Simultaneous Learning of Moving and Active Perceptual Policies for Autonomous Robot

Wataru Hatanaka, Fumihiro Sasaki, Ryota Yamashina, Atsuo Kawaguchi

RICOH

Motivation

Humans/animals can move their bodies, heads, and eyes actively to perceive the state of the environment they are surrounded by, autonomous robots should also do that.
However:

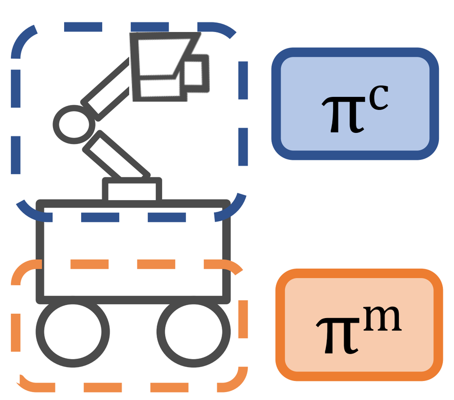
- Optimizing perceptual behaviors is not explicitly treated as a problem in a common setting, what the robot learns to perceive depends on the environment or task.
- Specifying what the robot should perceive every time according to the environment or task is not scalable.

Contribution

- A novel approach for improving the task achievement of a robot by integrating motion and perceptual planning.
- A novel policy update technique using a meta-evaluation makes autonomous robots optimize moving and perceptual policies simultaneously without rewards from an environment.

Problem setup

We assume that two policies exist in one robot, and each policy takes the same camera image as input.



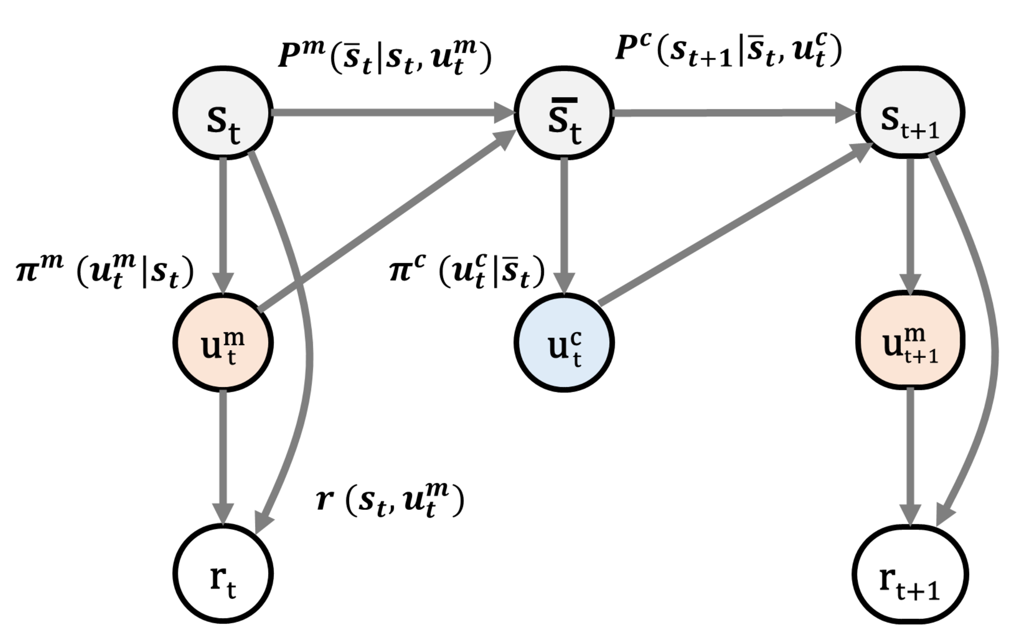
Perceptual policy that decides where to face a camera mounted on the robot.

Moving policy that controls the robot's movement.

Factorizing MDP

The state transition reflects a realistic environment:

- The robot's movements affect which direction the camera faces whereas the camera's movements do not affect the robot's movements.
- Only π^m gets a reward from an environment.



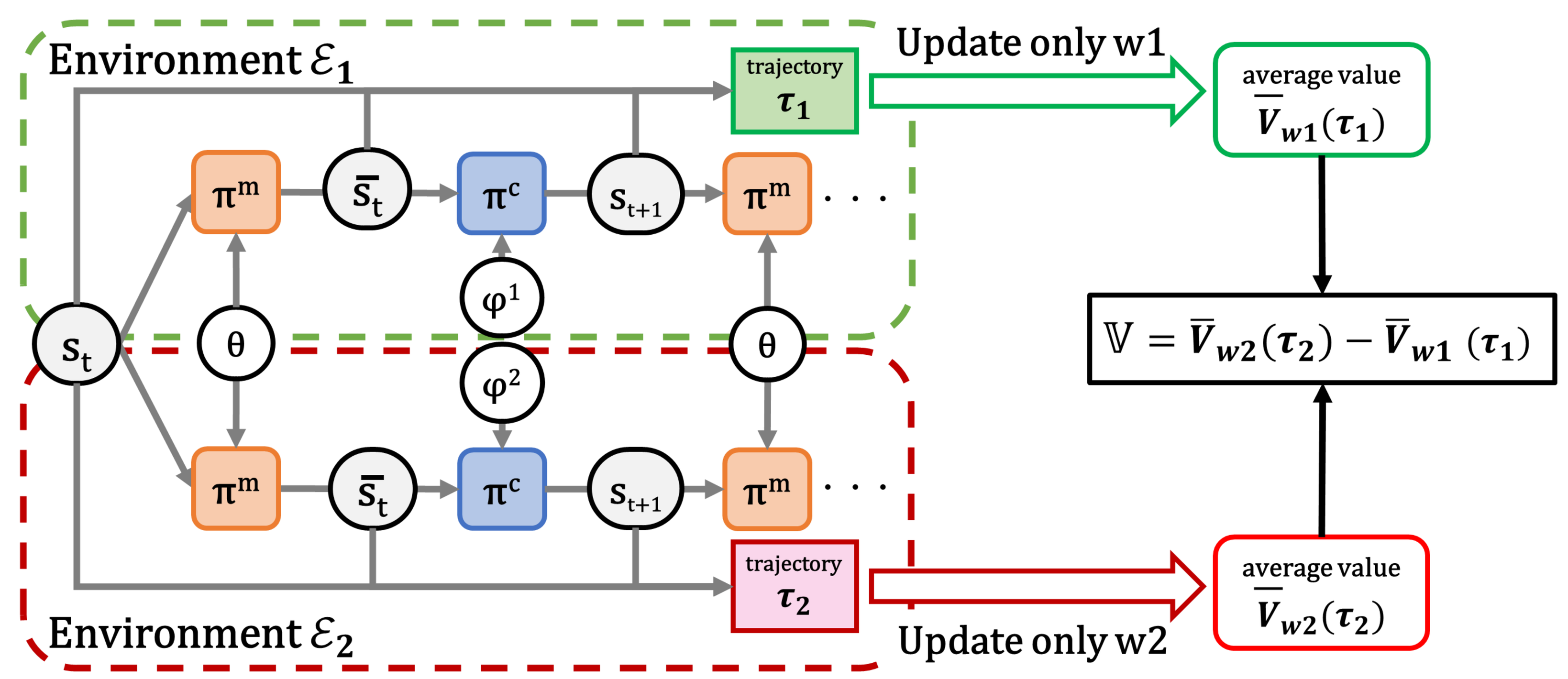
Meta-evaluation and optimization of policies

Preliminary

$\mathcal{E}_1, \mathcal{E}_2$: two same environments.
π_θ^m : the moving policy.
$\pi_{\phi_1}^c, \pi_{\phi_2}^c$: the perceptual policies in $\mathcal{E}_1$ and $\mathcal{E}_2$.
V_{w1}, V_{w2} : the value functions for π_θ^m in $\mathcal{E}_1$ and $\mathcal{E}_2$.

Evaluating a contribution of the perceptual policy

The meta-evaluator $\mathbb{V}$ quantifies the contribution of perceptual policies for a task achievement by comparing the values V_{w1} and V_{w2} in each environment.



Update rule of policies and value functions

$\pi_{\phi_2}^c$: REINFORCE with a cumulative reward replaced by the meta-evaluator $\mathbb{V}$.

$$\nabla_{\phi 2} J(\pi_{\phi 2}^c) = \mathbb{E}_{(\bar{s}_t, u_t^c) \in \tau_2}[\nabla_{\phi 2} \log \pi_{\phi 2}^c(u_t^c|\bar{s}_t)\mathbb{V}]$$

$\pi_{\phi_1}^c$: soft-update rule with updated $\pi_{\phi_2}^c$.

$$\phi_1 = \alpha\phi_2 + (1-\alpha)\phi_1$$

π_θ^m, V_{w1} : A2C [1] using a trajectory by a rollout of π_θ^m with updated $\pi_{\phi_1}^c$ in $\mathcal{E}_1$.

$$\nabla_\theta J(\pi_\theta^m) = \mathbb{E}_{(s_t, u_t^m) \in \tau'_1}[\nabla_\theta \log \pi_\theta^m(u_t^m|s_t) A_{w1}(s_t, u_t^m)]$$

$$\mathcal{L}(w_1) = (r(s_t, u_t^m) + \gamma V_{w1}(s_{t+1}) - V_{w1}(s_t))^2$$

V_{w2}: copy updated V_{w1}.

Empirical: controlling perceptual behaviors

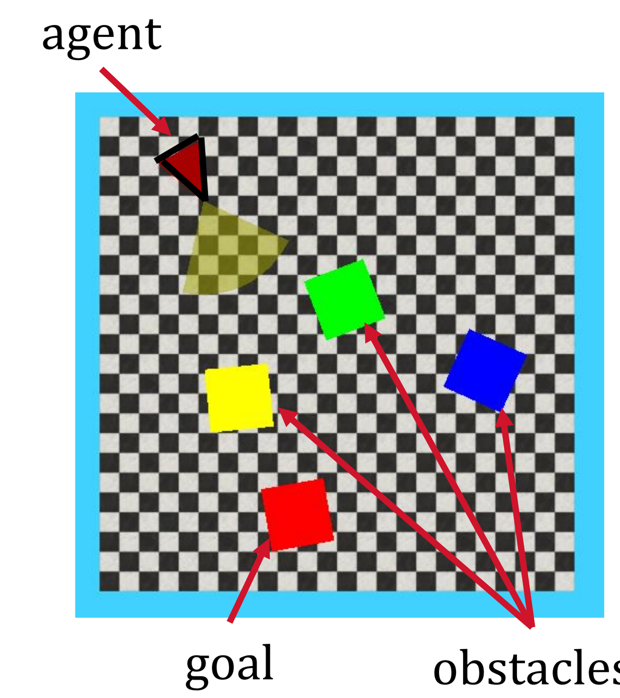
We found that ε-greedy exploration is beneficial to learn better perceptual policy $\pi_{\phi_2}^c$ rather than an entropy regularization of A2C.

$$u_t^{c2} = \begin{cases} \text{argmax}\, \pi_{\phi 2}^c(\bar{s}_t) & \text{with probability } 1-\epsilon \\ \text{a random action}(\bar{s}_t) & \text{with probability } \epsilon \end{cases}$$

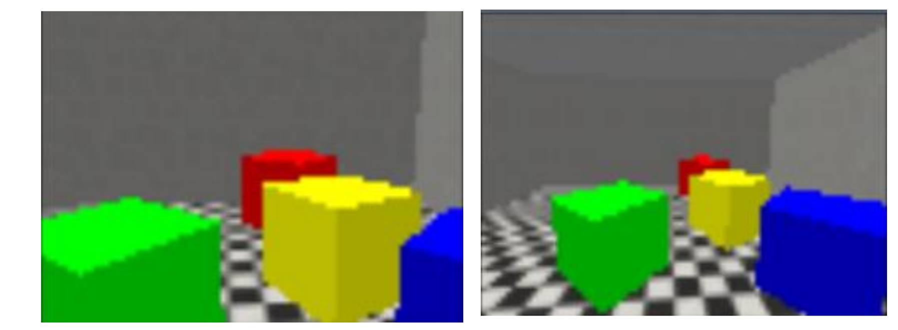
Experiments in partially observable environment

Settings

Map
Single room with three obstacles.

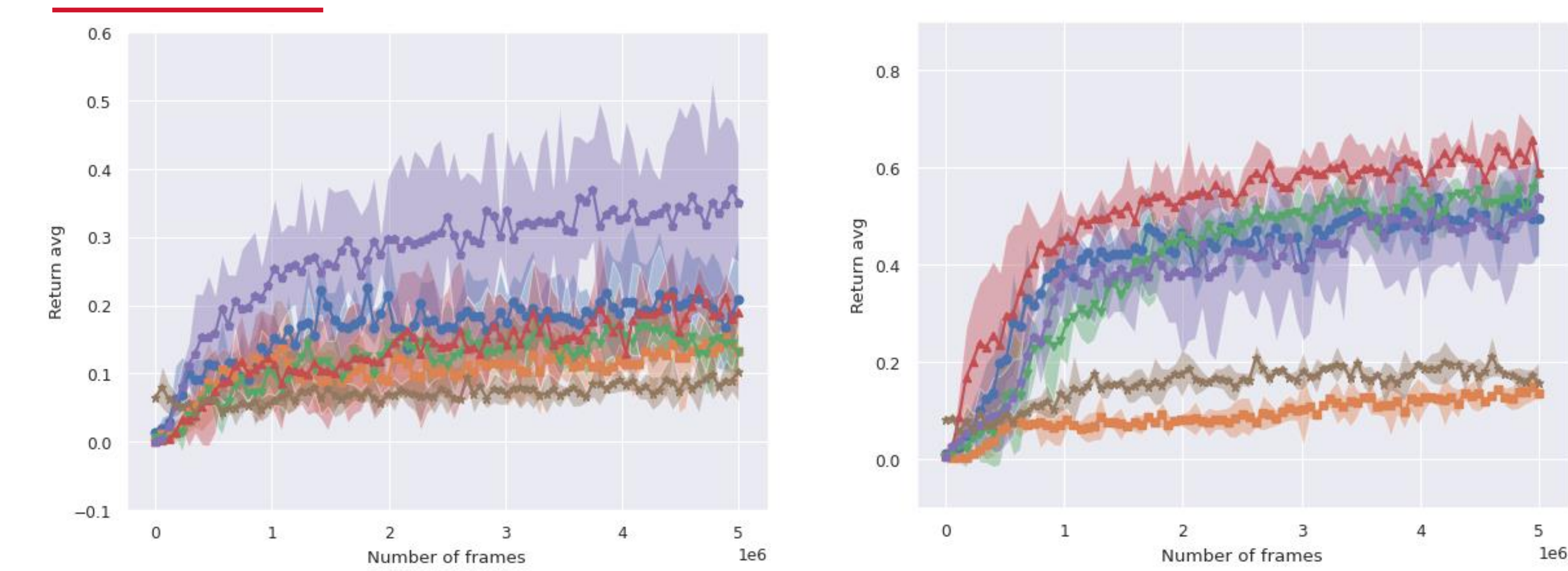


Observations
Two field-of-views: narrow and wide.



Actions
u^m : move-forward, turn-left/right
u^c : look-forward/left/right
Each agent rotates 15 degrees.

Results



Narrow fov   Wide fov

- **Ours(ε=0.1)**
- **Ours(ε=0.3)**
- Fixed_Forward: The camera is fixed in the forward direction.
- Joint: A single agent has a joint action $U^m \times U^c$.
- IA2C: π^m and π^c are trained separately by A2C.
- Curriculum[2]: Joint agent with pre-trained without any obstacles.

Conclusion

- The meta-evaluation process successfully allows the perceptual policy to acquire observations that are favorable to the moving policy for task achievement.
- ε-greedy exploration of perceptual behavior leads to intuitive results for us.

References

[1] Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning." *International conference on machine learning*. PMLR, 2016.

[2] Cheng, Ricson, Arpit Agarwal, and Katerina Fragkiadaki. "Reinforcement learning of active vision for manipulating objects under occlusions." *Conference on Robot Learning*. PMLR, 2018.