SFU SIMON FRASER UNIVERSITY

Embodied Human Activity Recognition

Sha Hu, Yu Gong, Greg Mori
Simon Fraser University
8888 University Drive Burnaby, B.C. Canada. V5A 1S6

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Contributions

- **A novel task** of intelligently **moving** an embodied agent in 3D environments to recognize **ongoing** human activities**;**
- **A reinforcement learning based approach** that learns an effective policy, outperforming heuristic baselines;
- Datasets preparation, code and videos are at:

Our Task

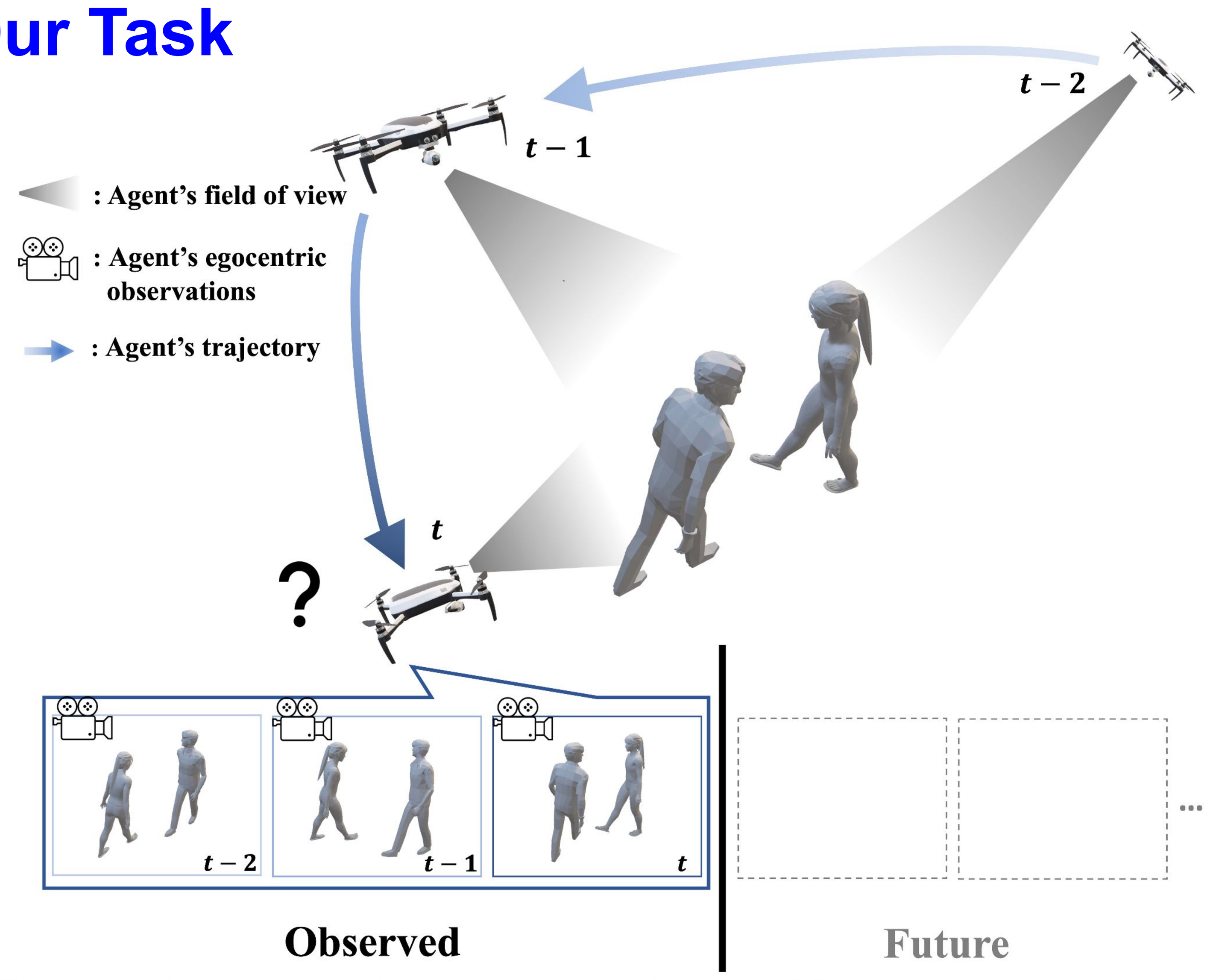
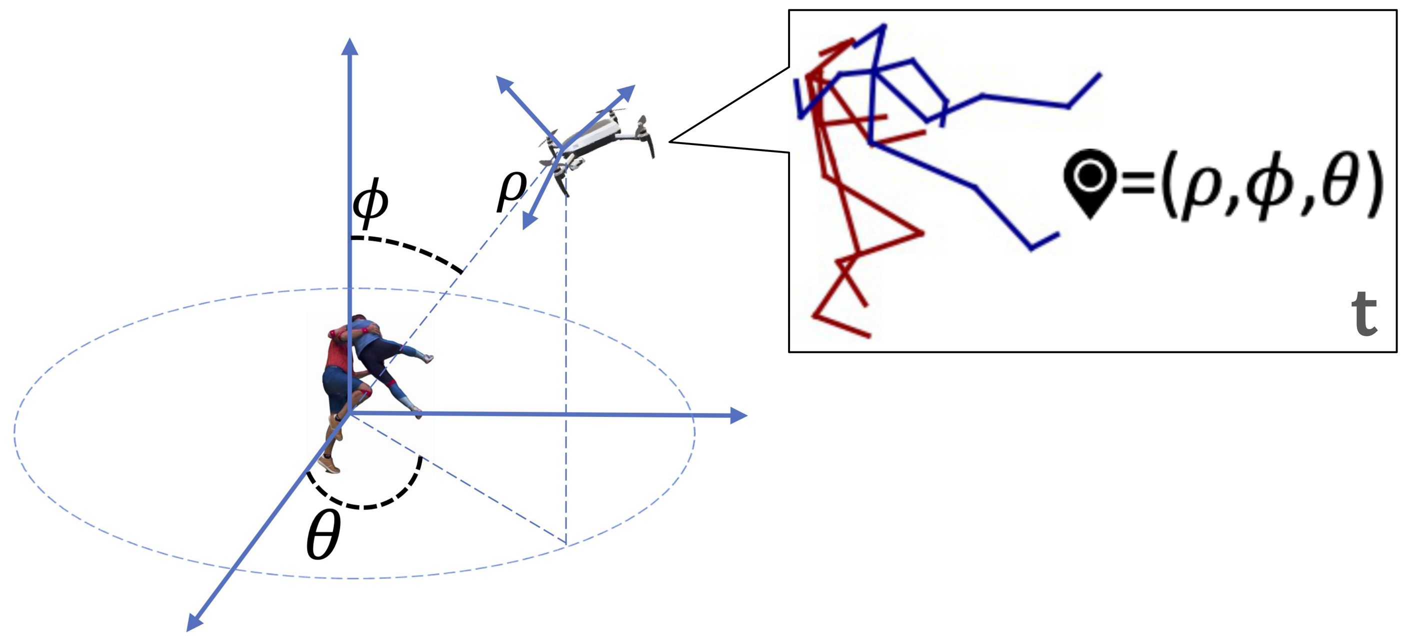


Figure 1. An embodied agent is operating in a 3D environment.

- **Embodied Human Activity Recognition (EHAR)**: The agent is tasked to intelligently **move around** using cues from its egocentric observations so that it can accurately **classify an ongoing human activity**.

Task Formulation

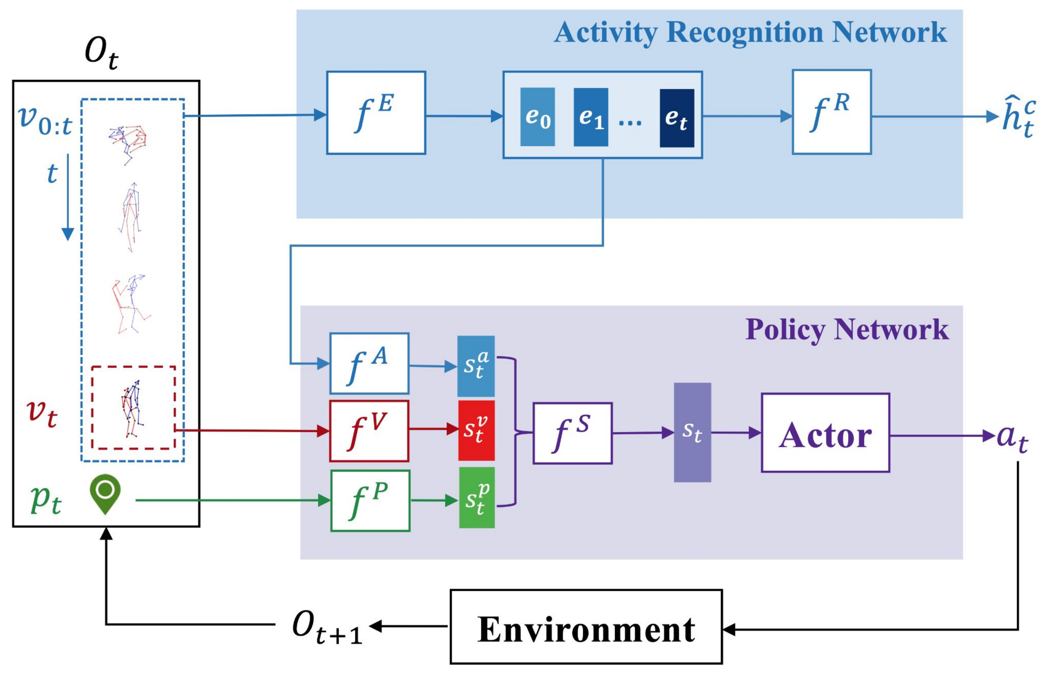
- EHAR as Contextual Partially-Observable Markov Decision Processes, defined by $\langle \mathcal{H} \times \mathcal{P}_0, L, \mathcal{S}, \mathcal{O}, \mathcal{A}, T, R \rangle$:
  - $\mathcal{H}$: human activity senarios; e.g., ExPI [1] & AIST++[2]
  - $\mathcal{P}_0$: the agent's starting positions;
  - L: the agent is allowed to take up to L actions;
  - $\mathcal{O}$: the agent receives 2D human skeletons and its positions as observations;
  - $\mathcal{A}$: the agent moves along axes in the spherical coordinate system to acquire new observations.



- Evaluation Metrics:
  - accuracy against observation ratios;
  - average accuracy across all time steps.

Approach

- The goal of the agent is to learn which movement to take based on its observations to maximize the chances of correctly classifying the activity.



- The agent contains:
  - An Activity Recognition Network that receives 2D human skeletons and predicts an activity label based on observations up to the current time;
  - A Policy Network that receives representations of the recognition state, current visual observation and spatial positions; and predicts an movement decision for the next time step.

Two-phase training:

- Phase 1: Imitation Learning from **Oracle**;
  Oracle is:
  - greedy: one-step cross-entropy reduction;
  - unrealistic: require the true label and act on hindsight.
- Phase 2: Fine-tuning by reinforcement learning.

Results and Findings

- Experimental setup: we use real motion capture data from ExPI[1] and AIST++[2] to simulate embodied agents and evaluate our method.

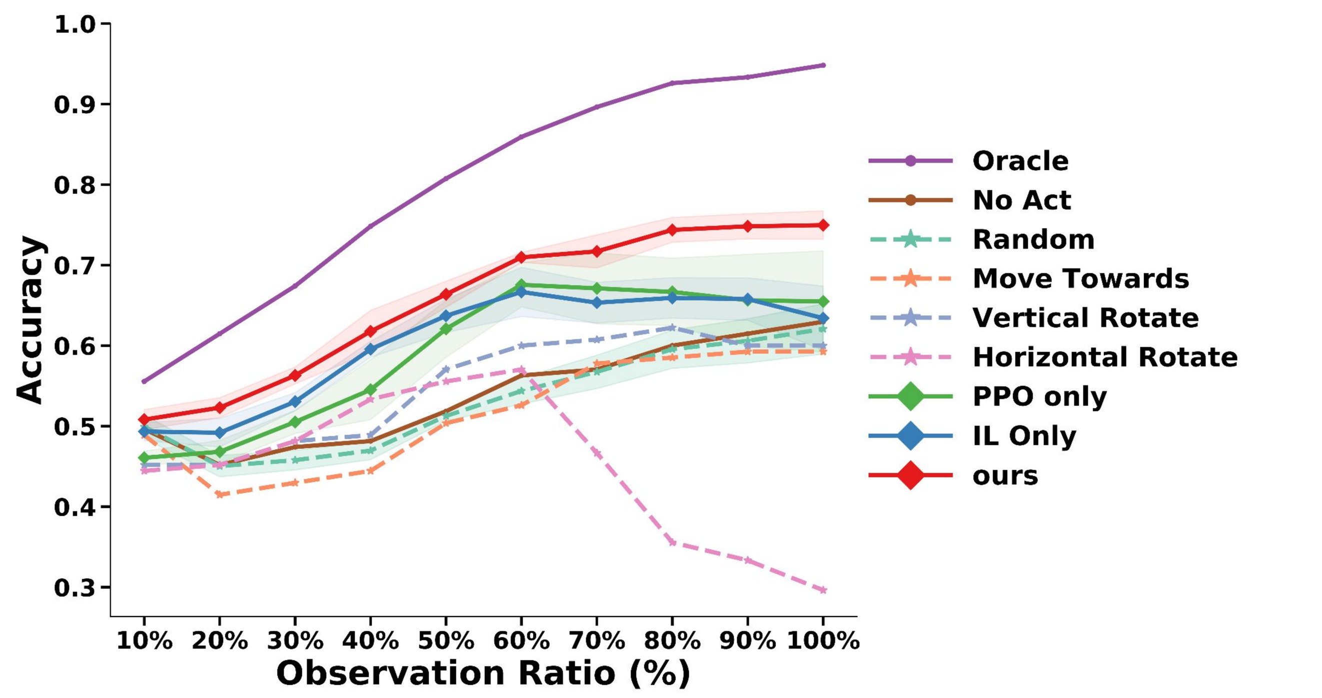


Figure 4. Curves of accuracy against observation ratios on ExPI.

			acc@10	acc@30	acc@50	acc@70	acc@90	acc@100	$\overline{\text{acc}}$
Oracle			55.56	67.41	80.74	89.63	93.33	94.81	77.53
Passive	No-act		49.63	47.41	51.85	57.04	61.48	62.96	52.79
Embodied	Heuristic	Towards	48.89	42.96	50.37	57.78	59.26	59.26	50.64
		H-rotate	44.44	48.15	55.56	46.67	33.33	29.63	45.23
		V-rotate	45.19	48.15	57.04	62.22	60.74	60.00	54.52
		Random	49.93 ± 1.44	45.78 ± 1.22	51.26 ± 0.62	56.74 ± 2.07	60.59 ± 2.74	62.07 ± 3.12	52.45 ± 0.86
	Learning	PPO only	46.07 ± 0.81	50.52 ± 1.43	62.07 ± 3.53	67.11 ± 4.40	65.63 ± 5.70	65.48 ± 6.28	58.51 ± 2.45
		IL only	49.33 ± 1.12	53.04 ± 1.12	63.70 ± 2.10	65.33 ± 2.53	65.78 ± 2.64	63.70 ± 3.98	59.61 ± 0.78
		Ours	**50.81** ± 1.24	**56.30** ± 1.05	**66.37** ± 1.62	**71.70** ± 2.06	**74.81** ± 1.57	**74.96** ± 1.77	**64.27** ± 0.33

- Oracle acts as "best case" performance of an agent
  - potential improvements exist: gap between ours and oracle.
- Comparisons:
  - passive agent **VS.** embodied agents;
  - heuristic agents **VS.** our agent;
  - ours **VS.** ablations of training phases.

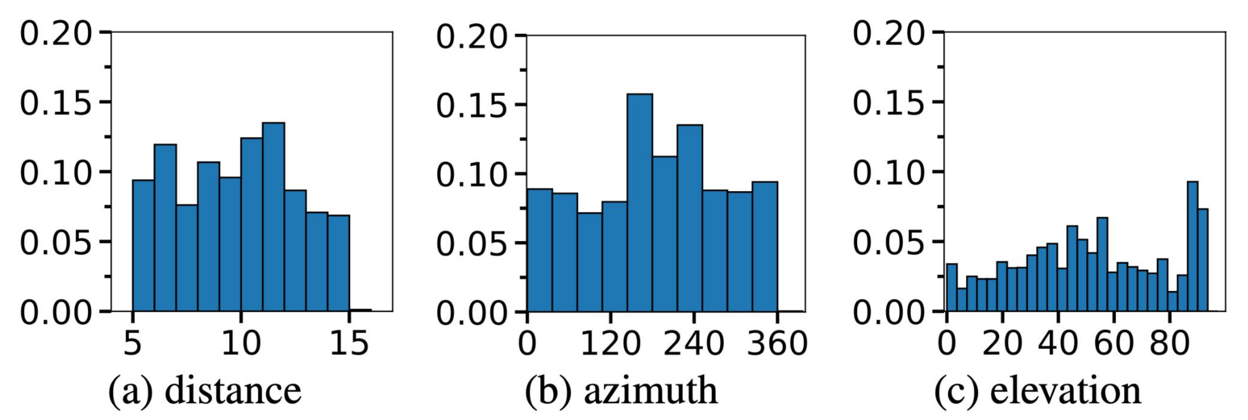
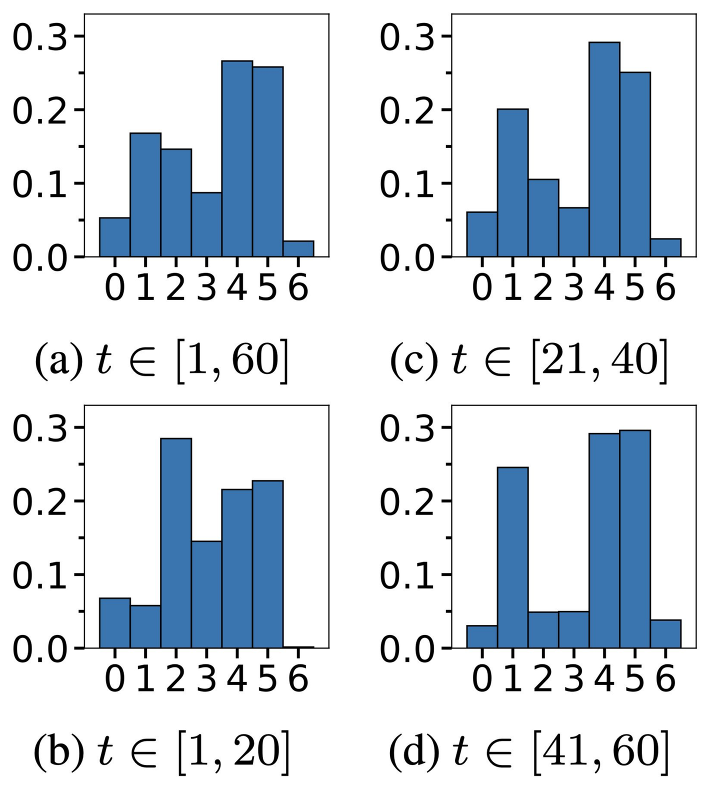


Figure 5. Histograms of our **learned** agent's positions.

- **Learned** agent's **spatial** behavior patterns:
  - agent's movements span widely;
  - no cannonical position favored for EHAR; good viewpoints might be diverse.



a=0/1 means $\rho \pm \Delta\rho$;
a=2/3 means $\Phi \pm \Delta\phi$;
a=4/5 means $\Theta \pm \Delta\theta$;
a=6 means hovering.

Figure 6. Histograms of our **learned** agent's actions across different stages of episodes.

- **Learned** agent's **temporal** behavior patterns:
  - agent's active adjustments of its 3D positions by movements;
  - good viewpoints change along with motions of human activities.

References:
[1] Guo, W., Bie, X., Alameda-Pineda, X., & Moreno-Noguer, F. (2022). Multi-Person Extreme Motion Prediction. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). [2] Li, R., Yang, S., Ross, D. A., & Kanazawa, A. (2021). Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. arXiv preprint arXiv:2101.08779.