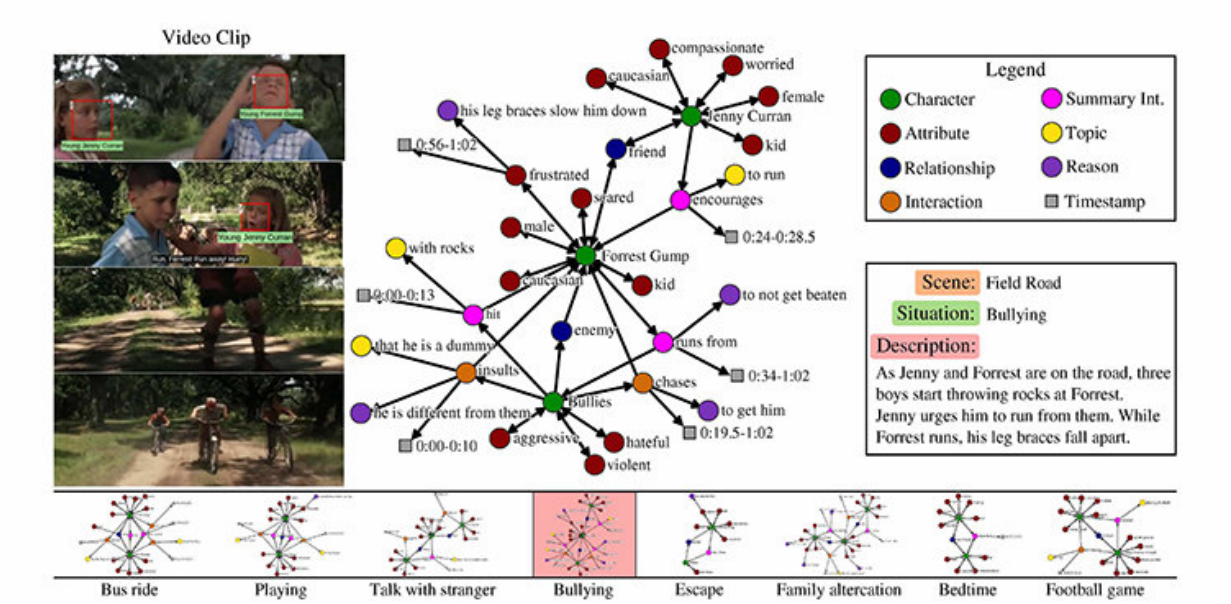
How to train your robot: Research provides new approaches

June 25 2018, by Nina Haikara



An example from the MovieGraphs dataset, scene from the film Forrest Gump. Credit: University of Toronto

If your friend is sad, you can say something to help cheer them up. If you ask your co-worker to make coffee, they know the steps to complete this task.

But how do artificially intelligent robots, or AIs, learn to behave in the same way humans do?

University of Toronto researchers are presenting new approaches towards socially intelligent AIs, at the Computer Vision and Pattern Recognition (CVPR) conference, the premier annual computer vision event this week in Salt Lake City, Utah.

How do we train a robot how to behave?

In their paper MovieGraphs: Towards Understanding Human-Centric Situations from Videos, Paul Vicol, a Ph.D. student in computer science, Makarand Tapaswi, a post-doctoral researcher, Lluis Castrejón, a master's graduate of U of T computer science who is now a Ph.D. student at the University of Montreal Institute for Learning Algorithms, and Sanja Fidler, an assistant professor at U of T Mississauga's department of mathematical and computational sciences and tri-campus graduate department of computer science, have amassed a dataset of annotated video clips from more than 50 films.

"MovieGraphs is a step towards the next generation of cognitive agents that can reason about how people feel and about the motivations for their behaviours," says Vicol. "Our goal is to enable machines to behave appropriately in social situations. Our graphs capture a lot of high-level properties of human situations that haven't been explored in prior work."

Their dataset focuses on films in the drama, romance, and comedy genres, like Forrest Gump and Titanic, and follows characters over time. They don't include superhero films like Thor because they're not very representative of the human experience.

"The idea was to use movies as a proxy for the real world," says Vicol.

Each clip, he says, is associated with a graph that captures rich detail about what's happening in the clip: which characters are present, their relationships, interactions between each other along with the reasons for

why they're interacting, and their emotions.

Vicol explains that the dataset shows, for example, not only that two people are arguing, but what they're arguing about, and the reasons why they're arguing, which come from both visual cues and dialogue. The team created their own tool for enabling annotation, which was done by a single annotator for each film.

"All the clips in a movie are annotated consecutively, and the entire graph associated with each clip is created by one person, which gives us coherent structure in each graph, and between graphs over time," he says.

With their dataset of more than 7,500 clips, the researchers introduce three tasks, explains Vicol. The first is video retrieval, based on the fact that the graphs are grounded in the videos.

"So if you search by using a graph that says Forrest Gump is arguing with someone else, and that the emotions of the characters are sad and angry, then you can find the clip," he says.

The second is interaction ordering, which refers to determining the most plausible order of character interactions. For example, he explains if a character were to give another character a present, the person receiving the gift would say "thank you."

"You wouldn't usually say 'thank you,' and then receive a present. It's one way to benchmark whether we're capturing the semantics of interactions."

Their final task is reason prediction based on the social context.

"If we focus on one interaction, can we determine the motivation behind

that interaction and why it occurred? So that's basically trying to predict when somebody yells at somebody else, the actual phrase that would explain why," he says

Tapaswi says the end goal is to learn behaviour.

"Imagine for example in one clip, the machine basically embodies Jenny [from the film Forrest Gump]. What is an appropriate action for Jenny? In one scene, it's to encourage Forrest to run away from bullies. So we're trying to get machines to learn appropriate behaviour."

"Appropriate in the sense that movies allow, of course."

Screenshot: MIT CSAIL/VirtualHome: Simulating Household Activities via Programs

How does a robot learn household tasks?

Led by led by Massachusetts Institute of Technology Assistant Professor Antonio Torralba and U of T's Fidler, VirtualHome: Simulating Household Activities via Programs, is training a virtual human agent using natural language and a virtual home, so the robot can learn not only through language, but by seeing, explains U of T master's student of computer science Jiaman Li, a contributing author with U of T Ph.D. student of computer science Wilson Tingwu Wang.

Li explains the high-level action may be "work on computer" and the description includes: turning on the computer, sitting in front of it, typing on the keyboard and grabbing the mouse to scroll.

"So if we tell a human this description, 'work on computer,' the human can perform these actions just like the descriptions. But if we just tell robots this description, how do they exactly do that? The robot doesn't have this common sense. It needs very clear steps, or programs."

Because there's no dataset that includes all this knowledge, she says the researchers built one using a web interface to gather the programs, which provide the action name and the description.

"Then we built a simulator so we have a virtual human in a virtual home who can perform these tasks," she says.

For her part in the ongoing project, Li is using deep learning – a branch of machine learning that trains computers to learn – to automatically generate programs from text or video for these programs.

However, it's no easy task to perform each action in the simulator, says

Li, as the dataset resulted in more than 5,000 programs.

"Simulating everything one does in a home is extremely hard, and we make a step towards this by implementing the most frequent atomic actions such as walk, sit, and pick up," says Fidler.

"We hope that our simulator will be used to train robots complex tasks in a virtual environment, before going on to the real world."

MovieGraphs was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and VirtualHome is supported in part by the NSERC COmputing Hardware for Emerging Intelligent Sensing Applications (COHESA) Network.

More information: MovieGraphs: Towards Understanding Human-Centric Situations from Videos, arxiv.org/abs/1712.06761

Provided by University of Toronto