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.

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.

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

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 plastic (*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/

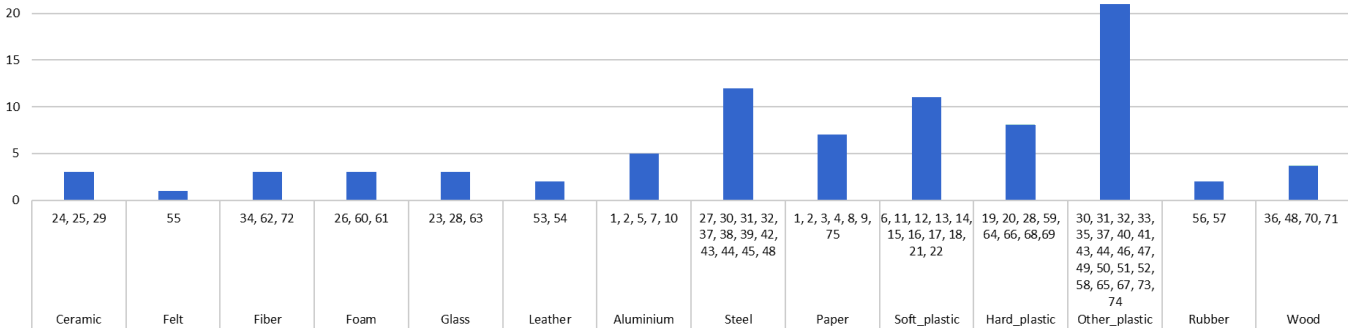


Fig. 3: Distribution of materials in the YCB dataset. Note that objects made of two materials appear twice. The IDs of the objects are those introduced in [29].

Fig. 4: Objects of the YCB dataset grouped by material.

D. Data post-processing

To detect the different impact sounds of the same object in the raw audio signal, we used a sliding window technique. We segmented uniformly the signal and for each segment, we extracted the highest absolute value and compared it with the highest absolute value in the previous segment. We assumed that an impact occurs when a large change in the highest absolute values between two temporally adjacent segments is produced. Knowing *a priori* the number of impact sounds present in the signal, the threshold could be easily tuned. The audio signal in each recording was denoised by relying on a classical denoising algorithm based on wavelet-threshold multitaper spectra [30]. The algorithm is also very useful to remove robot noise when the gripper is mounted on a robotic arm. An example of denoised signal corresponding to impact sounds generated on a paper object by a robotic arm can be seen in Figure 5.

After denoising, we computed a spectrogram for each detected impact sound from the denoised audio signal, with a FFT of size 400 by using the Torch audio library[6]. Since impact sounds are generally transient, we trimmed the waveform of audio samples over time to one second in length. We normalized the data to have a zero mean and standard deviation equal to one. Each spectrogram therefore

[6]https://pytorch.org/audio/stable/index.html

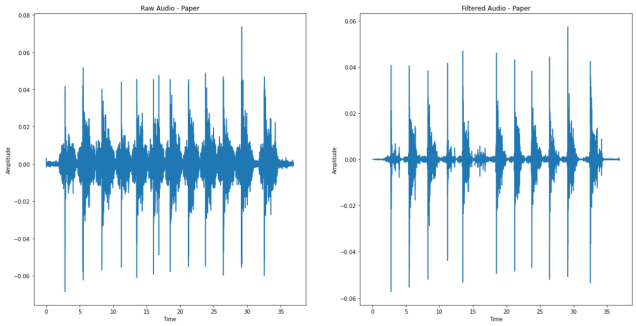


Fig. 5: Example of denoising, via wavelet-threshold multitaper spectra, of raw audio signal corresponding to an impact sound generated by a motor arm on a object of paper.

represents a concise 'snapshot' of an audio wave which is well suited to being input to CNN-based architectures. Examples of spectrograms corresponding to two different materials can be seen in Figure 6.

IV. Experiments and Results

A. Network architecture

We trained from scratch a modified version of the ResNet34 network [31] with a cross-entropy loss to predict object material among 11 classes. We found such network

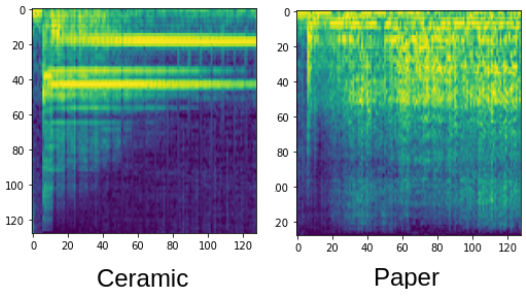


Fig. 6: Example of spectrograms corresponding to two different materials.

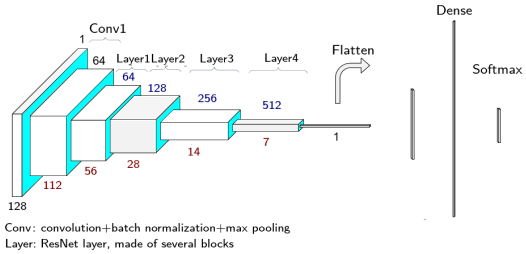


Fig. 7: ResNet34 used in the experiments. Upper blue numbers refer to the feature map dimension. Lower reddish numbers indicate the output size of the convolutional operator.

to have a capacity adapted to the complexity of the problem to solve. We used gray level spectrograms of dimension 128×128 as input. The first convolutional layer has 1 input channel and 64 output channels, kernel size 7, padding 3 and stride 2. The dimension of the last fully connected layer is 512 (see Figure 7).

B. Experiments on the manually generated dataset

Since the material class distribution in the YCB-model set is very unbalanced, we used data augmentation techniques [32] to create a more balanced training dataset. In the experiments reported in this section, we neglected objects of felt, leather and fiber to make results more directly comparable with those performed with data generated by a robot, where they were also discarded. We trained from scratch the network for 15 epochs using a learning rate of 0.001 which decreases by a factor gamma 0.1 every seven epochs, with momentum 0.9, and weight decay 0.02. When including impact sounds from the same object in training and testing, performance is very high and achieves over 90% accuracy. However, to ensure generalization capability to unknown objects, we split the dataset in train/validation and test set in a way such that the test set contains objects not presented during training. The list of unpresented objects and their material information is reported in Table II. Such split consists of 72% training, 15% validation and 13% testing. In this case, performance drops significantly but still achieved nearly 60% over 11 classes (9% chance level accuracy) for seventeen objects of the test set. In Figure 8, we show the

Material	Objects
Steel	padlock
Aluminium	master chef can, potted met can
Hard Plastic	pitcher base, chain, rope
Soft Plastic	banana, strawberry
Other Plastic	Golf ball, small marker
Ceramic	mug
Wood	Wood block
Paper/Cardboard	Gelatin box
Foam	Foam brick, washers
Glass	skillet lid
Rubber	racquetball

TABLE II: Independent testing set – not presented objects and their material.

confusion matrix obtained when testing on objects presented for the first time. As would be expected, materials that have a strongly attenuated response, foam in particular, are hard to identify. Paper and plastic have also lower recognition rates; for plastic, part of the drop in accuracy is due to misclassification within the subclasses of plastic (soft / hard / other).

Fig. 8: Confusion matrix – manually collected dataset (Section IV-B). Confusion matrix on test set of the YCB-impact sounds dataset.

C. Experiments on the robot-generated dataset

Experiments were carried out on the data collected from the robotic setup to gauge how effectively the network described in IV-B can transfer to similar impact sounds generated by a robotic setup. The following experiments were conducted:

1) Testing a randomly initialized network on audio samples generated by the robot.
2) Testing the network pre-trained in Section IV-B on audio samples generated by the robot:
   a) teleoperated vertical impact
   b) automated horizontal impact
   c) mixture of test samples from vertical and horizontal impact

3) Fine-tuning the network by further training it with a sparse subset of audio samples from the robot impact sound dataset. For every material, mixed audio samples of known objects collected from different exploratory behaviors were used for further training the network.
   a) teleoperated vertical impact
   b) mixture of test samples from vertical and horizontal impact

   The number of samples were chosen to be 50 samples per material, 100 samples per material, and the whole training set.
4) Retraining the network from randomized initial weights on only audio samples form the robot impact sound dataset.
   a) teleoperated vertical impact
   b) mixture of test samples from vertical and horizontal impact
   c) train only on sounds collected from vertical poking, and test only on sounds collected from horizontal poking

In all cases, the samples in the test set are only taken from objects not presented during training (see Table II). The results from three runs per experiment are averaged and shown in Table III. The confusion matrices for selected experiments are shown in Fig 9. For experiments 2) and 3), the hyperparameters for training were not changed from IV-B.

Experiment	Accuracy
1	26 %
2.a	42.01%
2.b	48.31%
2.c	45.92%
3.a (50 samples/material)	53.36%
3.b (50 samples/material)	55.41%
3.a (100 samples/material)	59.83%
3.b (100 samples/material)	62.14%
3.a (avg. 175 samples/material)	65.12%
3.b (avg. 175 samples/material)	67.68%
4.a	84.7%
4.b	78.9%
4.c	71.36%

TABLE III: Experimental results with data collected by a robotic arm.

We can see from the results that the network trained in Section IV-B achieves 42% accuracy versus 26% achieved using random weights. There is a slight difference between the prediction accuracy for vertical and horizontal impact sounds which is discussed later.

Fine-tuning the network with audio samples from the robot setup in the Experiment 3.x series can increase the accuracy from around 45% up to around 65% when more samples are added.

If the network is trained specifically on the data from the robot setup (Experiments 4.x), the accuracy increases up to 84.7% (Exp. 4.1) for the vertical impacts. This is a good recognition rate (recall that this is still on the test set – objects not presented to the network before). Moreover, the misclassifications are largely within the subcategories of plastic. The fact that training and testing was done using the same action—vertical pokes—may be realistic in an application scenario. When both robot actions are combined—top and side pokes—the performance drops to 78.9%.

Finally, we tested transfer from one robot action to the other in Experiment 4c. For vertical impacts, the motion of the objects against the direction of motion is restricted, but during horizontal impact, they are free to slide across the surface. The difference was investigated by retraining the proposed network from scratch on only vertical impact sounds and testing on only horizontal sounds. The accuracy of 71% is reasonable and may indicate also the accuracy that could be achieved when transferred to a different robot setup—different robot arm, gripper, table etc. For sounds from more rigid materials, an accuracy of 70% to 80% can be achieved. For materials that damp the impact sounds the accuracy drops, rubber being the most difficult material to recognize.

D. Recognizing individual objects from impact sounds

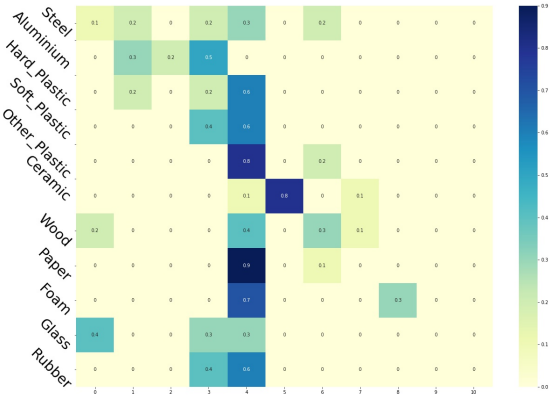
With the dataset of impact sounds generated manually, which include 75 YCB-objects, we trained the same ResNet34 from scratch to explore the potential of our impact sounds dataset for object recognition. We divided the dataset in training/validation/test in a way such that the test set contains only impact sounds of objects not presented during training. Our model achieves 79% accuracy at test time, which is significant considering that there are 75 different objects. This result is in line with previous work [33], which has shown that sounds generated by physical interaction with objects contain information indicative of the object. The confusion matrix we obtained is shown in Figure 10.

V. Conclusion and Future Work

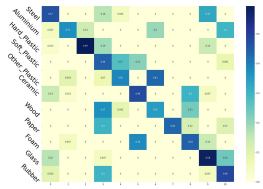
This paper proposed a learning framework and introduced an experimental study which shows how a robot can use impact sounds to recognize the types of surface material of objects not presented during training. To the best of our knowledge, this is the first experimental study which investigates and demonstrates the suitability of impact sounds to predict material in a robotic manipulation setting. In addition, we tested also recognition of individual objects.

The framework was evaluated using 49 objects from the YCB-object set made of 11 different materials and by using different types of behavioral interactions, namely poking from different directions under varying motion restrictions. Overall, our experiments demonstrate that impact sounds can be an important source of information to predict object material. Material information can subsequently be used in a perception-action learning loop alone or jointly with other modalities to adjust accordingly the manipulation force and grasp strategy, to estimate stiffness, or to recognize the object type.

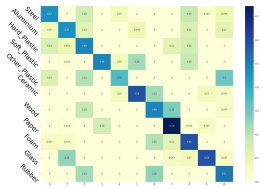
We achieved very good material recognition accuracy (85%) on the robot setup when the same exploratory action



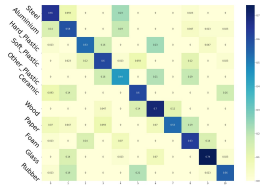
(a) Exp 1



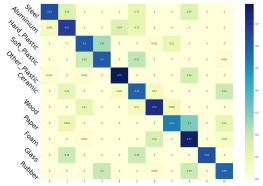
(b) Exp 2.b



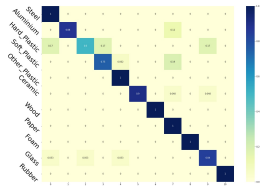
(c) Exp 3.b, 50 samples/material



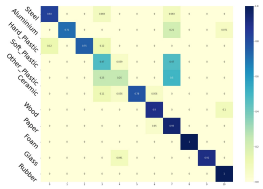
(d) Exp 3.b, 100 samples/material



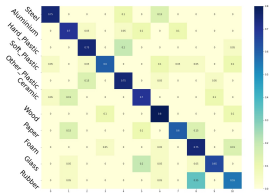
(e) Exp 3.b, entire training data



(f) Exp 4.a



(g) Exp 4.b



(h) Exp 4.c

Fig. 9: Confusion matrices – data collected by robot arm (Section IV-C).

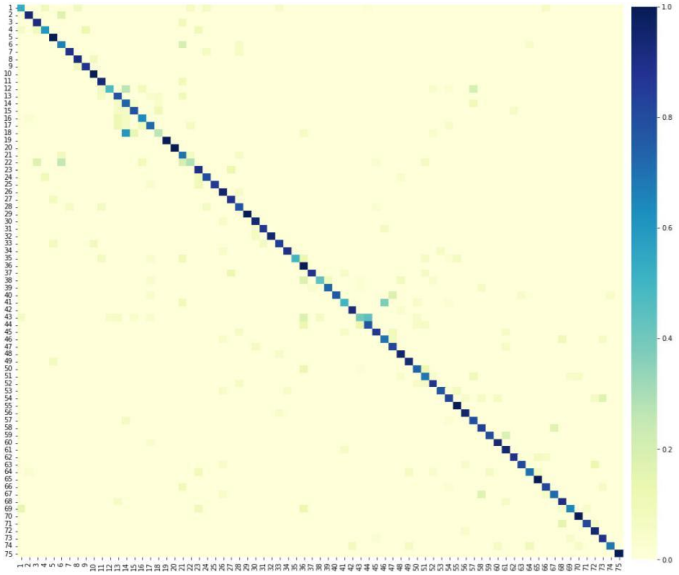


Fig. 10: Confusion matrix obtained when predicting objects from impact sounds not presented during training. For IDs of the objects, see [29].

(vertical poking) was used. This is acceptable in a setting when the actions is specifically designed for perception. We have also shown transfer between different robot actions and their parameters. Thus, material recognition from sound could also be employed in parallel to other tasks carried out by the robot (e.g. pick and place). As would be expected, materials that attenuate impact sounds like foam, rubber and paper were the hardest to classify.

We have made our dataset and trained model publicly available on the Open Science Foundation (OSF) platform (https://osf.io/YCB-impact-sounds). This dataset contains data from more material categories (14). In addition, additional actions (scratching and dropping the object – these were not used for classification in this work) are available, along with video recordings from the experiments. In the future, recognition using these additional actions or combined material estimation from audio and visual data can be performed.

VI. Acknowledgments

This work was supported by the project Interactive Perception-Action-Learning for Modelling Objects (IPALM) (H2020 – FET – ERA-NET Cofund – CHIST-ERA III / Technology Agency of the Czech Republic, EPSILON, no. TH05020001) and partially supported by the project MDM-2016-0656 funded by MCIN/ AEI /10.13039/501100011033. M.D. was supported by grant RYC-2017-22563 funded by MCIN/ AEI /10.13039/501100011033 and by "ESF Investing in your future". S.P. and M.H. were additionally supported by OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics". We thank Bedrich Himmel for assistance with sound setup, Antonio Miranda and Andrej Kruzliak for data

collection, and Lukas Rustler for video preparation.

REFERENCES

[1] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3479–3487.

[2] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.

[3] L. Chin, J. Lipton, M. C. Yuen, R. Kramer-Bottiglio, and D. Rus, "Automated recycling separation enabled by soft robotic material classification," in *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*. IEEE, 2019, pp. 102–107.

[4] D. Guo, H. Liu, B. Fang, F. Sun, and W. Yang, "Visual affordance guided tactile material recognition for waste recycling," *IEEE Transactions on Automation Science and Engineering*, 2021.

[5] C. A. Fowler, "Auditory perception is not special: We see the world, we feel the world, we hear the world," *The Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2910–2915, 1991.

[6] A. Arnab, M. Sapienza, S. Golodetz, J. Valentin, O. Miksik, S. Izadi, and P. Torr, "Joint object-material category segmentation from audio-visual cues," 2015.

[7] Z. Zhang, J. Wu, Q. Li, Z. Huang, J. Traer, J. H. McDermott, J. B. Tenenbaum, and W. T. Freeman, "Generative modeling of audible shapes for object perception," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1251–1260.

[8] F. Shi, J. Guo, H. Zhang, S. Yang, X. Wang, and Y. Guo, "Glavnet: Global-local audio-visual cues for fine-grained material recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 433–14 442.

[9] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, 2016.

[10] Z. Zhang, Q. Li, Z. Huang, J. Wu, J. Tenenbaum, and B. Freeman, "Shape and material from sound," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7053–7062.

[12] B. Spice, "Sounds of action: Using ears, not just eyes, improves robot perception," *ScienceDaily*, 2020.

[13] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellström, "Audio-visual classification and detection of human manipulation actions," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3045–3052.

[14] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2405–2413.

[15] A. Sterling, J. Wilson, S. Lowe, and M. C. Lin, "Isnn: Impact sound neural network for audio-visual object classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 555–572.

[16] X. Fan, D. Lee, Y. Chen, C. Prepscius, V. Isler, L. Jackel, H. S. Seung, and D. Lee, "Acoustic collision detection and localization for robot manipulators," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9529–9536.

[17] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?" *Journal of Vision*, vol. 9, no. 8, pp. 784–784, 2009.

[18] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Opensurfaces: A richly annotated catalog of surface appearance," *ACM Transactions on graphics (TOG)*, vol. 32, no. 4, pp. 1–17, 2013.

[19] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 951–958.

[20] K.-J. Langenberg, R. Marklein, and K. Mayer, *Ultrasonic nondestructive testing of materials: theoretical foundations*. CRC Press, 2012.

[21] A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, F. Durand, and W. T. Freeman, "Visual vibrometry: Estimating material properties from small motion in video," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 5335–5343.

[22] A. Narten and H. Levy, "Liquid water: Molecular correlation functions from x-ray diffraction," *The Journal of Chemical Physics*, vol. 55, no. 5, pp. 2263–2269, 1971.

[23] F. Pfeiffer, T. Weitkamp, O. Bunk, and C. David, "Phase retrieval and differential phase-contrast imaging with low-brilliance x-ray sources," *Nature physics*, vol. 2, no. 4, pp. 258–261, 2006.

[24] S. Stryker, J. A. Greenberg, S. J. McCall, and A. J. Kapadia, "X-ray fan beam coded aperture transmission and diffraction imaging for fast material analysis," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[25] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Transactions On Graphics (TOG)*, vol. 18, no. 1, pp. 1–34, 1999.

[26] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, "On the significance of real-world conditions for material classification," in *European conference on computer vision*. Springer, 2004, pp. 253–266.

[27] B. Caputo, E. Hayman, and P. Mallikarjuna, "Class-specific material categorisation," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1597–1604.

[28] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," in *Conference on Robot Learning*. PMLR, 2017, pp. 157–166.

[29] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 36–52, 2015.

[30] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE transactions on Speech and Audio processing*, vol. 12, no. 1, pp. 59–67, 2004.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

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