Multi-Modal Object Detection and Depth Estimation with Audio Queries

Omkar Shailendra Vengurlekar ***Tharun Poobalan** **Shrinidhi Dnyaneshwar Kumbhar**
MS in RAS (AI)
Arizona State University
ovengurl@asu.edu skumbha4@asu.edu

Abstract

The ability of humans to selectively attend to auditory information about objects in multi-object scenes has significant implications for the development of robots and other artificial agents. However, the extent to which robots are capable of attending to auditory cues to accurately locate and identify objects remains largely unexplored in literature. While much research has focused on developing visual object recognition systems for robots, auditory perception has received comparatively little attention. By better understanding how humans allocate attention in multi-object scenes during auditory perception, we can begin developing effective strategies for robots to localize and identify objects using auditory information. This has the potential to greatly enhance the capabilities of robotic systems. Overall, our study demonstrates the use of auditory perception when developing robotic systems by carefully integrating audio information in the spatial domain and, thus, harnessing the potential of auditory cues for robotic object localization and identification.

1 Introduction

Our multi-modal approach integrates audio encoder into object detection model, reducing the sequential dependency and leveraging parallel compute to provide accurate object detection based on acoustic queries. The proposed neural network architecture combines information from both audio and visual modalities, including object names and visual features, respectively. We use a stereo vision camera setup for object detection and depth estimation, which is crucial for robotics applications such as object manipulation.

The approach enables robots to detect or localize specific objects, using audio descriptions to identify the target object. Accurate depth information is crucial for object manipulation as it allows the robot to understand the spatial relationships between objects and plan its movements accordingly. Our proposed approach enhances the accuracy and robustness of object detection systems but incooperating stereo-vision pipeline and also leveraging human perception's natural biases toward audio information. Overall, the proposed approach has the potential to significantly improve the performance of object detection systems in real-world scenarios in various fields.

2 Related work

There are numerous studies in the literature that have attempted to use speech recognition and object detection for robotic manipulation.

For example, Alami et al. (1), proposed an object-grasping system using a depth camera and speech recognition to facilitate human-robot interaction. Similarly, Redmon et al. (2), presented a system for object detection and grasping using a deep learning-based approach and speech commands.

3 Baseline results

In this study, we utilized the Tacotron model to generate audio data for the Wav2Vec model. The generated data set consisted of 5,400 data points from 108 different speakers, with a sampling rate of 16,000 Hz. To ensure compatibility with the Wav2Vec model, we pre-processed the data by re-sampling it to a 16,000 Hz sampling rate. We fine-tuned the wav2vec-large-xlsr-53-english model on English audio files and their corresponding transcripts for 10 epochs, resulting in a Word Error Rate (WER) of 0.002865 which indicates a good generalization capability of our fine-tuned Wav2Vec model. As this approach has not been tested or implemented before, we do not have any baseline results to compare our findings with.

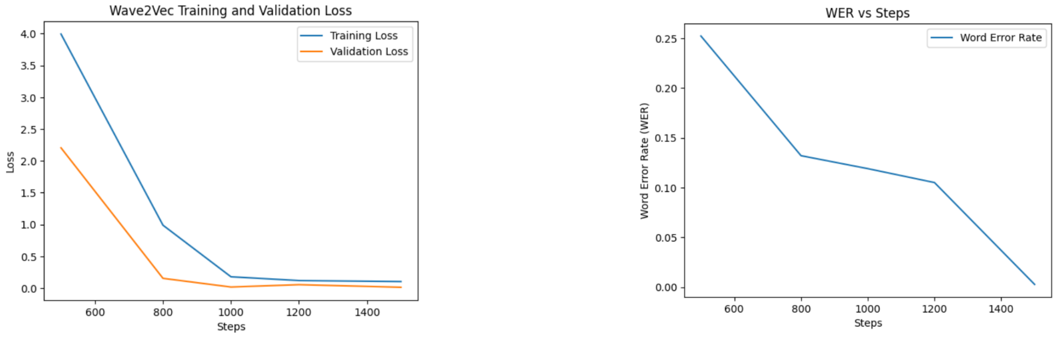

Figure 1: Wav2Vec training and validation losses along with the WER

The training and validation losses were also low, indicating good generalization. Figure 1 shows the Wav2Vec training and validation loss.

A fine-tuned Wav2vec predicts the intended audio perfectly, thus helping to create stable embeddings to pass to the attention mechanism.

We conducted testing on the model using audio files from the test set and comparing the predicted transcription with the original, demonstrating good results.

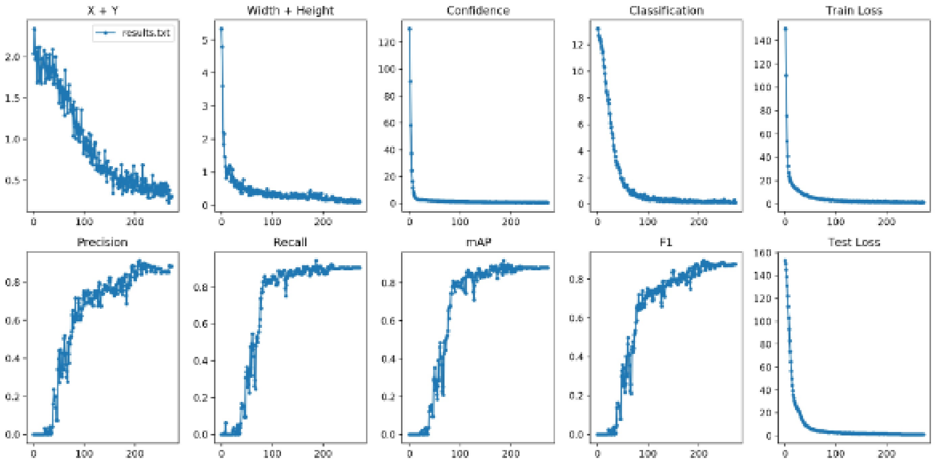

Figure 2: YOLOv3 training

For inference, we input an audio file into the model and confirmed that the transcription matched the ground truth.

4 Approach

Our proposed method involves the integration of two models into a single unified architecture that can detect the objects based on the audio descriptions and estimate the depth of the respective objects. The two models used are Wav2Vec (3)and YOLOv3 (4). The Wav2Vec model is employed for converting audio commands into transcriptions. The YOLOv3 model then detects the objects in the frame and estimates depth based on the object labels identified in the transcription.

Predicted string	Ground truth
Pick up the cup	Pick up the cup
Lift the cup	Lift the cup
Lift the bottle	Lift the bottle
Find the ball	Find the ball
Locate the bottle	Locate the bottle

Table 1: Wav2Vec inference results

Step	Training Loss	Validation Loss	Wer
500	3.993800	0.206009	0.252149
1000	0.179400	0.018349	0.010506
1500	0.104300	0.014598	0.002865

Table 2: YOLOv3 training results

Step	Training Loss	Validation Loss	Wer
500	3.993800	0.206009	0.252149
1000	0.179400	0.018349	0.010506
1500	0.104300	0.014598	0.002865

Table 3: Wav2Vec training results

4.1 Data

To train our baseline Wav2Vec model for audio-based object detection, we used a data set that was generated using the Tacotron2 (5) model to synthesize human audio recordings. The data set consists of 5 different actions (find, grab, lift, pick up, and locate). Ten different objects (bottle, person, apple, banana, cup, phone, bowl, keyboard, book, mouse) from the COCO image classification dataset. We used 108 speaker's audios which ensured that the data set was diverse with different speaker accents.Created 5400 examples by taking permutations of these actions and objects. Used SRGAN (4X) to create high resolution images 4 times the original size. Images for these classes were collected from Roboflow data pool. Converted low resolution images (<416) to high resolution. Rescaled the bounding boxes to new image resolution. Preprocessed each image to get it back to 416x416 resolution (original weights of darknet53 were trained on 416x416).

4.2 Speech recognition and object detection

We preprocessed synthetic audio with 16000 Hz sampling rate and created audio tokens using HuggingFace tokenizers.We fine-tuned wav2vec2-large-xlsr-53-english for our task such that the output of the model is a sequence embedding spanned over it's vocab. We further used wave2vec decoder to convert raw logits to transcripts. For object detection pipeline, we fine-tuned YOLOv3 to detect the 10 classes using the normalized and upscaled images from SRGAN. Output of YOLOv3 are the bounding boxes, classes and their probabilities.

4.3 Unified architecture

To run the unified model, we wrapped the two models using an ONNX wrapper. The batch size was kept 2 during inference for the object detection model where as it was kept 1 for the audio encoder. This gave a multimodal nature to the wrapped model which now accepted two images and an audio array in a single forward pass, distributed over the GPU memory. Initially the stereo vision calibration and rectification was done. We, then built logic for computing depth of the detected objects using disparity. If both the cameras are detecting the same number of objects, we sort the list of detected objects on their x coordinates from left to right. Then, we find the center of the bounding boxes and their depths for each of the detected objects. We end the script by drawing bounding boxes, highlighting their centers, and mentioning the depth of the aforementioned centers.

Figure 3 shows the architecture in a flowchart.

5 Conclusion

Our project presents an efficient solution for multi-modal object detection and depth estimation using audio queries. By combining Wav2Vec and YOLOv3 models into a unified architecture, we achieved remarkable accuracy in object detection and depth estimation. The integration of stereo vision calibration and rectification added another layer of accuracy to our approach. Our proposed method has immense potential in areas such as assistive technology, robot-object localization and

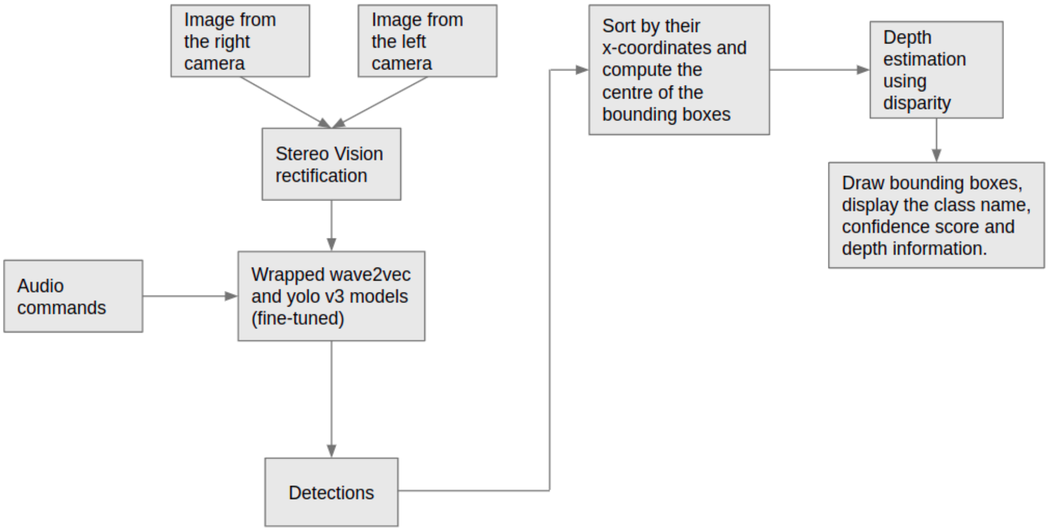
Image from the right camera
Image from the left camera
Stereo Vision rectification
Audio commands
Wrapped wave2vec and yolo v3 models (fine-tuned)
Detections
Sort by their x-coordinates and compute the centre of the bounding boxes
Depth estimation using disparity
Draw bounding boxes, display the class name, confidence score and depth information.

Figure 3: Flowchart explaining the unified architecture

identification, and human-robot interaction. Overall, we believe that our project makes a significant contribution to the field of perception in robotics.

6 Future work

6.1 Cross-modal-attention

The fundamental idea behind using Wav2Vec is to leverage the acoustic information to localize a specific object in a given scene. This is achieved through the implementation of a cross-modal attention mechanism. The intermediate image embedding is first extracted from the Darknet53 backbone of YOLOv3. A series of convolution layers help align the shape of the image embeddings with the audio embeddings. This is done because the attention mechanism requires all tensors to have the same shape.

As Wav2vec employs a vocabulary size of 33 (inclusive of special tokens), the attention heads are divided into three heads, resulting in a multi-headed approach that facilitates a more diverse retrieval of information. The output of the multi-head attention is then projected back to a shape that is similar to that of the output of the Darknet53 backbone. This ensures a seamless transition of the cross-modality to the heads of YOLOv3.

The utilization of this architecture allows for the extraction of perfect semantics from the audio, which can be used to enhance the object detection process. By leveraging the acoustic information in conjunction with the image embedding, the cross-modal attention mechanism can effectively highlight important features of the scene, leading to more accurate object localization.

6.2 Partial audio matches or visibility

Oftentimes, the speech instructions might be muffled and that might lead the Wav2Vec model to be unable to capture the speaker's true intent. For example, if the speaker said "Find the broccoli" and there was some disturbance, the model might hear "Bind the brokoly". This transcription does not match any intent the model was trained to capture and has incorrect English. Instead of ignoring this input like the model does today, we plan to match the tokens to their nearest recognizable counterparts. Hence, this transcription would get corrected to "Find the broccoli".

Similarly, the object being manipulated might lie on the periphery of the field of view of the camera. Today, the model fails to recognize the object. However, we would like it to partially match the objects at the periphery.

References

[1] R. Alami, R. Chatila, S. Fleury, M. Ghallab, and F. Ingrand, “An architecture for autonomy,” *The International Journal of Robotics Research*, vol. 17, no. 4, pp. 315–337, 1998.

[2] J. Redmon and A. Angelova, “Real-time grasp detection using convolutional neural networks,” 2015.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” 2019.

[4] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” 2018.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” 2017.