Contact Classification for Agriculture Manipulation in Cluttered Canopies

Moonyoung Lee, Kevin Zhang, George Kantor, Oliver Kroemer

Carnegie Mellon University
Robotics Institute
{moonyoul, klz1, gkantor, okroemer}@andrew.cmu.edu

Abstract

In this paper, the authors present a novel way to classify contact objects using audio signals in a highly cluttered canopy environment for agriculture manipulation. Rather than solely relying on visual data to represent the dense canopies as obstacles, we investigate whether robot can observe latent properties about safe interactions such as brushing against leaves using audio signals. We developed a hand-held device to facilitate the data collection process to distinguish between three classes: leaf, twig, trunk. Of the time domain, frequency domain, and cepstrum representations (MFCC), MFCC comparisons showed the most distinguishable patterns across the classes. The provided results present a promising direction to expand this research to leverage deep learning networks to consistently classify the extracted audio inputs that can lead to safe and robust agriculture manipulation.

Introduction

With increasing world population and rapid decrease in available arable land, there is an urgent need to enhance agriculture productivity. While there has been growing efforts to adopt robotics and AI-enabled automation in agriculture, many existing approaches focus predominantly in monitoring but not on harvesting. Given that fruit harvesting is a labor-intensive task that constitutes significant portion of the production costs (Whitney 1995), and there are growing labor shortages on the field (Sario 1993), introducing robots in this trend suggests to be a cost-effective solution. However, robot manipulation in a heavily cluttered environment where picking fruits in dense canopies can be an extremely challenging problem. Traditional robot manipulation approaches, commonly applied in warehouses and factories, often only use visual data to model the robot's surroundings as rigid body objects with which the robot must avoid colliding. This would not transfer well in cluttered canopies because such approach would be too restrictive to generate feasible paths.

Rather than relying only on visual data to model the environment, we propose to also utilize vibrotactile or audio signals obtained from contact microphones attached to the robot's end-effector to classify the obstacle the robot has come in contact with. Our key insight is that leafs, twigs, and trunks have different audio responses upon contact. As leaves are more permeable obstacles than twigs or trunks, the robot can use the audio contact classification to supplement its original trajectory. Combining both visual and audio signals provides intelligent approach for the robot to obtain latent information about the environment and ultimately to robust manipulation even in cluttered and occluded settings.

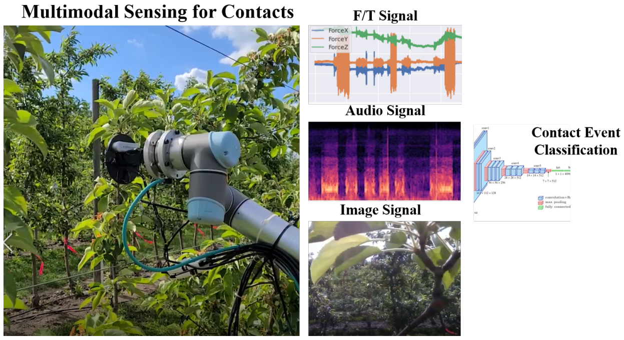

Figure 1: Robotic manipulation in cluttered agricultural environments. Multi modal sensing inputs with deep learning network for classifying type of collision events.

Related Works

There has been numerous studies on tackling various aspects of robotic fruit harvesting. Common areas of focus are robot perception for fruit counting (Yandun, Silwal, and Kantor 2020) or motion planning to pick fruits (Willigenburg, Hol, and Henten 2004), (Cao et al. 2019). Many of the manipulation planning approaches, however, drastically reduce the model complexity of the interacting environment by ignoring obstacles that are not fruits. The underlying assumption is that leaves and twigs don't drastically affect the robot trajectory. This assumption works for only limited cases of well-pruned canopies or where fruits are clearly exposed for picking but performs poorly in realistic field environments where excessive forces could damage the plants.

(Nemlekar et al. 2021) explicitly included leaves as permeable objects with low weights in their RRT-based path planner. However, their perception pipeline to distinguish permeable and non-permeable objects required time-consuming human intervention to label all the leaves from the generated point cloud.

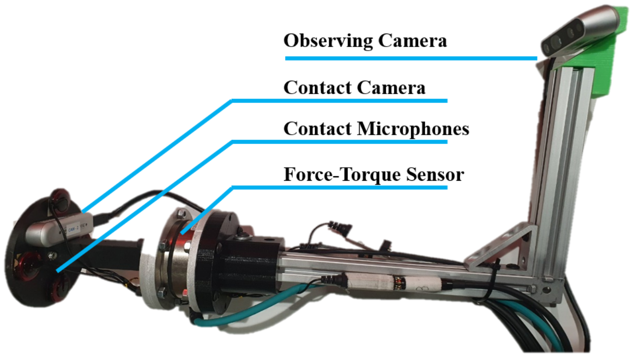

Figure 2: Multi modal sensor suite that can be hand-held in order to facilitate data collection process.

Instead of relying on visual data for classifying contact properties, audio signals can provide distinguishable features as shown in (Zhang et al. 2019), (Sawhney et al. 2020). These studies resemble this paper most closely in that contact microphones are utilized to distinguish between objects of interests with aid of a deep learning network. However, these works focus on significantly more structured data collected from cutting tasks rather than the cluttered structures of plant canopies. Lastly, as deep learning network performance benefits from a prolific dataset, the authors develop a hand-held device to facilitate the data collection process similar to studies in (Song et al. 2020).

Data Collection

Pilot data for contact interaction was captured at UMass Farm on University of Massachusetts Amherst Center for Agriculture with the UR5e robot on a mobile base. As safely collecting prolific data with the robot on the field is difficult, the authors developed a hand-held multi-modal sensor suite as shown in Fig. 2 to facilitate data collection process. The sensor suite has four piezo contact microphones evenly distributed on a contact plane of the device. In addition, the device captures RGB images from two Intel Realsense D435 cameras and Force/Torque readings from the FT24252 sensor.

With the hand-held device, we collected data for various leaf, twig, and trunk types across campus of Carnegie Mellon University. We used mainly pushing motions to move the device into the plants with some lateral motion.

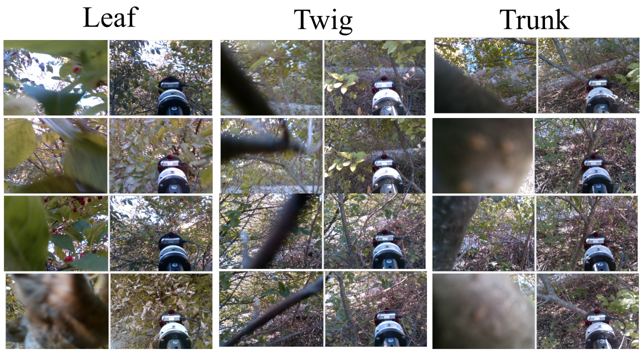

Figure 3: Sensor collection on various contacts

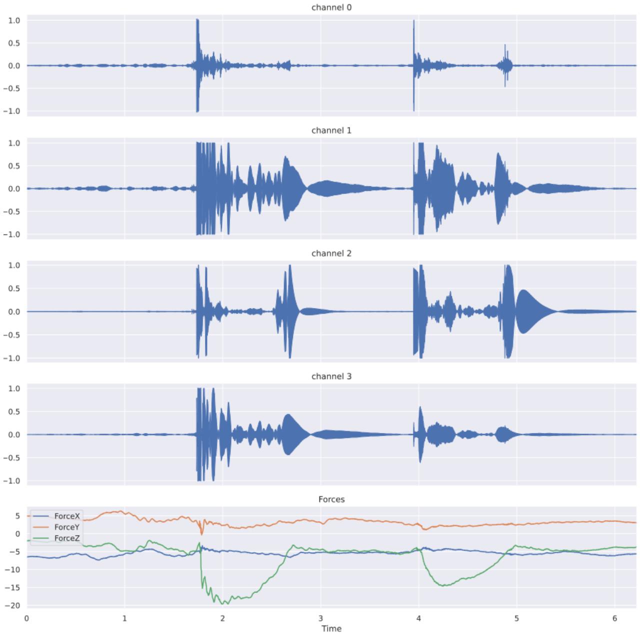

Figure 4: Audio signal response of contacting against trunk from each of the microphone (top,bot,left,right) and F/T sensor

As the purpose of this work is to classify various contact objects according to the audio input, we focus on time intervals of the data where distinct brushing or rustling sounds are made upon contact. As shown in Fig. 3, upon such contact time intervals, we observe that there are characteristic visual features distinct for each class. The aim of the following sections is to also find similar distinctive audio patterns for each class upon contact.

Data Processing

Audio signal from the microphone arrays in the time domain provides amplitude of contact sound over time. When plotted with the Force/Torque sensor data as shown in Fig. 4, we observe their responses are shared along the same time duration, indicating periods of contact interaction. The low-cost piezo microphones used are high impedance and weak signal that requires to be amplified through the UMC404HC audio interface with 24-bit/192kHz resolution converters. We manually tune the audio input gain to mitigate signal clipping while still detecting minute rustling signals. However, as both the audio amplitude only provide signal intensity over time, the data is limited and difficult to classify between various object classes. We therefore want to extract audio features in the frequency domain where vibration frequencies detected would be distinctive between rustling of light leaves as opposed to heavier scratching noises of thicker branches.

To analyze in the frequency domain, we first filter out dead-space in the time domain in order to best extract useful audio features after applying the short time fourier transform

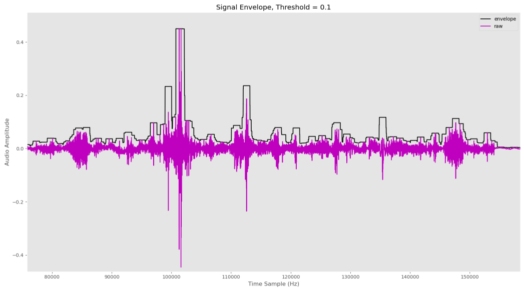

Figure 5: Envelope of the audio signal to filter out dead-space in the time-amplitude plot

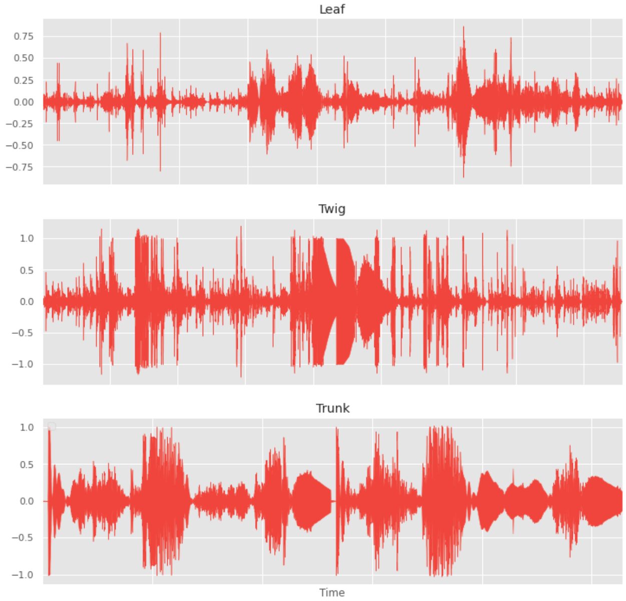

Figure 6: Filtered audio signal extracted for each of the classes

across extracted time intervals. To do so, a rolling window that averages the absolute mean of the audio signal is used to create a signal envelope as shown in Fig. 5. Once the signal envelope is created, we can apply an amplitude threshold to smoothly extract dominant audio signals without distorting the audio signal. With the dead-space filtered out, we can then compare the spectrograms of contact sounds among the various objects. As seen in Fig. 7, we can observe the nuance differences in the spectrograms as rustling leaves display lowered energy with continuous signals compared to the high energy with impulse-like signals for the striking twigs upon contact.

Results

The current nascent stage of the dataset is composed of three different classes (leaf, twig, trunk), three different type variations among each classes, four trials of various contact motions, and for four individual microphones, resulting in a total of approximately 2880 seconds of audio signals. Although the four audio channels could provide spatial information about the contact position, for this work, we focus only on one of the audio channel in order to classify contact objects.

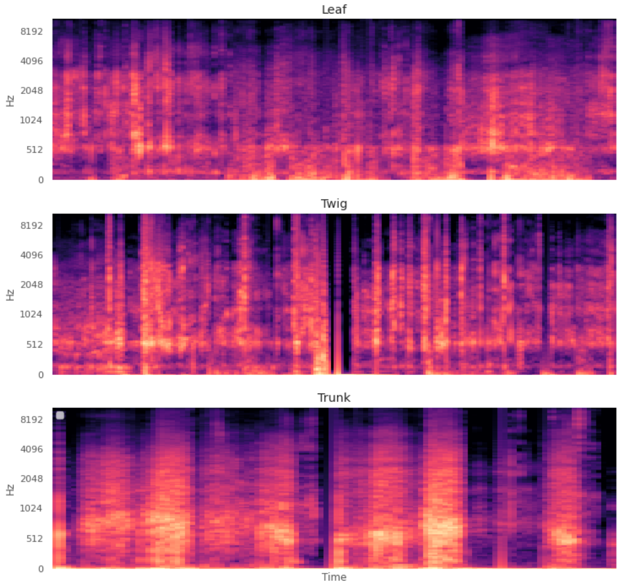

Figure 7: Spectrogram comparison of different classes

We compare audio signals in three different signal representations to distinguish between the object classes: time domain, frequency domain, and Mel Frequency Cepstrum Coefficients (MFCC). As seen in Fig. 6, contact with leaves are mainly rustling motions that result in continuous signals with lower amplitude. Contact with twigs, on the other hand, are mainly snapping motions (like releasing a spring, we are not damaging the plants) that result in jagged signals with high amplitude. Contact with trunks are ”thumps” or signals high amplitude but with longer duration than snapping as seen with twigs. We can more clearly observe this described pattern in the frequency domain as shown in the Spectrogram comparison shown in Fig. 7. As we are ultimately interested in online classification using short duration of audio signals, we use a more dense representations of MFCC.

Across the time duration of 0.1 second, we represent the frequency and energy spectrum into 12 mel coefficients, as shown Fig. 8. As contact sounds are mostly dominated in

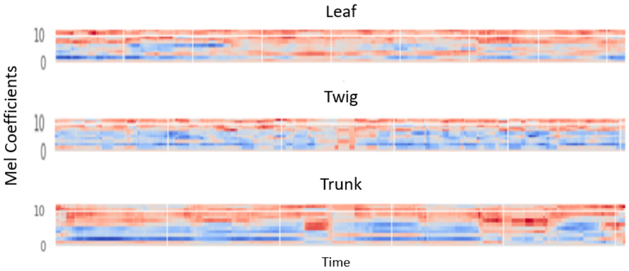

Figure 8: Audio features represented as MFCC plots for for different classes

the low frequency, we see the lower coefficints displaying the strongest responses similarly across all classes. However, we can still visually observe unique patterns across the extracted audio features during contact. As the MFCC responses of the extracted audio signals are visually distinguishable, the authors hypothesize that a classifier could be trained to distinguish between these different types of signals for online contact classification. We leave the training of this classifier to future work.

Discussion

Relying only on vision sensing for tasks such as fruit harvesting in dense canopies will struggle with occlusions and observing latent properties about whether the robot can safely brush against the environment. As such, in this paper, the authors present a novel way to classify contact objects using audio signals in a highly cluttered canopy environment for agriculture manipulation.

To investigate how audio features can help classify contact information, we developed a hand-held device to facilitate the data collection process. From the preliminary dataset, we analyzed audio signals to distinguish between three classes: leaf, twig, trunk. Of the time domain, frequency domain, and MFCC representations, MFCC comparison showed the most distinguishable patterns across the classes.

The provided results present a promising direction to expand this research to leverage deep learning networks to consistently classify the extracted audio inputs. Depending on the reliability of the feature input and network performance, future works could utilize this research in various ways.

For example, when the robot is collecting data in the field, The audio classification could help to automatically label the contact data provided and therefore simplify the robot's motions planning. In addition, after the dataset is augmented with field data with the robot, audio classification could be used to extract more information about the nature of collision, and adjust the trajectory accordingly. In this manner we plan to increase the robustness and the safety of the arm while interacting with plants. By the time of the workshop, the authors intend to augment the preliminary dataset and make it available to public.

References

Cao, X.; Zou, X.; Jia, C.; Chen, M.; and Zeng, Z. 2019. RRT-based path planning for an intelligent litchi-picking manipulators. In *Computers and Electronics in Agriculture*, vol 156. 105–118.

Nemlekar, Z., Heramb. Liu; Kothawade, S.; Niyaz, S.; Raghavan, B.; and Nikolaidis, S. 2021. Robotic Lime Picking by Considering Leaves as Permeable Obstacles. In *International Conference on Intelligent Robots and Systems*.

Sario, S. 1993. Robotics of fruit harvesting: A sate-of-the-art review. In *Journal of agricultural engineering research*, vol. 54, no. 4, pp. 265– 280.

Sawhney, A.; Lee, S.; Zhang, K.; Veloso, M.; and Kroemer, O. 2020. Playing with food: Learning food item representations through interactive exploration. In *International Symposium on Experimental Robotics*, 309–322.

Song, S.; Zeng, A.; Lee, J.; and Funkhouser, T. 2020. Grasping in the Wild: Learning 6DoF Closed-Loop Grasping From Low-Cost Demonstration. In *IEEE ROBOTICS AND AUTOMATION LETTERS*, vol 5. No.3.

Whitney, J. D. 1995. A review of citrus harvesting in florida. In *ASME Citrus Engineering Symposium*, vol. 99823, 33–59. American Society of Mechanical Engineers.

Willigenburg, L.; Hol, C.; and Henten, E. 2004. On-line near minimum-time path planning andcontrol of an industrial robot for picking fruits. In *Computers and Electronics in Agriculture*, vol 44. 223–237.

Yandun, F.; Silwal, A.; and Kantor, G. 2020. Visual 3D Reconstruction and Dynamic Simulation of Fruit Trees for Robotic Manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 54–55.

Zhang, K.; Sharma, M.; Veloso, M.; and Kroemer, O. 2019. Leveraging multimodal haptic sensory data for robust cutting. In *IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*.