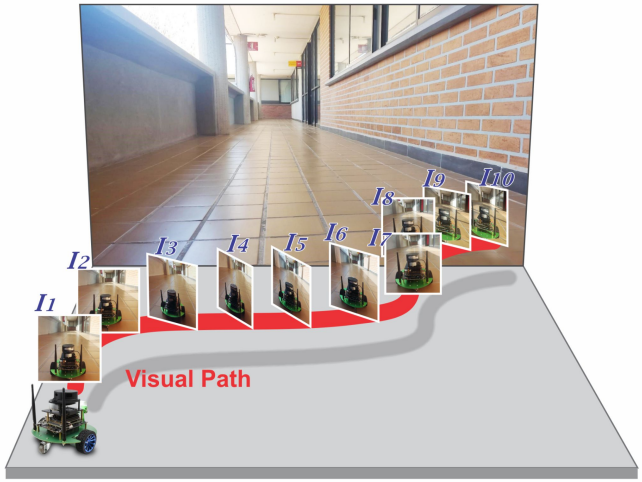
SoundSpaces: Audio-Visual Navigation in 3D Environments

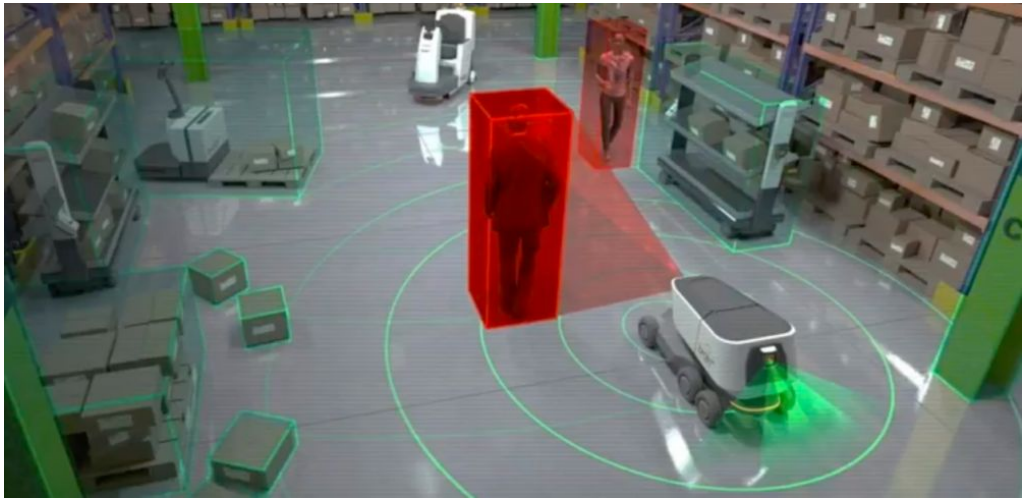
Presenter: Kevin Yang

9/14/2023

Audio-Visual Embodied Navigation

- Embodied AI
  - Navigation task
  - Interact and understand the physical world through sensory inputs (Vision, Sound, Touch, …)



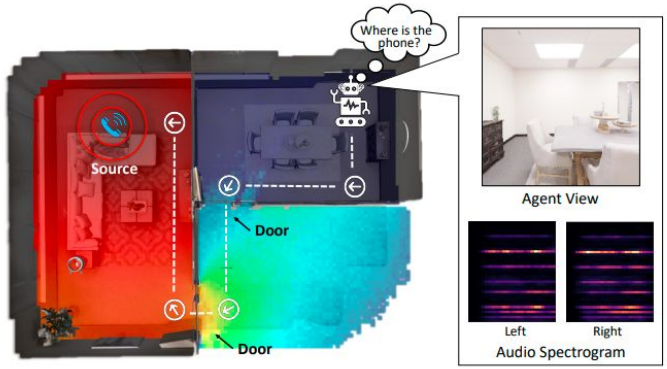


Motivation and Main Problem

- Today's embodied agents are **restricted** solely on their visual cues
- Sound is **key** to:
  - Understanding a physical space
  - Localizing sound-emitting targets
- Aural cues are critical when visual cues are unreliable
  - Lights turn off, person calling from upstairs, ringing phone occluded by the sofa, …

Motivation and Main Problem

- How do we enable agents to leverage both modalities for perception and decision making when navigating?
- How can these same agents learn to generalize to new environments and new sounds?



Related Work

- Audio-visual learning
  - Recent work focuses on human-captured video rather than embodied perception
  - Localizing pixels in video frames associated with sound
- Vision-based navigation
  - Training agents to navigate an environment based on visual inputs
  - Question answering, active visual recognition, and instruction following
- Audio-based navigation
  - Audio-based AR/VR equipment used for auditory sensory substitution
  - AV environments are non-photorealistic and are used specifically for human navigators

Related Work - Cont’d

- Sound localization in robotics
  - Prior work fuses AV cues for surveillance, speech recognition, human robot interaction, …
  - None attempt AV navigation in unmapped environments
- 3D environments
  - Photorealistic scenes that portray 3D scenes that real people and robots can interact with
  - Popularly used environments do not provide audio rendering

Contributions

- Introduce SoundSpaces - AV platform for embodied AI
- Introduce the task of AV navigation by autonomous agents in visually and acoustically realistic 3D environments
- Generalize deep RL visual framework to accommodate audio observations
- Create a benchmark suite of tasks for audio-visual navigation

SoundSpaces: Background

- Audio platform that augments the Habitat Simulator, focusing specifically on creating realistic sound rendering in Matterport3D and Replica datasets
  - Habitat - open-source 3D simulator with an API that supports fast RGB, depth, and semantic rendering
  - Matterport3D, Replica - real world indoor environments that contain 3D meshes and image scans

SoundSpaces

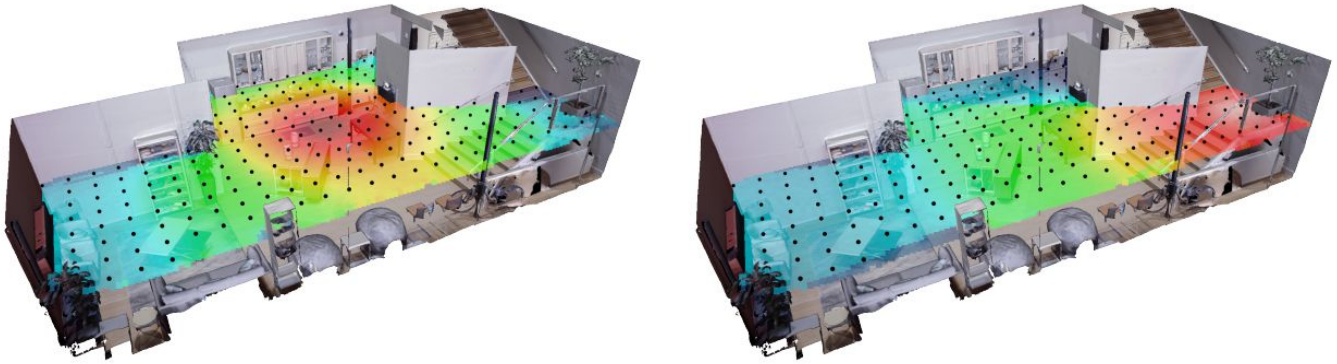
- How does SoundSpaces generate **realistic** sound rendering in a 3D environment?
  - State-of-the-art algorithm (Bidirectional Sound Transport) for room acoustics modeling
  - Bidirectional path tracing algorithm to model sound reflection in the room geometry
- Materials influence sounds received by the environment
  - Each material has different absorption, scattering, and transmission coefficients that affect sound propagation
  - Set acoustic material properties of major surfaces (meshes' semantic labels - materials)

SoundSpaces - Cont’d

- Simulate the acoustics of the environment by pre-computing room impulse responses (RIR)
  - RIR - transfer function between a sound source and microphone
  - Generate ambisonic audio by convolving a desired waveform with a RIR (ambisonic -> binaural)
- Reachability

$$\mathcal{S} = \{(x_i^s, y_i^s, z_i^s)\}_{i=1}^N$$

$$\mathcal{L} = \{(x_i^r, y_i^r, z_i^r)\}_{i=1}^N$$



Task Definitions for AV Navigation

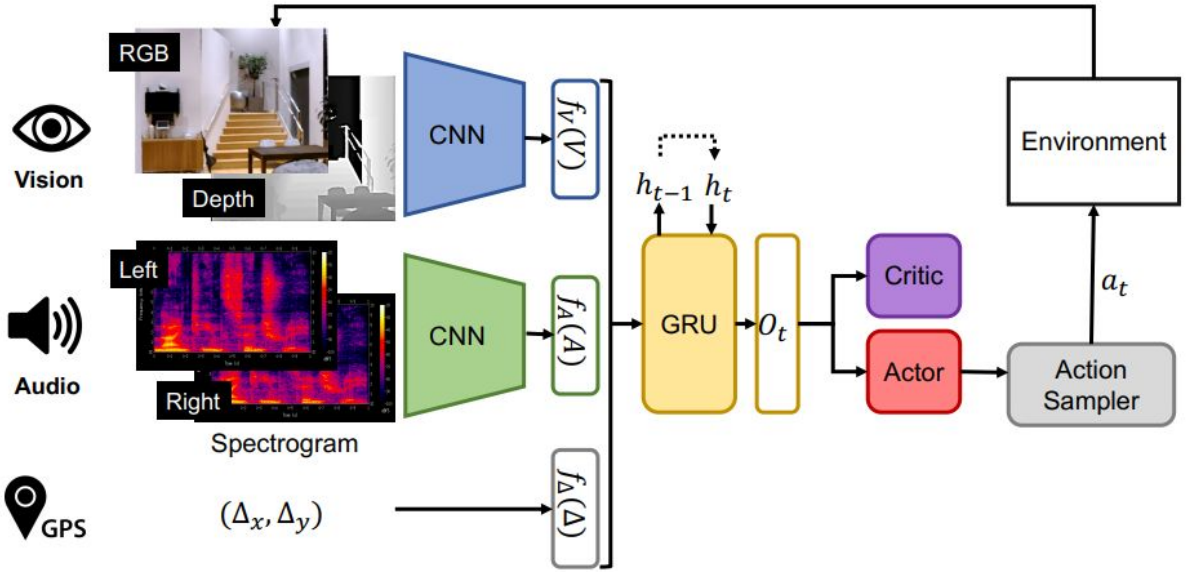
- Task Definitions:
  - PointGoal - agent is tasked with navigating to a point goal that is defined by a displacement vector (Δ_x^0, Δ_y^0)
  - **AudioGoal** - agent instead receives audio from the sounding target
  - **AudioPointGoal -** agent receives audio and a point vector
- Agent and goal embodiment
  - Goal does not have a visible embodiment
  - Vision is important to detect and avoid obstacles
  - All tasks have a vision component

Task Definitions for AV Navigation - Cont’d

- Action Space:
  - Forward, Left, Right, and Stop
  - Forward - invalid if it attempts to traverse a pair of nodes without an edge connect between the pair
- Sensors:
  - PointGoal - GPS, RGB, and depth
  - AudioGoal - binaural sound, RGB, and depth
  - AudioPointGoal - binaural sound, GPS, RGB, and depth
- Episode specification:
  - PointGoal - defined by a random scene, agent start location, agent start rotation, and goal location
  - AudioGoal and AudioPointGoal - source audio waveform that is convolved with the RIR corresponding with the parameters above
  - Episode is successful if agent executes Stop at the location of the goal

Navigation Network

$\pi_\theta(a_t|o_t, h_{t-1})$ $V_\theta(o_t, h_{t-1})$



Training Procedure

- Proximal Policy Optimization(PPO)
- Agent Reward Function:
  - Receives a reward of +10 for executing Stop at the goal location
  - Negative reward of -0.01 per time step
  - +1 for reducing the geodesic distance of the goal
  - -1 for increasing the geodesic distance of the goal
- Entropy maximization term
  - Used for cumulative reward optimization for better action space exploration

Experimental Setup

- Datasets:
  - Each episode consists of a tuple
    - <Scene, agent start location, agent start rotation, goal location, and audio waveform>
  - Prune episodes that are either too short or can be completed by moving in a straight line
  - Ensure that at the onset of each episode the agent can hear sound

Table 1: Summary of SoundSpaces dataset properties

Dataset	# Scenes	Resolution	Sampling Rate	Avg. # Node	Avg. Area	# Training Episodes	# Test Episodes
Replica	18	0.5m	44100Hz	97	47.24 m^2	0.1M	1000
Matterport3D	85	1m	16000Hz	243	517.34 m^2	2M	1000

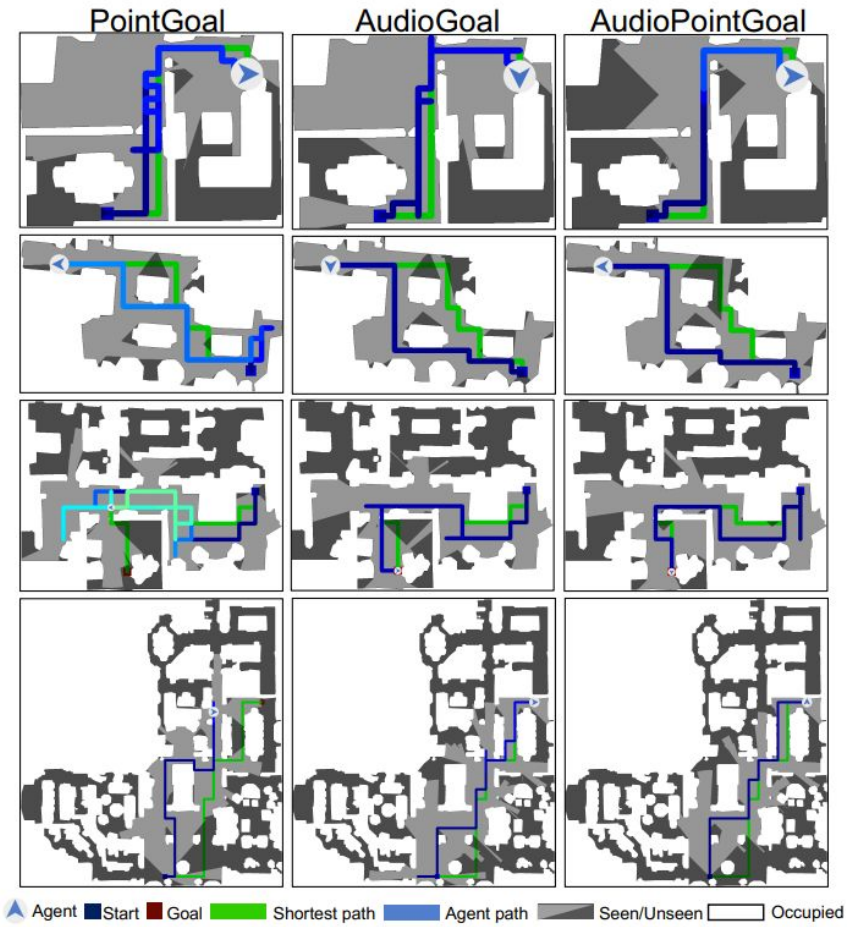
Experimental Setup

- Sound sources
  - RIRs can be convolved with a random input waveform, which allows for varying sounds across episodes
- Metrics
  - SPL - success rate normalized by inverse path length
  - Episode is successful if the agent reaches the goal and executes the Stop action
- Non-learning Baselines
  - RANDOM - chooses random action from {MoveForward, TurnLeft, TurnRight}
  - FORWARD - calls MoveForward and if it hits and obstacle, it calls TurnRight
  - GOAL FOLLOWER - orients itself towards the goal and calls MoveForward

Experimental Results

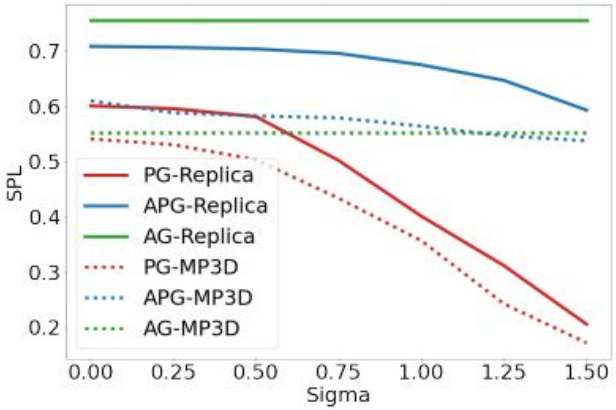
- Does audio help navigation?

		Replica		Matterport3D	
		PointGoal	AudioPointGoal	PointGoal	AudioPointGoal
Baselines	RANDOM	0.044	0.044	0.021	0.021
	FORWARD	0.063	0.063	0.025	0.025
	GOAL FOLLOWER	0.124	0.124	0.197	0.197
Varying visual sensor	Blind	0.480	**0.681**	0.426	**0.473**
	RGB	0.521	**0.632**	0.466	**0.521**
	Depth	0.601	**0.709**	0.541	**0.581**

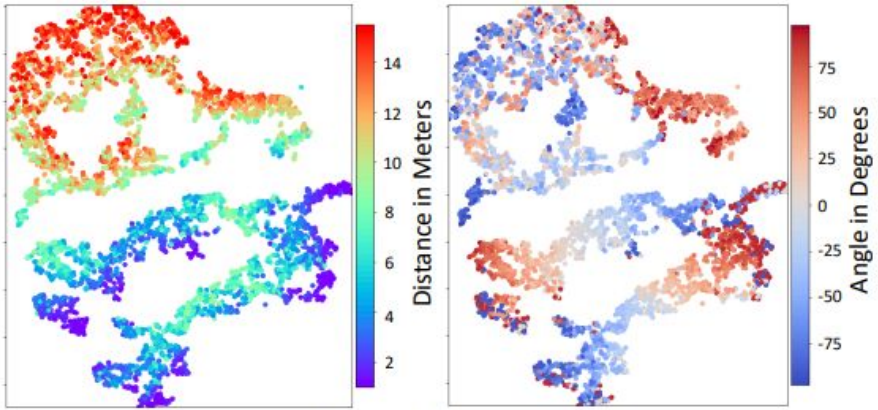


Experimental Results

- Can audio supplant GPS for an audio target?



(a) From perfect to noisy GPS



(b) t-SNE of AudioGoal audio features

Experimental Results

- What is the effect of different sound sources?

Dataset		PG	*Same sound* AG	*Same sound* APG	*Varied heard sounds* AG	*Varied heard sounds* APG	*Varied unheard sounds* AG	*Varied unheard sounds* APG
Replica	Blind	0.480	0.673	0.681	0.449	0.633	0.277	0.649
	RGB	0.521	0.626	0.632	0.624	0.606	0.339	0.562
	Depth	0.601	0.756	0.709	0.645	0.724	0.454	0.707
Matterport3D	Blind	0.426	0.438	0.473	0.352	0.500	0.278	0.497
	RGB	0.466	0.479	0.521	0.422	0.480	0.314	0.448
	Depth	0.541	0.552	0.581	0.448	0.570	0.338	0.538

Discussion of Results

- Audio enhances navigation in complex 3D environments
  - Synergy of audio and vision improves agent performance
  - Faster learning
  - More accurate navigation outcomes
- Audio competes well with traditional GPS-like methods
  - Reduced reliance on perfect GPS odometry
- Agents demonstrate the ability to generalize to various sound scenarios

Limitations

- Agent in the simulator can only move or hop between discrete grid points
  - Abstracts away difficult parts of the navigation task
- Simulations do not generalize to novel environments
- Only supports audio rendering for simple 3D environments

Future Work

- Multi-agent scenarios
- sim2real transfer
- Moving sound-emitting targets
- Navigating in the context of dynamic audio events

Extended Readings

- Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)
  - Audio-visual learning
- Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
  - Habitat
- Chang, A., Dai, A., Funkhouser, T., , Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: Proceedings of the International Conference on 3D Vision (3DV) (2017)
  - Matterport3D
- Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv (2017)
  - Replica

Summary

- Problem: Existing work on embodied agents focuses too heavily on visual data, which can be limited or unreliable for specific scenarios
- SoundSpaces enables audio rendering for Habitat in Replica and Matterport3D environments
- When closely integrated with egocentric visual observations, audio enhances directional cues for sound sources while also enriching spatial information about the environment
- SoundSpaces RL model enables embodied agents to generalize audio-based navigation across previously unheard sound scenarios

Thank You!