felt, fiber, leather) were omitted as their response may be too weak compared to e.g. the sound generated by robot motors. The distribution for number of YCB objects explored per material is shown in Table I.

The recording hardware comprised of a Rode VideoMic Pro along with the Creative Extigy SoundBlaster[5] sound card. All audio was recorded at 44.1 kHz.

1) Teleoperated robot – impacts from the top: The position of the object with respect to the microphone is fixed. The Kinova robot pokes the objects from the vertical direction, and the audio is recorded. For each object, 50 such samples were collected for a total of 2190 impact sounds after removing a few instances.

2) Autonomous robot operation – impacts from the side: The robot approaches the object from the horizontal direction. Since there is a possibility that the object moves, the audio produced is different from the vertical poke and helps in preventing the classifier from over-fitting to one specific audio profile. For the horizontal poke action, two different speeds were used: 25mm=s and 14mm=s. We collected 50 samples per material for training and 10 samples per material for testing, for a total of 660 samples.

Material	# Objects
Steel	9
Aluminium	4
Hard Plastic	5
Other Plastic	6
Soft Plastic	5
Ceramic	5
Wood	3
Paper/Cardboard	5
Foam	3
Glass	2
Rubber	2

TABLE I: Material distribution of the 49 objects used in the experiments with a gripper mounted on a robot arm.

[1]https://www.rode.com/microphones/videomicpro

[2]https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h1n-handy-recorder/

[3]https://github.com/pal-robotics/pal_gripper

[4]https://gopro.com/es/es/update/hero4

[5]http://ixbtlabs.com/articles/extigy/