Learning to Guide Human Attention on Mobile Telepresence Robots with 360° Vision

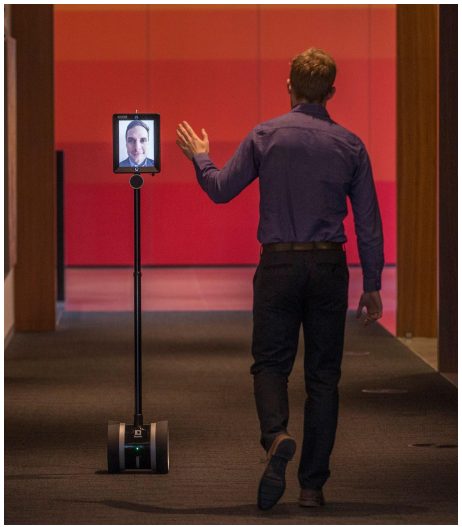
Kishan Chandan, Jack Albertson, Xiaohan Zhang, Xiaoyang Zhang, Yao Liu, Shiqi Zhang

SUNY Binghamton, Binghamton, NY 13902 USA

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021

Mobile Telepresence Robots (MTR)

Telepresence is an illusion of spatial presence, at a place other than the true location. [Minsky 1980]



Double Robotics 2016

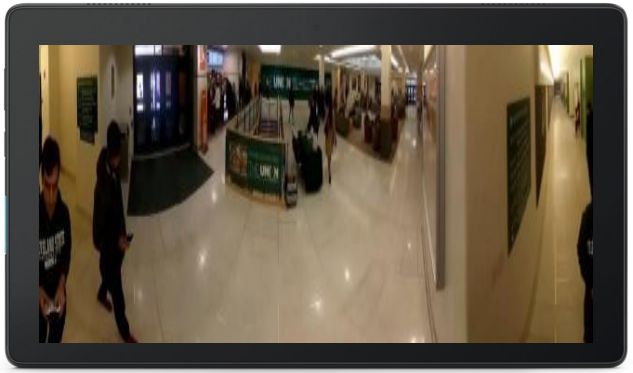
Robocup Rescue 2018

360-degree Vision

- Goal
  - To perceive the entire sphere of the remote environment
- Visualization
  - Mobile devices
  - Head mounted displays (HMDs)

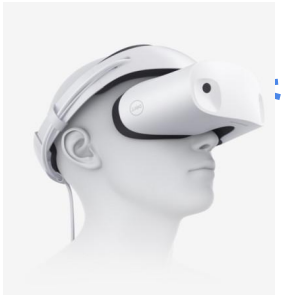
Problems in Visualizing 360-degree Videos

- Mobile devices
  - Difficulty in perceiving frame due to the resolution



Problems in Visualizing 360-degree Videos

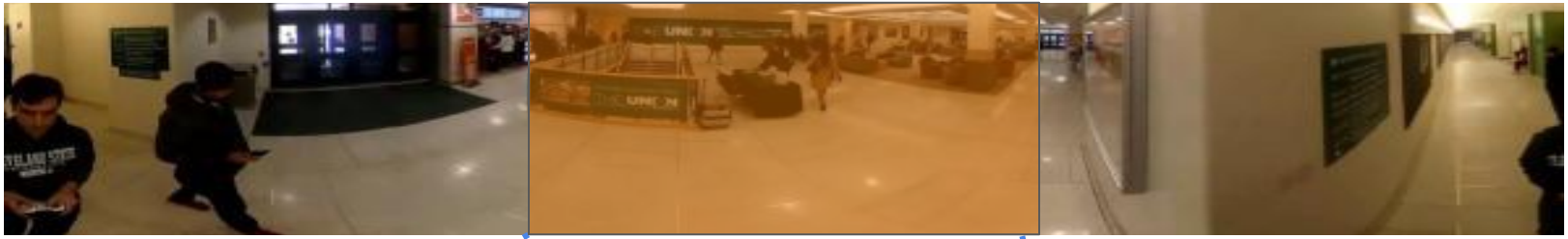
- Mobile devices
  - Difficulty in perceiving frame due to the resolution
- Head mounted displays (HMDs)
  - Can visualize only a part of remote environment

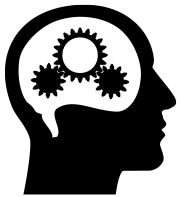


Problems in Visualizing 360-degree Videos

- Humans have a peripheral vision of 120°

Peripheral vision 120°





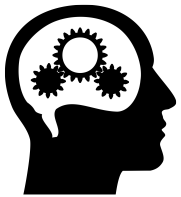
[1] T. C. Ruch and J. F. Fulton, “Medical physiology and biophysics,”Academic Medicine, vol. 35, no. 11, 1960

Problems in Visualizing 360-degree Videos

- Out of 360°, about 240° is out of the field of view of the human

Outside the field of view

Outside the field of view

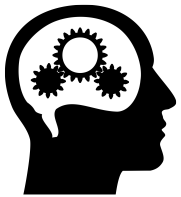


Problems in Visualizing 360-degree Videos

- Human can miss the areas with potentially rich visual information

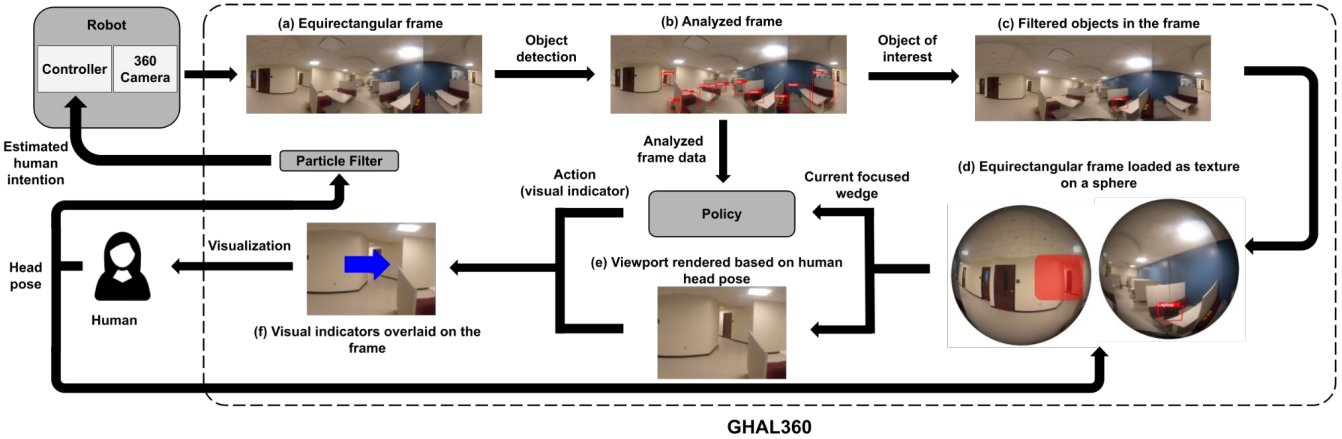
Outside the field of view

Outside the field of view



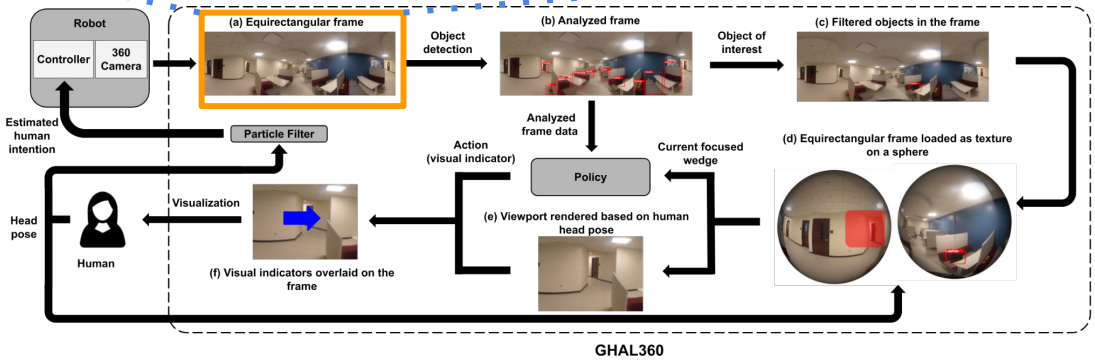
GHAL360 Framework

- Guiding Human Attention with Learning in 360° vision (GHAL360)
  - Bridge human-MTR observability gap
  - Scene analysis using 360-degree videos
  - Learning based guiding of human attention
  - Particle filter to estimated human motion intention



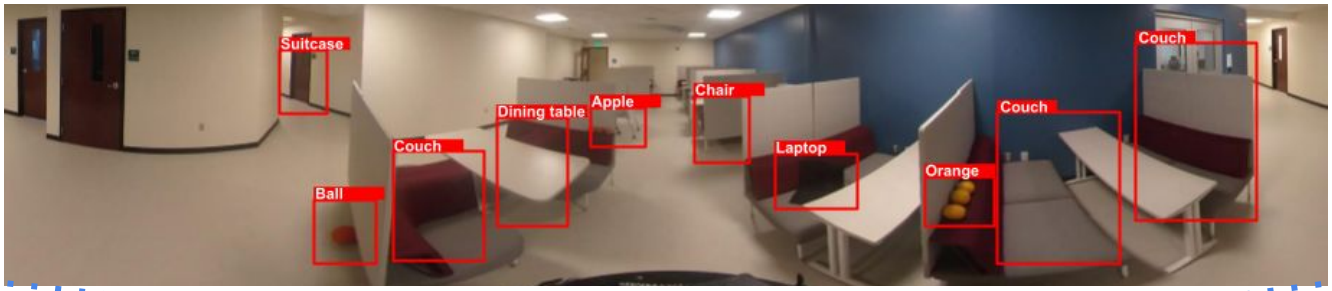
GHAL360 Framework

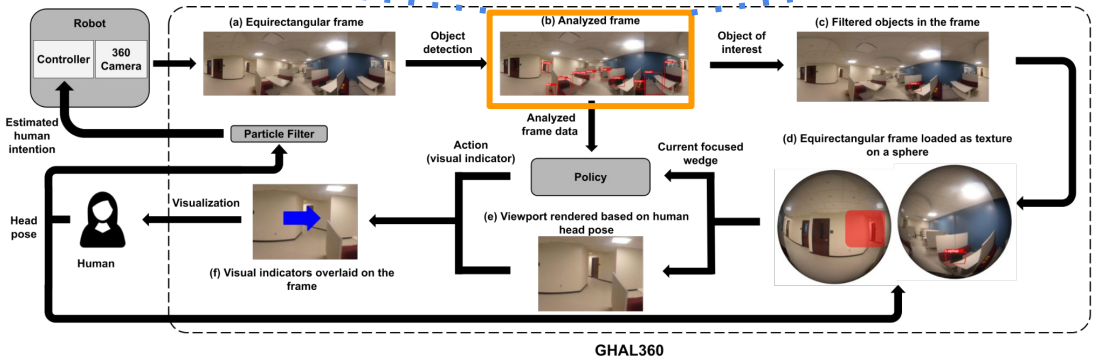
(a) Equirectangular frame



GHAL360 Framework

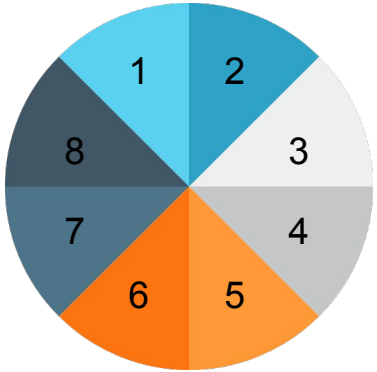
(b) Analyzed frame





GHAL360 - Scene Analysis

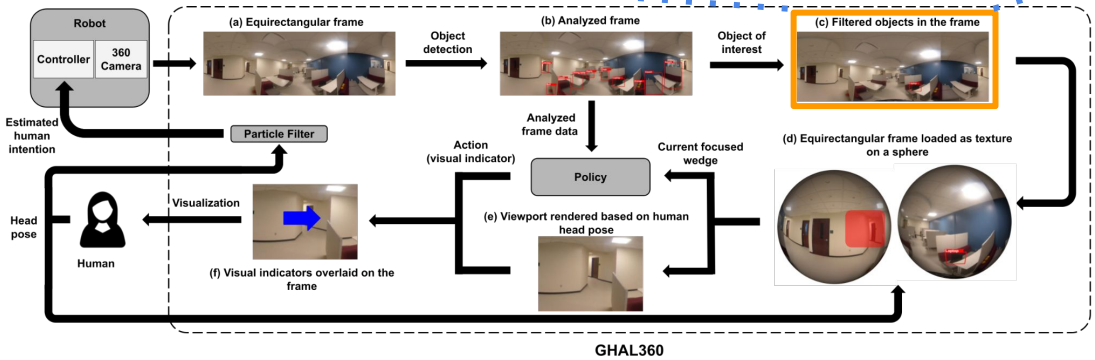
- CNN-based scene analysis using **YOLOv3**
- The 360° sphere is divided into eight wedges. Each wedge is in one of the following four states:
  - No objects detected
  - With clutter
  - Containing only the object of interest
  - Containing clutter as well as the object of interest
- Egocentric state representation to mark the locations of objects with respect to the wedge human is currently focusing on (W_0)



Eight wedges

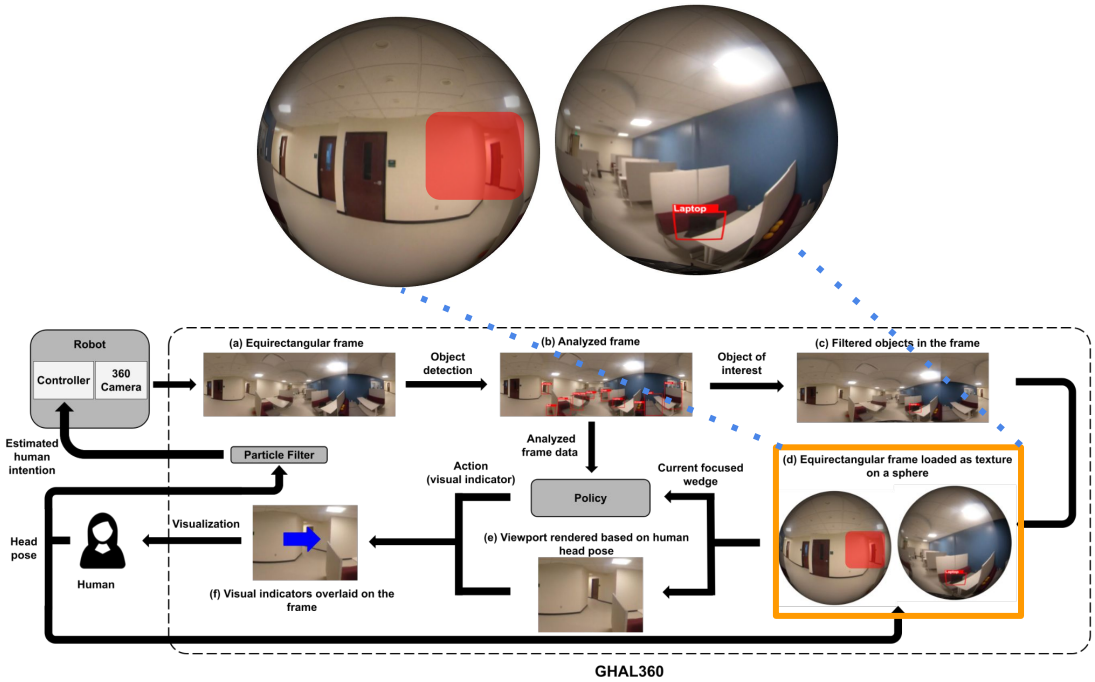
GHAL360 Framework

(c) Filtered objects in the frame



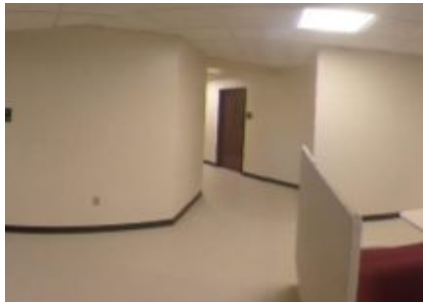
GHAL360 Framework

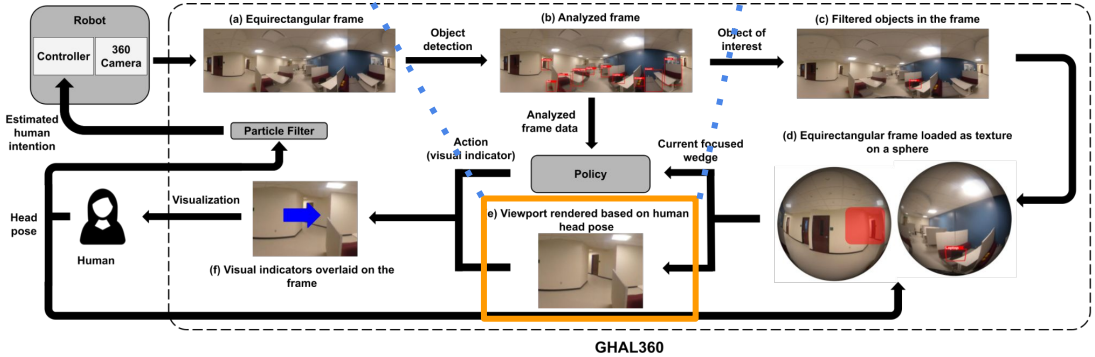
(d) Equirectangular frame loaded as texture of frame on a sphere



GHAL360 Framework

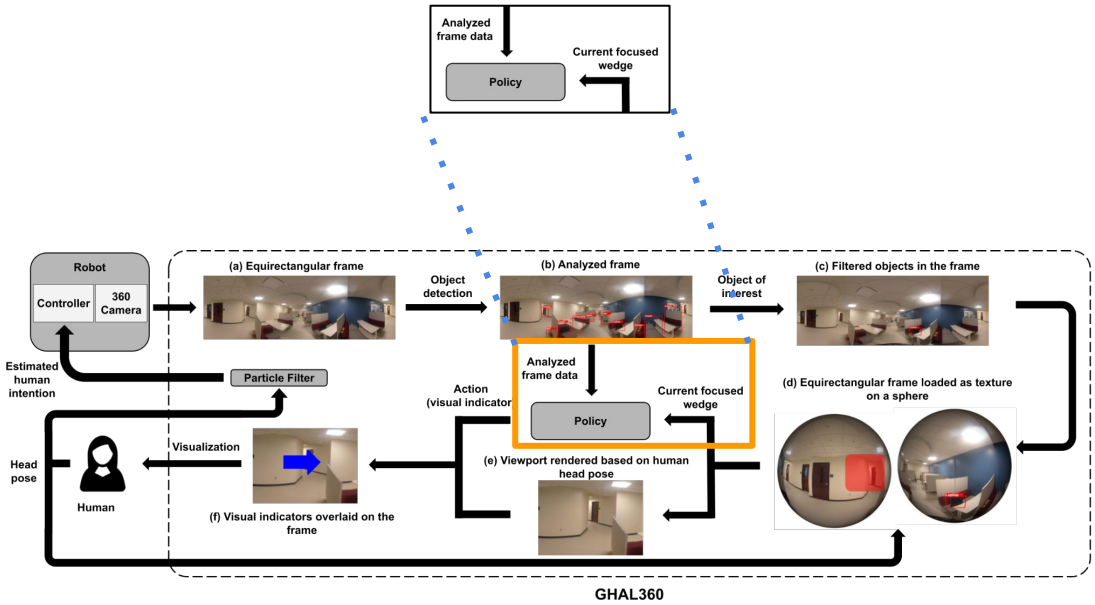
(e) Viewport rendered based on human head pose





GHAL360 Framework

Reinforcement learning-based guidance strategy (Policy)



GHAL360 - Guidance Policy Learning

- Reinforcement Learning (Q-Learning) to learn a policy for guiding human attention
- State space:
  - **S**: $W_0 \times W_1 \times \ldots \times W_{N-1}$, where W_i represents state of a wedge
  - N = 8, so $|S| = 4^8 = 65536$ states
- Action space:
  - **A**: {***left, right, confirm***}
  - ***Left*** and ***right*** actions are guidance indicators, while the ***confirm*** action checks if the object is in the current wedge

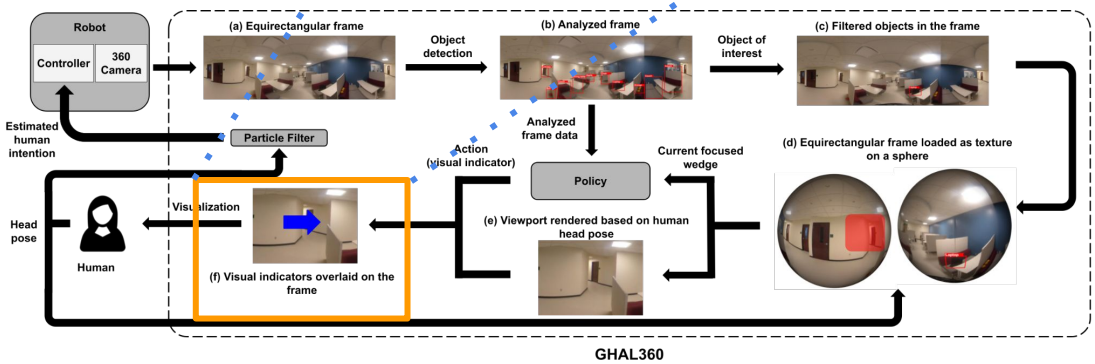
GHAL360 - Guidance Policy Learning

- Based on *s*, the policy generated from the RL approach π returns an action (*a*), which can be a guidance indicator to enable the human to find the object of interest in the remote environment
- Once the agent performs the action, it observes the immediate reward (*r*) and the next state (*s'*)
- Using the tuple (*s*, *a*, *r*, *s'*), the agent updates the Q-value for the state-action pair
- We use the BURLAP library[1] for the implementation of our reinforcement learning agent and the domain

[1]http://burlap.cs.brown.edu/

GHAL360 Framework

(f) Visual indicators overlaid on the frame



GHAL360- Human Motion Intent Estimation

- Particle filter-based state estimator to identify the the human operator's motion intention
- Each particle represents represents a state, which in our case is one of the eight wedges
- Particle filter initialization (M particles):
  - $X = \{x^0, x^1, \dots, x^{M-1}\}$
  - Weight of particles: $\omega^i = 1/M$
- Belief state of the particle filter:
  - $B(X_t) = P(X_t \mid e_{1:t})$

GHAL360- Human Motion Intent Estimation

- Evidence to the particle filter is the current change in human head orientation (left or right), and current focused wedge
- Belief updation:
  - $B'(X_{t+1}) = \sum_{xt} P(X'|x_t)\, B(x_t))$
- Threshold density of particles: 70%
- Particle filter outs the predicted human intention once the density is above the threshold
- Controller sends a control signal to the robot based on the predicted human intention to move the robot in the remote environment

Telepresence Interface

Experiments

- Target search scenario
- Segway-based mobile robot platform (RMP 110)
- 360-degree camera (Ricoh Theta V)
- 2 Matterport3D[1] environment (Home and office), and 1 Real-world office environment as dataset

(a) Front

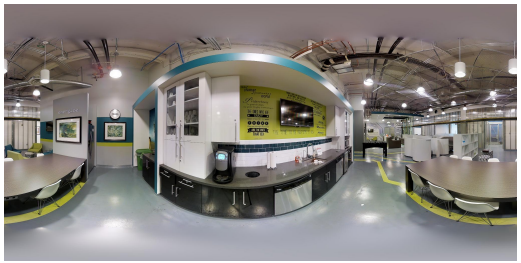
(b) Back

[1] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva,S. Song, A. Zeng, and Y. Zhang, “Matterport3d: Learning from rgb-d data in indoor environments,”International Conference on 3D Vision (3DV), 2017

Experiments

- 2 Matterport3D[1] environment (a,b,c,d)
- Real-world office environment (e,f)

(a)



(c)

(e)

(b)

(d)

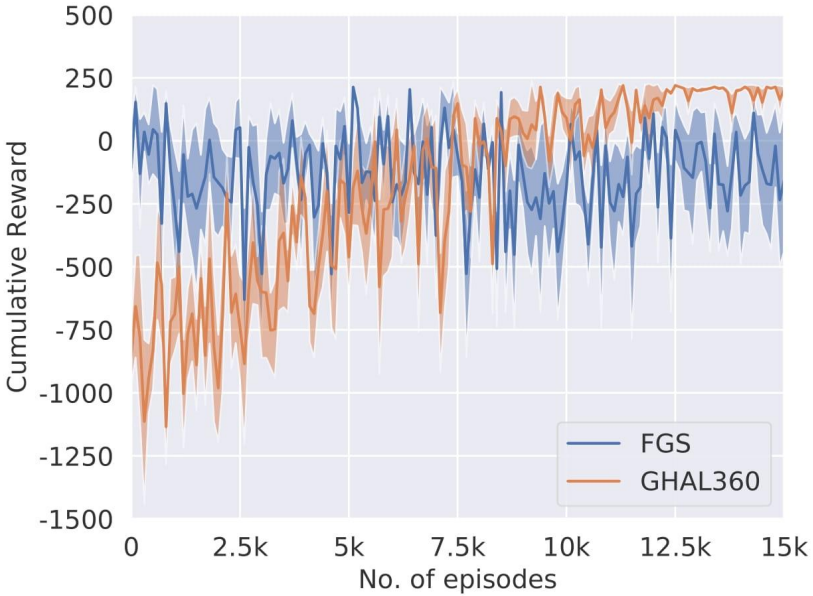
(f)

Baselines

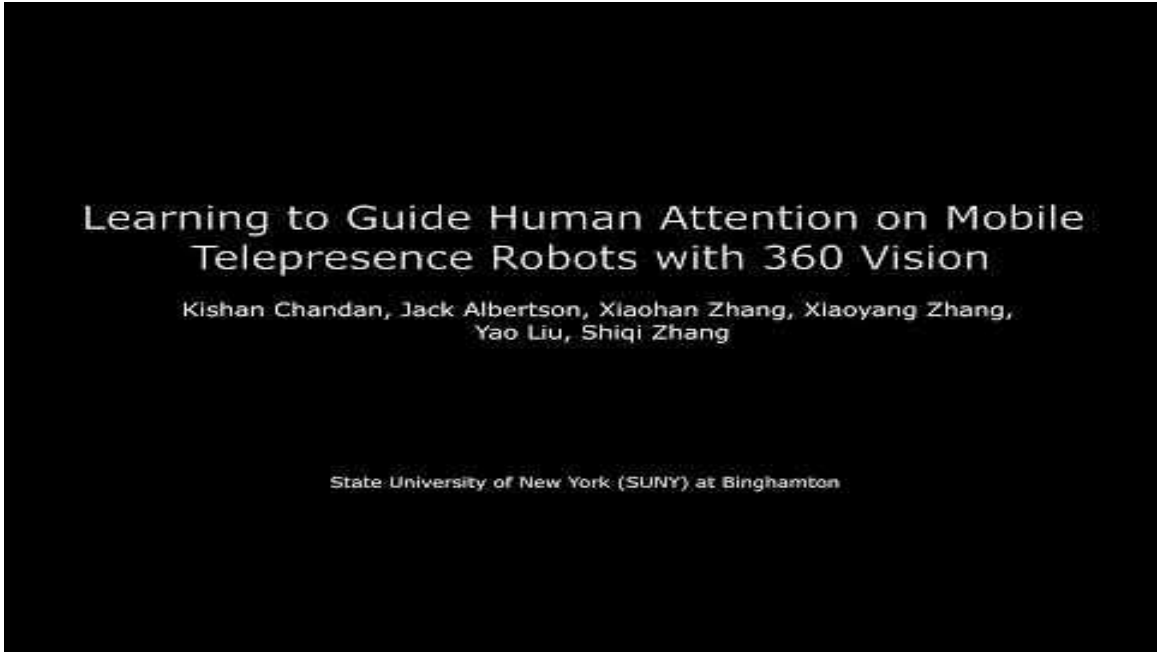
- Monocular Fixed Orientation (MFO) - 120° FOV of remote environment
- All-Degree View (ADV): 360° FOV of remote environment
- Fixed Guidance Strategy (FGS): hand-coded guidance strategy
- Reinforcement Learning-based Guidance Strategy (RLGS) - ablation of GHAL360 without the particle filter based human motion intent estimation

GHAL360 - Policy Learning

- Cumulative reward comparison between FGS and GHAL360
- We tested 150 different policies after each batch of 100 episodes (15000 training episodes) in our realistic simulation
- Learned policy performs better than FGS (hand-coded baseline)

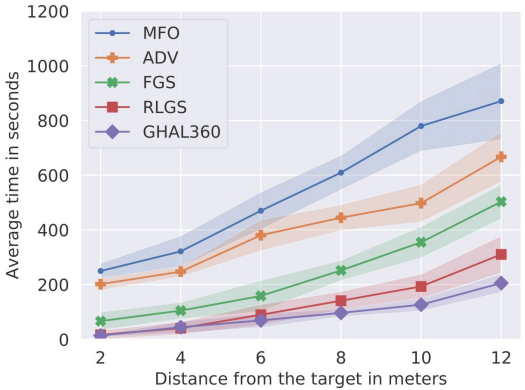


Demonstration

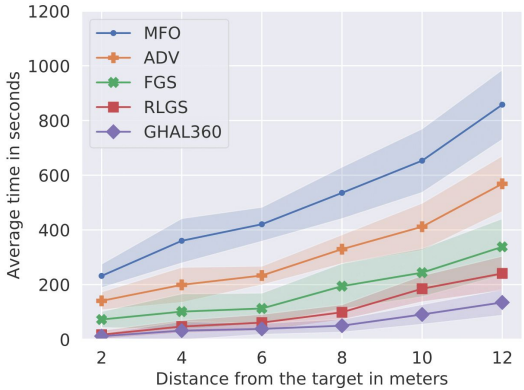


Results

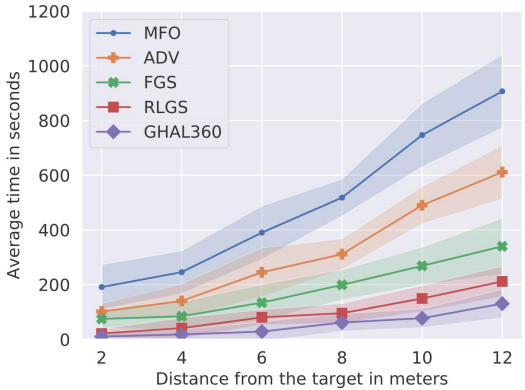
- Comparison of average task completion time at six different distances
- GHAL360 outperforms the baseline at all distances



Env 1: Matterport3D (Home)



Env 2: Matterport3D (Office)



Env 3: Real Robot (Office)

Contributions of our GHAL360 Framework

- A novel framework (called GHAL360) for mobile telepresence robots that enables 360° scene analysis while simultaneously guiding human attention
  - Learning-based guidance strategy
  - Particle filter-based human motion intent estimation
- We evaluated GHAL360 using target search tasks in photo-realistic simulation (two envs from an open-source dataset and one env constructed using a real robot)
- Achieved significant improvements in the efficiency of the human-robot team in comparison to competitive baselines.

Thank you for watching!