Researchers create new model that aims to give robots human-like perception of their physical environments

July 15 2020

MIT researchers have developed a representation of spatial perception for robots that is modeled after the way humans perceive and navigate the world. The key component of the team's new model is Kimera, an open-source library that the team previously developed to simultaneously construct a 3D geometric model of an environment. Kimera builds a dense 3D semantic mesh of an environment and can track humans in the environment. The figure shows a multi-frame action

sequence of a human moving in the scene. Credit: Image courtesy of the researchers

Wouldn't we all appreciate a little help around the house, especially if that help came in the form of a smart, adaptable, uncomplaining robot? Sure, there are the one-trick Roombas of the appliance world. But MIT engineers are envisioning robots more like home helpers, able to follow high-level, Alexa-type commands, such as "Go to the kitchen and fetch me a coffee cup."

To carry out such high-level tasks, researchers believe robots will have to be able to perceive their physical environment as humans do.

"In order to make any decision in the world, you need to have a mental model of the environment around you," says Luca Carlone, assistant professor of aeronautics and astronautics at MIT. "This is something so effortless for humans.

But for robots it's a painfully hard problem, where it's about transforming pixel values that they see through a camera, into an understanding of the world."Now Carlone and his students have developed a representation of spatial perception for robots that is modeled after the way humans perceive and navigate the world.

The new model, which they call 3-D Dynamic Scene Graphs, enables a robot to quickly generate a 3-D map of its surroundings that also includes objects and their semantic labels (a chair versus a table, for instance), as well as people, rooms, walls, and other structures that the robot is likely seeing in its environment.

The model also allows the robot to extract relevant information from the

3-D map, to query the location of objects and rooms, or the movement of people in its path.

"This compressed representation of the environment is useful because it allows our robot to quickly make decisions and plan its path," Carlone says. "This is not too far from what we do as humans. If you need to plan a path from your home to MIT, you don't plan every single position you need to take. You just think at the level of streets and landmarks, which helps you plan your route faster."

Beyond domestic helpers, Carlone says robots that adopt this new kind of mental model of the environment may also be suited for other high-level jobs, such as working side by side with people on a factory floor or exploring a disaster site for survivors.

He and his students, including lead author and MIT graduate student Antoni Rosinol, will present their findings this week at the Robotics: Science and Systems virtual conference.

A mapping mix

At the moment, robotic vision and navigation has advanced mainly along two routes: 3-D mapping that enables robots to reconstruct their environment in three dimensions as they explore in real time; and semantic segmentation, which helps a robot classify features in its environment as semantic objects, such as a car versus a bicycle, which so far is mostly done on 2-D images.

Carlone and Rosinol's new model of spatial perception is the first to generate a 3-D map of the environment in real-time, while also labeling objects, people (which are dynamic, contrary to objects), and structures within that 3-D map.

The key component of the team's new model is Kimera, an open-source library that the team previously developed to simultaneously construct a 3-D geometric model of an environment, while encoding the likelihood that an object is, say, a chair versus a desk.

"Like the mythical creature that is a mix of different animals, we wanted Kimera to be a mix of mapping and semantic understanding in 3-D," Carlone says.

Kimera works by taking in streams of images from a robot's camera, as well as inertial measurements from onboard sensors, to estimate the trajectory of the robot or camera and to reconstruct the scene as a 3-D mesh, all in real-time.

To generate a semantic 3-D mesh, Kimera uses an existing neural network trained on millions of real-world images, to predict the label of each pixel, and then projects these labels in 3-D using a technique known as ray-casting, commonly used in computer graphics for real-time rendering.

The result is a map of a robot's environment that resembles a dense, three-dimensional mesh, where each face is color-coded as part of the objects, structures, and people within the environment.

A layered scene

If a robot were to rely on this mesh alone to navigate through its environment, it would be a computationally expensive and time-consuming task. So the researchers built off Kimera, developing algorithms to construct 3-D dynamic "scene graphs" from Kimera's initial, highly dense, 3-D semantic mesh.

Scene graphs are popular computer graphics models that manipulate and

render complex scenes, and are typically used in video game engines to represent 3-D environments.

In the case of the 3-D dynamic scene graphs, the associated algorithms abstract, or break down, Kimera's detailed 3-D semantic mesh into distinct semantic layers, such that a robot can "see" a scene through a particular layer, or lens. The layers progress in hierarchy from objects and people, to open spaces and structures such as walls and ceilings, to rooms, corridors, and halls, and finally whole buildings.

Carlone says this layered representation avoids a robot having to make sense of billions of points and faces in the original 3-D mesh.

Within the layer of objects and people, the researchers have also been able to develop algorithms that track the movement and the shape of humans in the environment in real time.

The team tested their new model in a photo-realistic simulator, developed in collaboration with MIT Lincoln Laboratory, that simulates a robot navigating through a dynamic office environment filled with people moving around.

"We are essentially enabling robots to have mental models similar to the ones humans use," Carlone says. "This can impact many applications, including self-driving cars, search and rescue, collaborative manufacturing, and domestic robotics.

Another domain is virtual and augmented reality (AR). Imagine wearing AR goggles that run our algorithm: The goggles would be able to assist you with queries such as 'Where did I leave my red mug?' and 'What is the closest exit?'

You can think about it as an Alexa which is aware of the environment

around you and understands objects, humans, and their relations."

"Our approach has just been made possible thanks to recent advances in deep learning and decades of research on simultaneous localization and mapping," Rosinol says. "With this work, we are making the leap toward a new era of robotic perception called spatial-AI, which is just in its infancy but has great potential in robotics and large-scale virtual and augmented reality."

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/*), a popular site that covers news about MIT research, innovation and teaching.*

Provided by Massachusetts Institute of Technology