Single-Target Audio-Visual Learning and Navigation in Search and Rescue Scenarios: Transfer Application to Physical Robot

by

Md Amanullah Kabir Tonmoy

Nathanael Oliver

12.05.2023

UNIVERSITY OF South Carolina

College of Engineering and Computing

Introduction

- **Locating sound in 3D Spaces with RIRs**
- **Introducing variables into environment**
  - Robots
  - Humans
  - Predictive RIRs using Camera and Binaural Audio Source

Introduction

- **Room Impulse Response (RIR)**
  - Room Geometry
  - Objects in Room
  - Position of Objects
  - Material of Objects
- **Applications with Humans in Room**
  - SoundCam Dataset
  - Estimating Environment Geometry with Sound and Vision

Application in Robotics

- **Motivation**
  - In autonomous navigation, the robot operates in an environment without having access to any reference map.
  - Active Simultaneous Localization and Mapping(ASLAM)
    - relying on information collected from sensors, such as camera and lidar,
    - construct the map while planning a path through the environment which is
    - Time consuming and often inaccurate due to sensor
  - End to End Policy learning
    - Can be applied to extract semantic feature to direct search
      - Smell
      - Audio
  - Audio-Visual cues
    - Multi-modal deep reinforcement learning by Chen et al. [1]
    - Find a single sound source in unknown environment
  - Singhal et al. extended the work [2]
    - To navigate multiple audio sources
    - Used transfer learning to reduce training cost
  - In this project, we try to replicate Chen et al.'s work using Replica Dataset and implement it in physical robot.

[1] Chen, Changan, et al. "Soundspaces: Audio-visual navigation in 3d environments." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer International Publishing, 2020.
[2] K. Singhal, M. Yaghouti, P. Jamshidi, Multi-Sense-Rescuer: Multi-Target Audio-Visual Learning and Navigation in Search and Rescue Scenarios

Application in Robotics

- **Problem Definition**
  - The problem at hand is to train an agent to navigate to a single audio source based on audio and visual cues with no reference map and evaluate its performance in real world.
- **Related Works**
  - Audio-Visual Learning: Focuses on human captured video rather than embodied perception
    - Synthesizing sounds for video[1]
    - Spatializing sound [2]
    - Sound source separation [3]
    - Cross-modal feature learning [4]
    - AV tracking [5]
  - Vision-based navigation:
    - AI agents can navigate based on visual inputs combined with spatio-temporal memory [6]
    - Visual Navigation can be tied to other tasks to get intelligent behavior [7]
  - Audio-based navigation:
    - Audio based equipment has been used to avoid obstacle and navigation [8]

[1] Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. ACM (2017)
[2] Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. In: NeurIPS (2018)
[3] Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV (2018)
[4] Yusuf Aytar, Carl Vondrick, A.T.: Learning sound representations from unlabeled video. In: NeurIPS (2016)
[5] Gebru, I.D., Ba, S., Evangelidis, G., Horaud, R.: Tracking the active speaker based on a joint audio-visual observation model. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 15–21 (2015)
[6] Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning. In: ICRA (2017)
[7] Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2616–2625 (2017)
[8] Massiceti, D., Hicks, S.L., van Rheede, J.J.: Stereosonic vision: Exploring visualto-auditory sensory substitution mappings in an immersive virtual reality navigation paradigm. PloS one (2018)

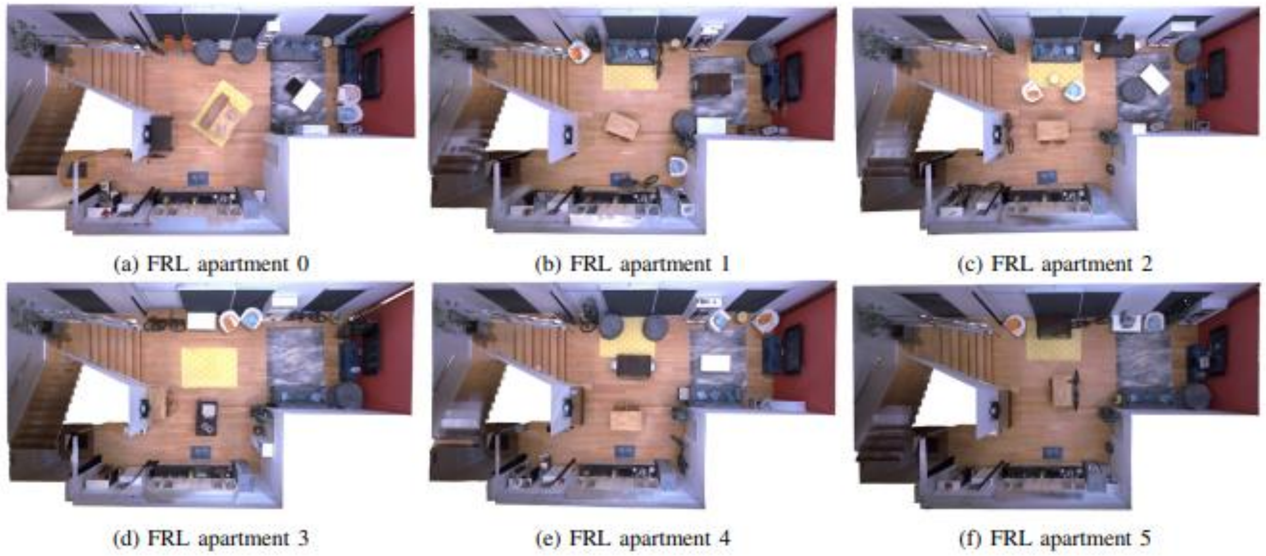
Technical Approach

- **Dataset**
  - Replica3D
- **Packages**
  - Sound Space
    - high-level APIs for navigation tasks
  - Habitat Simulator
    - offers a range of sensors, including
      - a RGB camera,
      - a depth sensor,
      - and a GPS, which provides the target location in the agent
- **Action Space**
  - moving forward 0.5 meters,
  - rotating 10 degrees clockwise or anticlockwise
  - STOP
- **Sensors**
  - Depth sensor, RGB and Audio sensor

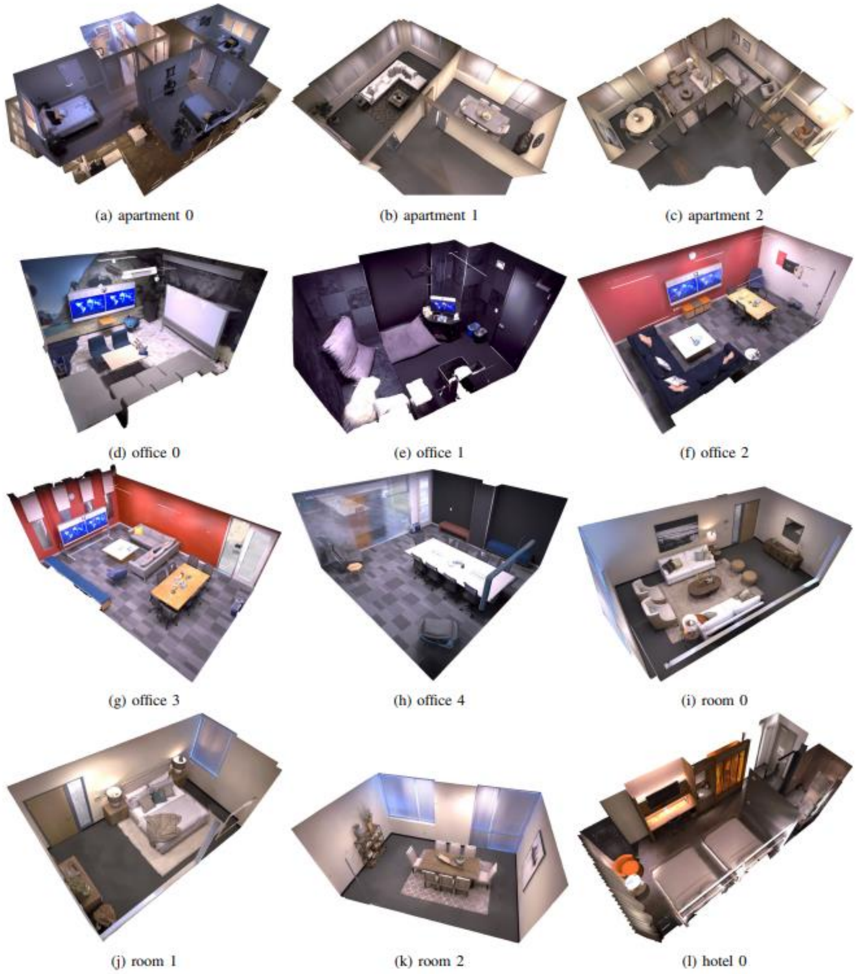
- **Metrices**
  - Average Success = $\frac{1}{N}\sum_{i=1}^{N} S_i$
  - Average SPL = $\frac{1}{N}\sum_{i=1}^{N} \frac{S_i l_i}{\max(p_i, l_i)}$
    - S_i = Flag whether the i-th episode is successful or not
    - l_i is the shortest path distance to succeed in i-th episode
    - p_i is the path length traversed by agent in i-th episode
- **Experiments**
  - Evaluate a pretrained model on the dataset.
  - Train the model from scratch and evaluate the performance.

Technical Approach

- **Dataset**
  - Replica3D



6 scenes of Apartment

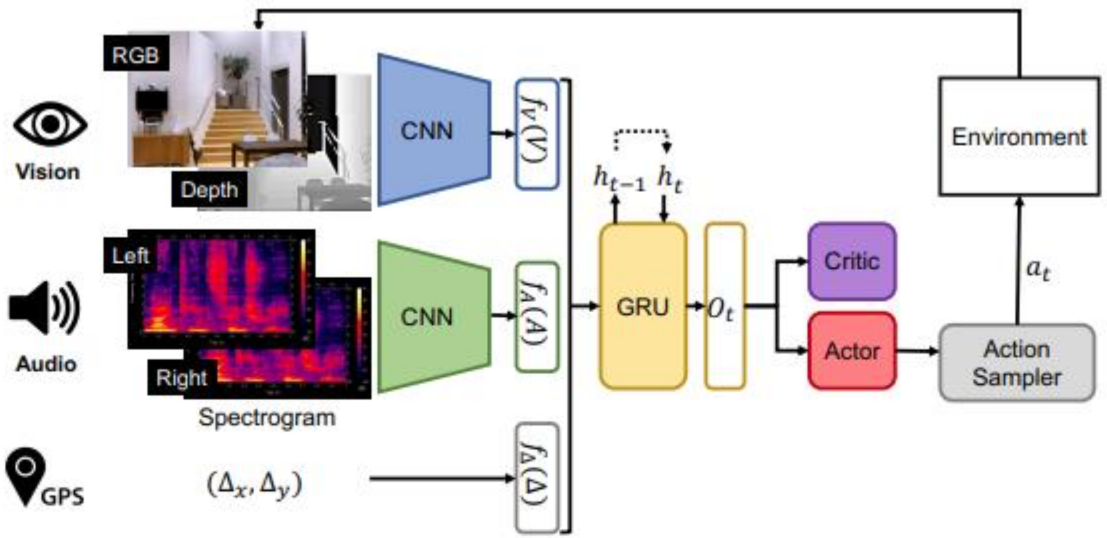


12 scenes of different constructions

[1] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)

Technical Approach

- **Model**
  - Audio and visual cues are used
  - Trained based on Proximal Policy Optimization
  - The agent is rewarded for reaching the goal quickly.
  - Specifically, it receives a
    - reward of +10 for executing Stop at the goal location,
    - a negative reward of −0.01 per time step,
    - +1 for reducing the geodesic distance to the goal, and the equivalent penalty for increasing it.
    - add an entropy maximization term to the cumulative reward optimization, for better action space exploration
  - Adam Optimizer: 2.5e-4 learning rate



[1] Chen, Changan, et al. "Soundspaces: Audio-visual navigation in 3d environments." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer International Publishing, 2020.

Results

- **Results from pretrained model**
  - We used a pretrained model from sound-space and evaluate it on Replica3D dataset.

Model	Average Success	Average SPL
Pre-trained Model	0.946	0.793

Rendered video is from Test Case (Apartment 1). Green Line indicates optimal path and Blue Line indicates the agent's path.

Results

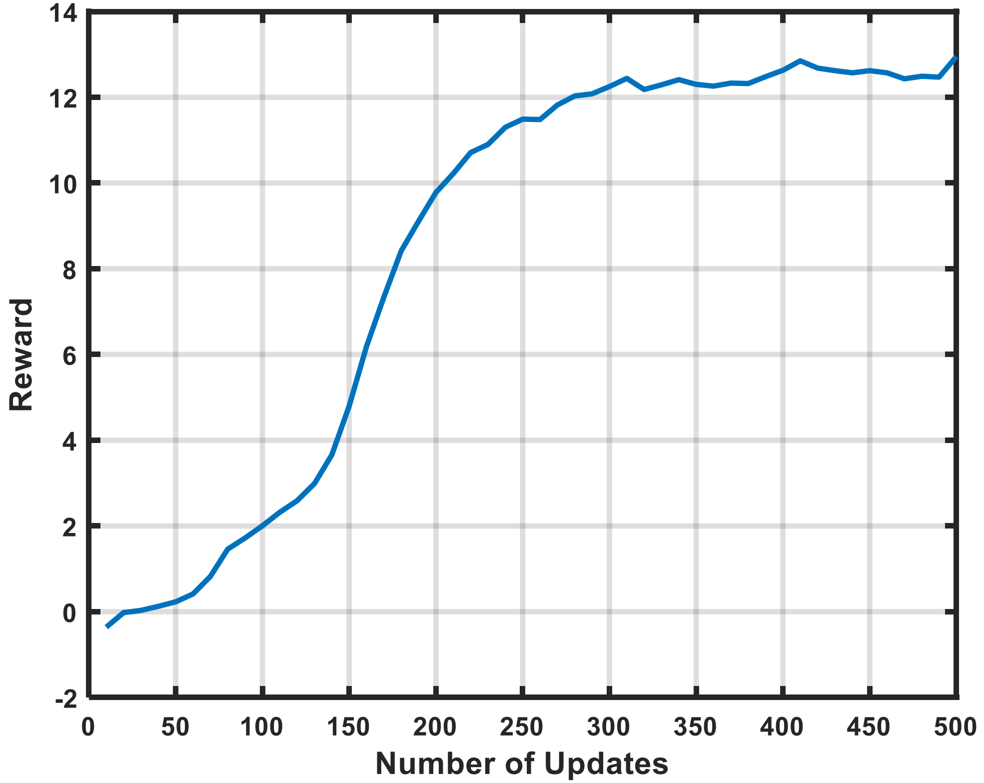
- **Training the model from scratch**
  - We trained a model from scratch using Replica3D dataset and test it on this dataset. Test cases are not used in training.

Model	Average Success	Average SPL
Training from scratch	0.644	0.35

Rendered video is from Test Case (Apartment 1). Green Line indicates optimal path and Blue Line indicates the agent’s path.

Discussion

- **Discussion:**
  - To train a model from scratch takes a longer time. The actual model was trained for 40000 updates whereas our model is trained for only 500 updates (a week).
  - Evaluation of the checkpoints also takes time (Every checkpoint is evaluated for around 3 hours).
  - If we train our model for that longer time, we can achieve similar results as pretrained model.



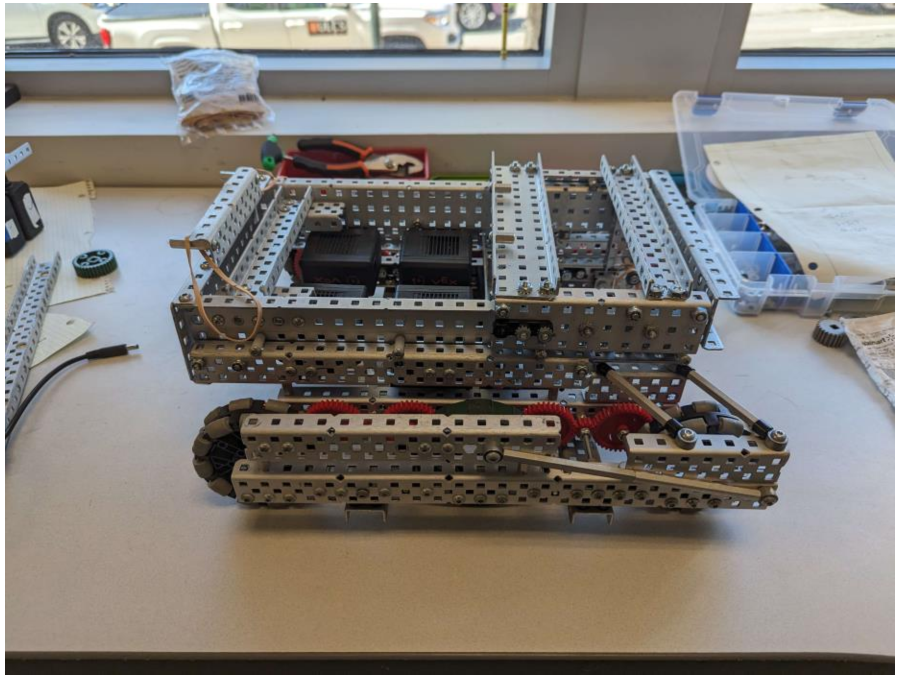
From the graph, it is shown that our model is not converged yet.

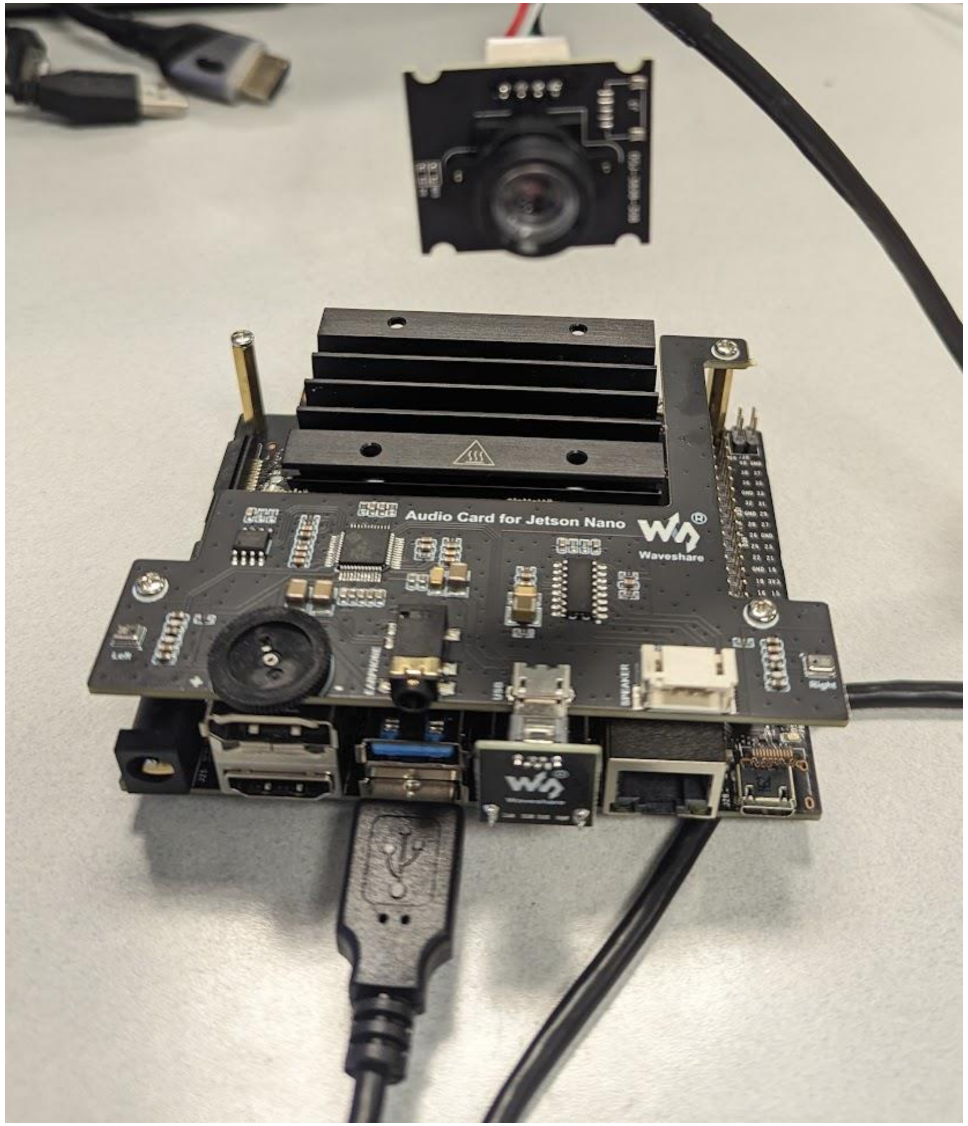
Challenges & Takeaway

- **Challenges:**
  - Version control of Habitat-Simulator and Sound spaces
  - Hardware Requirement Fulfillment
  - Longer time for training and validation of model
- **Takeaway:**
  - Use Cloud computation
  - Use GitHub
  - Reinforcement learning
  - Proximal Policy Optimization
  - Recent advancement in autonomous navigation using RL

Physical Implementation

- **Hardware**
  - Jetson Nano
  - Spectrogram Sensor
    - Detects Sound in 3D space
  - RGB Camera
  - Depth Sensor





Spectrogram Sensor

- **Sensor**
  - Uses 2 Microphones a distance apart to detect changes in waves and locate source of sounds
  - Uses Short Term Fourier Transform to analyze frequency content
  - Extracts features for use in Model

Audio Sample taken from Testing, showing left and right microphone values

Physical Implementation

- **Deintegrating Agent from Habitat-sim**
- **Integration with Physical Sensors**
- **Integration with Physical Robot**

Thank You

UNIVERSITY OF
South Carolina
College of Engineering and Computing