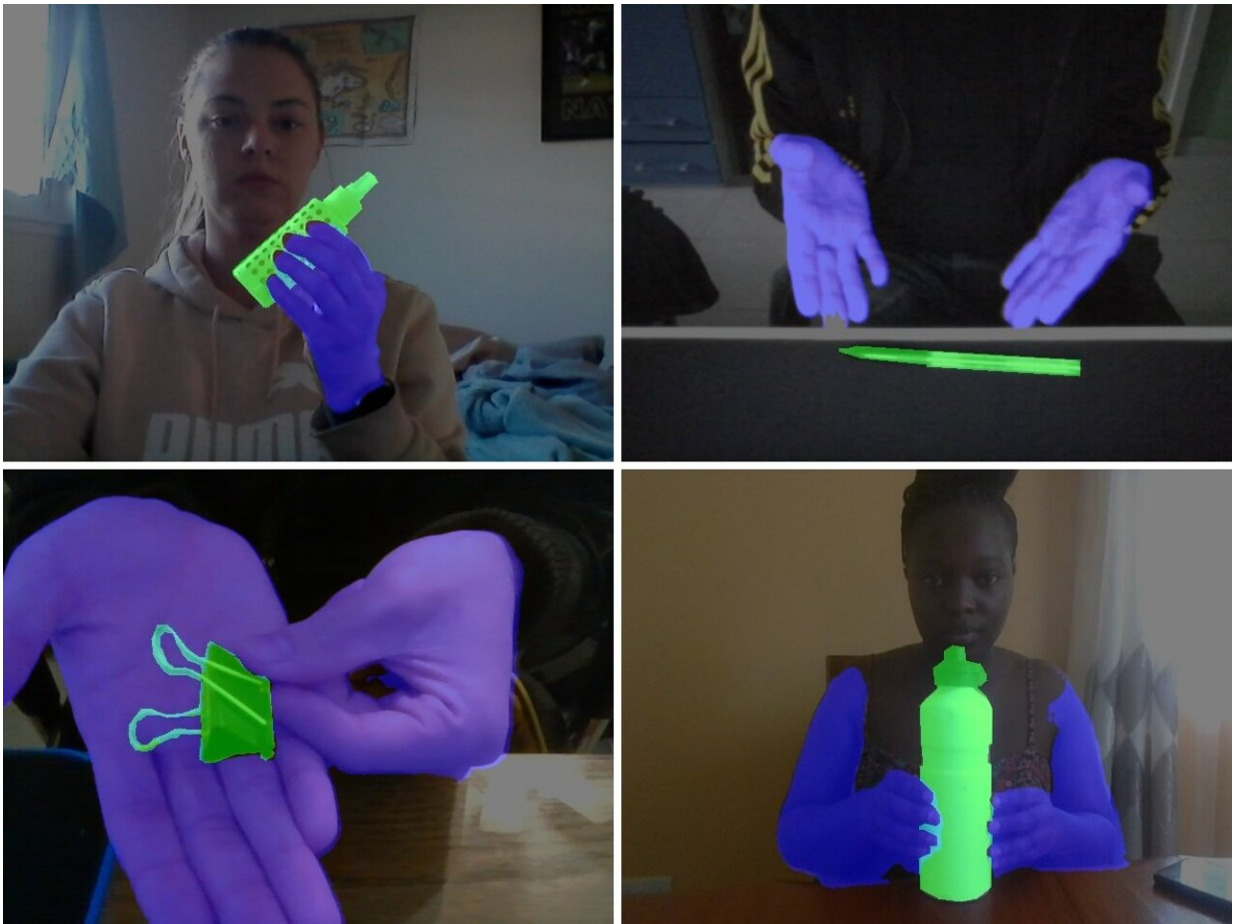# New software allows nonspecialists to intuitively train machines using gestures

October 31 2022

In each image of the HuTics custom data set, the users' hands are visualized in blue and the object in green. HuTics is used to train a machine learning model. Credit: ©2022 Yatani and Zhou

Many computer systems that people interact with on a daily basis require knowledge about certain aspects of the world, or models, to work. These systems have to be trained, often needing to learn how to recognize objects from video or image data. This data frequently contains superfluous content that reduces the accuracy of models. So, researchers found a way to incorporate natural hand gestures into the teaching process. This way, users can more easily teach machines about objects, and the machines can also learn more effectively.

You've probably heard the term machine learning before, but are you familiar with machine teaching? Machine learning is what happens behind the scenes when a computer uses input data to form models that can later be used to perform useful functions. But machine teaching is the somewhat less explored part of the process, which deals with how the computer gets its input data to begin with.

In the case of visual systems, for example ones that can recognize objects, people need to show objects to a computer so it can learn about them. But there are drawbacks to the ways this is typically done that researchers from the University of Tokyo's Interactive Intelligent Systems Laboratory sought to improve.

"In a typical object training scenario, people can hold an object up to a camera and move it around so a computer can analyze it from all angles to build up a model," said graduate student Zhongyi Zhou.

"However, machines lack our evolved ability to isolate objects from their environments, so the models they make can inadvertently include unnecessary information from the backgrounds of the training images. This often means users must spend time refining the generated models, which can be a rather technical and time-consuming task. We thought there must be a better way of doing this that's better for both users and computers, and with our new system, LookHere, I believe we have found

it."

Zhou, working with Associate Professor Koji Yatani, created LookHere to address two fundamental problems in machine teaching: first, the problem of teaching efficiency, aiming to minimize the users' time, and required technical knowledge. And second, of learning efficiency—how to ensure better learning data for machines to create models from.

LookHere achieves these by doing something novel and surprisingly intuitive. It incorporates the hand gestures of users into the way an image is processed before the machine incorporates it into its model, known as HuTics. For example, a user can point to or present an object to the camera in a way that emphasizes its significance compared to the other elements in the scene. This is exactly how people might show objects to each other. And by eliminating extraneous details, thanks to the added emphasis to what's actually important in the image, the [computer](#) gains better input data for its models.

"The idea is quite straightforward, but the implementation was very challenging," said Zhou. "Everyone is different and there is no standard set of hand gestures. So, we first collected 2,040 example videos of 170 people presenting objects to the camera into HuTics. These assets were annotated to mark what was part of the object and what parts of the image were just the person's hands.

"LookHere was trained with HuTics, and when compared to other object recognition approaches, can better determine what parts of an incoming image should be used to build its models. To make sure it's as accessible as possible, users can use their smartphones to work with LookHere and the actual processing is done on remote servers. We also released our [source code](#) and data set so that others can build upon it if they wish."

Factoring in the reduced demand on [users](#)' time that LookHere affords

people, Zhou and Yatani found that it can build models up to 14 times faster than some existing systems. At present, LookHere deals with teaching machines about [physical objects](#) and it uses exclusively visual data for input. But in theory, the concept can be expanded to use other kinds of input data such as sound or scientific data. And models made from that data would benefit from similar improvements in accuracy, too.

The research was published as part of *The 35th Annual ACM Symposium on User Interface Software and Technology*.