Vision-Based Social Robot Navigation with 6D Head Pose Estimation and F-Formation Analysis

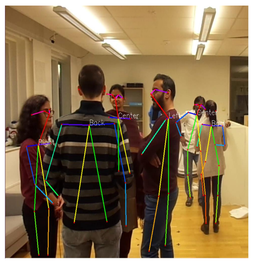
Georgina Woo[1,2], Raj Korpan[2], Ioannis Stamos[1]

[1] Computer Vision Lab, Department of Computer Science, Hunter College

[2] TIER Lab, Department of Computer Science, Hunter College

Robots navigating human spaces must understand not just obstacles, but **social groupings and situations**, inferred from **visual attention cues**.

How can robots recognize the **structure of social formations** using vision to navigate safely and naturally?



Some existing approaches assume perception from human eye level (left)

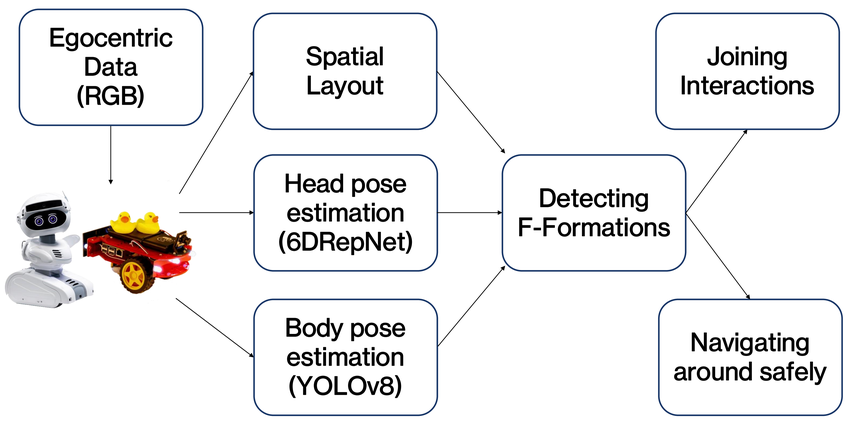
Smaller robots, like the Misty or Duckiebot, perceive from a lower perspective (below)

Misty's POV

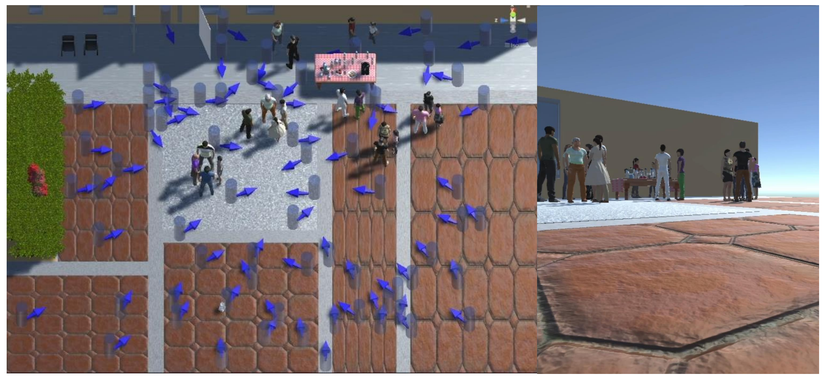
Duckiebot's POV

This makes visual reasoning more challenging due to more frequent distortions and occlusions.

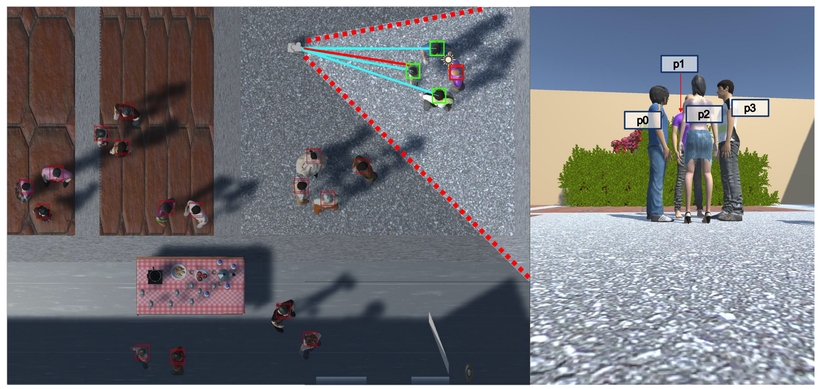
Proposed Framework



Automated dataset generation



At each of the 62 positions marked by blue arrows, Misty makes 20 forward steps and takes a picture from its point of view.



A frustum representing the robot's **field of view** is projected from the camera to identify and **record visible groupings**.
Raycasting rules out **occluded** heads, and records **Euler angles** of head poses in relation to the camera.

Extracting spatial information

Using the spatial information from the bounding boxes of each identified person, we can estimate groups based on **relative distance** and **height thresholds**. These groups can be represented as a graph, where spatial relationships are treated as **transitive links** between individuals.

Extracting pose information

Visualizing head pose estimation with 6DRepNet

Visualizing full body pose estimation with YOLOPosev8

Results

Evaluated on the Simulation dataset (1240 labeled images) and the Real dataset (143 labeled images)

Dataset	Precision	Recall	F1 Score
Before including pose information			
Simulation	78.40%	77.50%	77.90%
After including pose information			
Simulation	90.85%	97.00%	93.82%
Real	96.79%	89.35%	92.92%
Combined	93.82%	93.17%	93.37%

Acknowledgement

SK Pathi – Machine Perception and Interaction Lab, Örebro University
Hunter College Computer Science Department
This work was partially supported by the John P. McNulty Scholars Program for Leadership in Science and Math at Hunter College.

Conference materials available via QR code