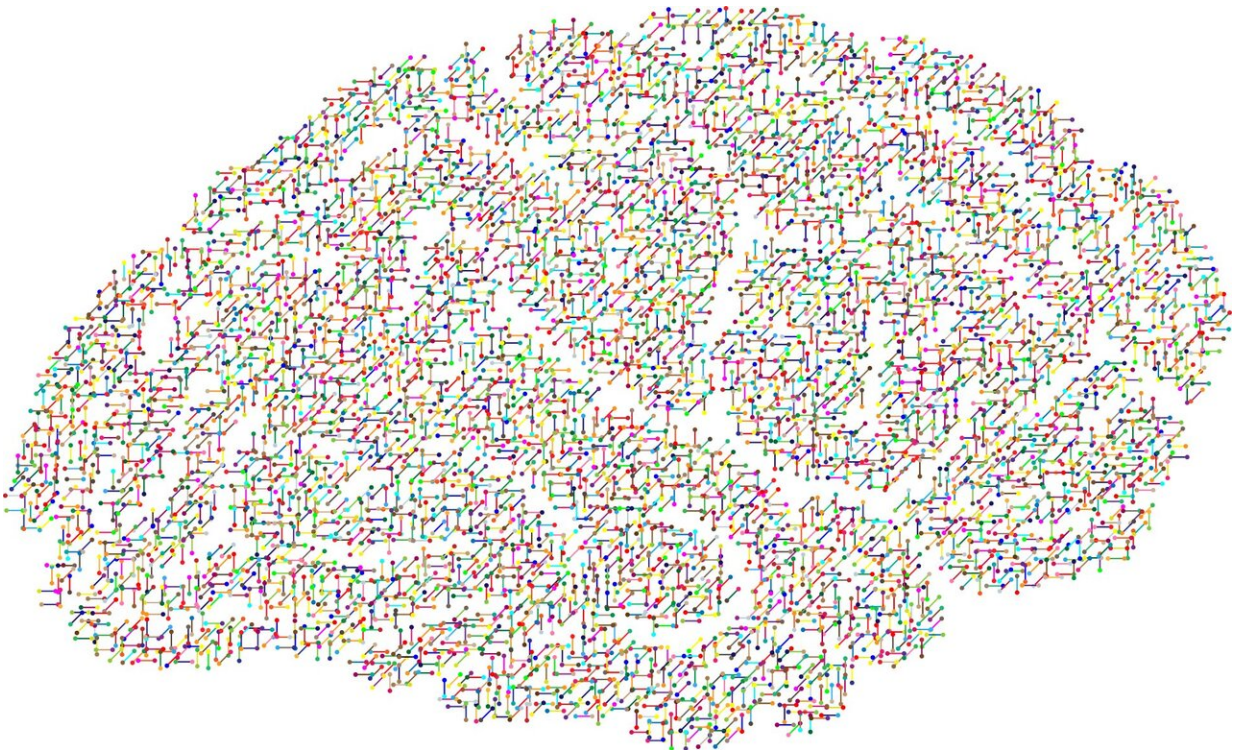# Teaching artificial intelligence to connect senses like vision and touch

June 17 2019, by Rachel Gordon

Credit: CC0 Public Domain

In Canadian author Margaret Atwood's book *The Blind Assassin*, she says that "touch comes before sight, before speech. It's the first language and the last, and it always tells the truth."

While our sense of touch gives us a channel to feel the physical world,

our eyes help us immediately understand the full picture of these tactile signals.

Robots that have been programmed to see or feel can't use these signals quite as interchangeably. To better bridge this sensory gap, researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have come up with a predictive artificial intelligence (AI) that can learn to see by touching, and learn to feel by seeing.

The team's system can create realistic tactile signals from visual inputs, and predict which object and what part is being touched directly from those tactile inputs. They used a KUKA robot arm with a special tactile sensor called GelSight, designed by another group at MIT.

Using a simple web camera, the team recorded nearly 200 objects, such as tools, household products, fabrics, and more, being touched more than 12,000 times. Breaking those 12,000 video clips down into static frames, the team compiled "VisGel," a dataset of more than 3 million visual/tactile-paired images.

"By looking at the scene, our model can imagine the feeling of touching a flat surface or a sharp edge," says Yunzhu Li, CSAIL Ph.D. student and lead author on a new paper about the system. "By blindly touching around, our model can predict the interaction with the environment purely from tactile feelings. Bringing these two senses together could empower the robot and reduce the data we might need for tasks involving manipulating and grasping objects."

Recent work to equip robots with more human-like physical senses, such as MIT's 2016 project using deep learning to [visually indicate sounds,](visually indicate sounds) or a model that predicts objects' responses to physical forces, both use large datasets that aren't available for understanding interactions between vision and touch.

The team's technique gets around this by using the VisGel dataset, and something called generative adversarial networks (GANs).

Yunzhu Li is a PhD student at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). Credit: Massachusetts Institute of Technology

GANs use visual or tactile images to generate images in the other modality. They work by using a "generator" and a "discriminator" that compete with each other, where the generator aims to create real-looking images to fool the discriminator. Every time the discriminator "catches" the generator, it has to expose the internal reasoning for the decision, which allows the generator to repeatedly improve itself.

Vision to touch

Humans can infer how an object feels just by seeing it. To better give machines this power, the system first had to locate the position of the touch, and then deduce information about the shape and feel of the region.

The reference images—without any robot-object interaction—helped the system encode details about the objects and the environment. Then, when the robot arm was operating, the model could simply compare the current frame with its reference image, and easily identify the location and scale of the touch.

This might look something like feeding the system an image of a computer mouse, and then "seeing" the area where the model predicts the object should be touched for pickup—which could vastly help machines plan safer and more efficient actions.

Touch to vision

For touch to vision, the aim was for the model to produce a visual image based on tactile data. The model analyzed a tactile image, and then figured out the shape and material of the contact position. It then looked back to the reference image to "hallucinate" the interaction.

For example, if during testing the model was fed tactile data on a shoe, it could produce an image of where that shoe was most likely to be touched.

This type of ability could be helpful for accomplishing tasks in cases where there's no visual data, like when a light is off, or if a person is blindly reaching into a box or unknown area.

Looking ahead

The current dataset only has examples of interactions in a controlled environment. The team hopes to improve this by collecting data in more unstructured areas, or by using a new MIT-designed tactile glove, to better increase the size and diversity of the dataset.

There are still details that can be tricky to infer from switching modes, like telling the color of an object by just touching it, or telling how soft a sofa is without actually pressing on it. The researchers say this could be improved by creating more robust models for uncertainty, to expand the distribution of possible outcomes.

In the future, this type of model could help with a more harmonious relationship between vision and robotics, especially for object recognition, grasping, better scene understanding, and helping with seamless human-robot integration in an assistive or manufacturing setting.

"This is the first method that can convincingly translate between visual and touch signals," says Andrew Owens, a postdoc at the University of California at Berkeley. "Methods like this have the potential to be very useful for robotics, where you need to answer questions like 'is this object hard or soft?", or 'if I lift this mug by its handle, how good will my grip be?" This is a very challenging problem, since the signals are so different, and this model has demonstrated great capability."

 More information: Connecting Touch and Vision via Cross-Modal Prediction. visgel.csail.mit.edu/

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