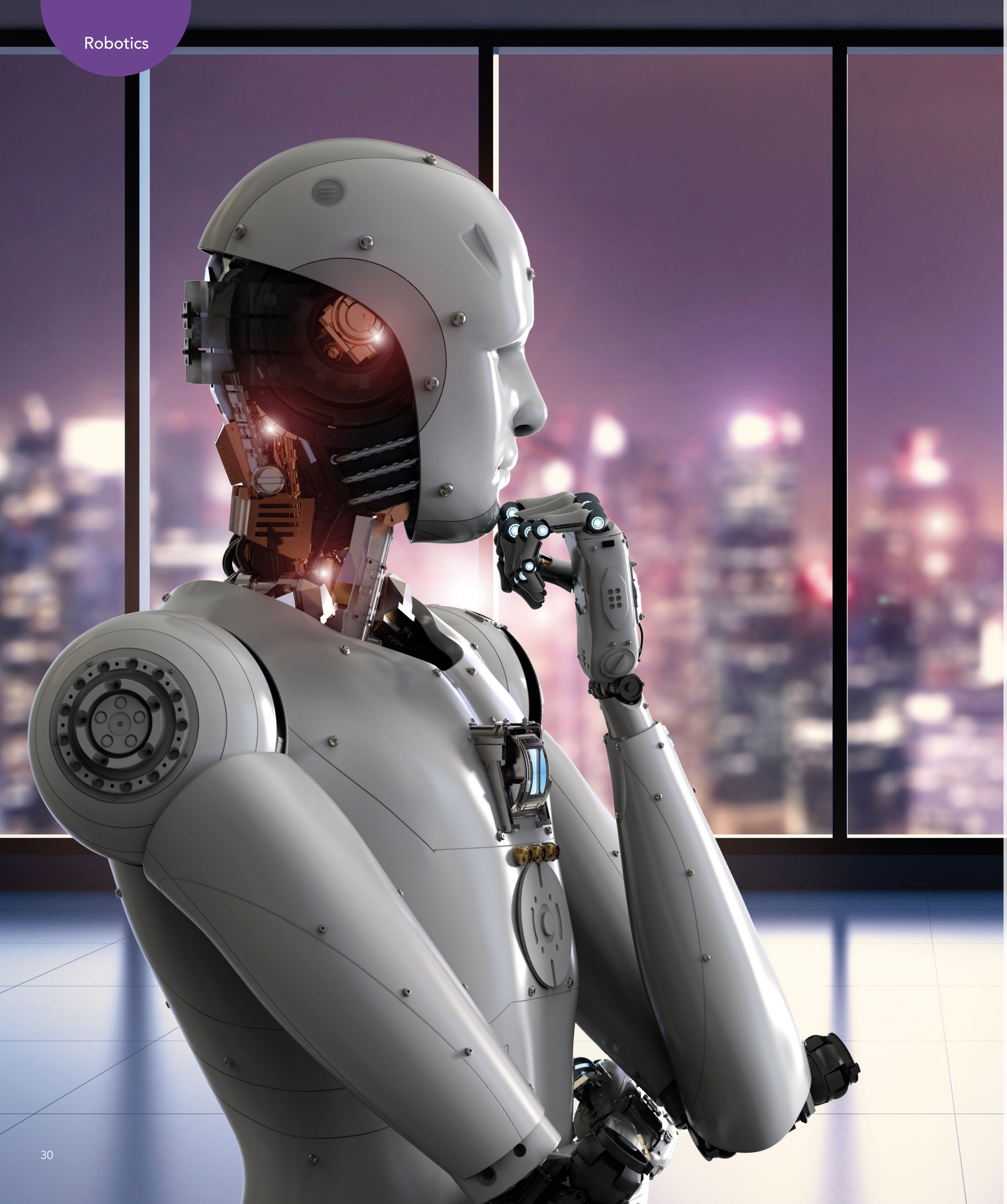


Understanding the world as a robot

While it is simple enough for a computer to capture and digitise an event through video recording technology, interpreting and using that information is much more challenging. Finding ways of translating to machine code the human ability to see and understand visual information is the domain of **Dr Cornelia Fermüller's** research at the University of Maryland at College Park. As part of this work, she has created systems that allow robots to interpret and predict human actions, an important step in developing robots capable of collaborating with people.

Advances in robotics technology have made it possible to create robots that are not confined to the lab, but can adventure across real-world terrain. From forests, to rivers, to nuclear plants, modern robots are capable of traversing regions inaccessible to humans to aid in important tasks such as carrying out rescue operations.

However, in order to take full advantage of the possibilities offered by more mobile robots, the robot needs to be able to interact and interpret its environment. For robots such as drones, all piloting and control is done via a remote human operator. On-board sensors and cameras will report information live so that the human operator then issues the necessary series of commands. The disadvantages of this are numerous. A reliable connection between the remote user and robot is essential and the data transfer also needs to be sufficiently quick to avoid introducing any unnecessary time-delays into the command-transfer and reporting process.

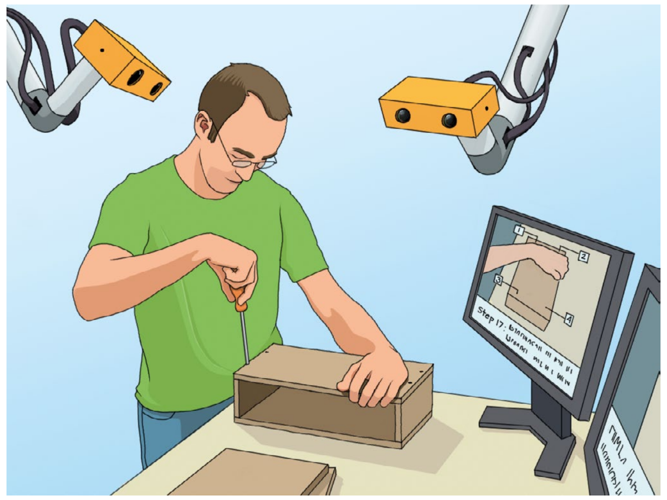
With advances in artificial intelligence, it is now possible to create machines that have some level of autonomous decision making. The question is, how can the complexity of the real world be translated into a form that can be understood by machines. Dr Cornelia Fermüller and her team at University of Maryland at College Park are experts in the development of such cognitive systems for machines. These are essentially a translation kit that converts external stimuli into something a robot can process and act upon. Through development of these systems it is possible to create robots that can not only recognise and distinguish different kinds of tools, but know which type of actions are required to interact with and utilise the tool to achieve particular tasks.

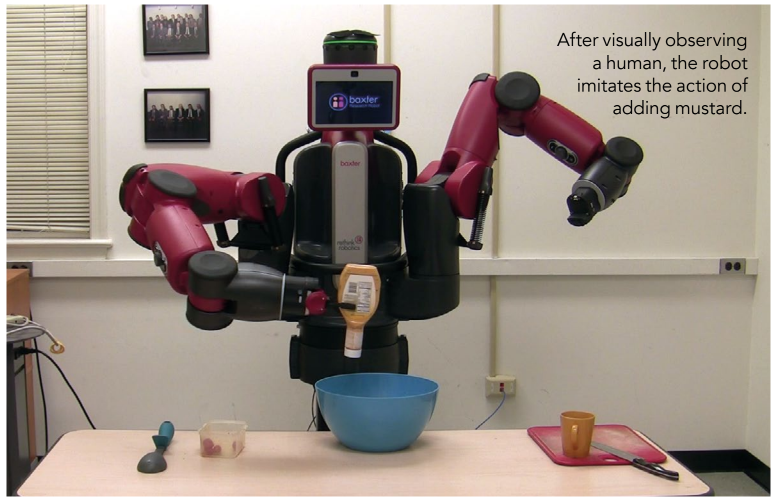
LEARNING TO LEARN

For humans, image and pattern recognition is a relatively trivial problem. Although the underlying chemical and biological processes for vision are complex, we are capable of processing and responding to visual information efficiently. Cornelia Fermüller has studied how humans perceive and reason about complex events and applied these insights to robots.

In English, a sentence usually consists of a subject, object and a verb. If we can identify each of these components, we can work out who (the subject) carried out which specific action (the verb) on what type of item (the object). How can this construction be applied to visual data though?

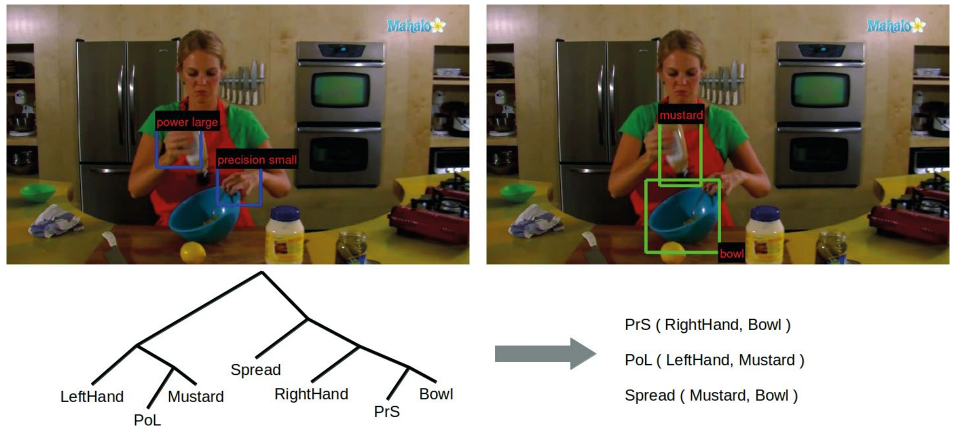
Mathematics and computations can tell us what is possible, but we need inspiration from nature on what we should study





After visually observing a human, the robot imitates the action of adding mustard.

Above: Illustration of the cyberphysical cognitive manufacturing assistant. The system monitors humans performing assembly tasks, and in the case of potential error, it will give suggestions to the assembler via displays.



Above: The team processed internet cooking videos to automatically obtain descriptions of the featured actions.

If we enter a room, and see a cat sitting on a mat and want to translate this to a sentence, we could view this scene through the 'subject, object, action' framework. The object is the cat, the action or verb is sitting, and the mat is the object. However, for us to be able to do this, we need to know several pieces of information: What do a cat and a mat look like? What movement pattern can be interpreted as sitting?

Dr Fermüller has been using this idea of describing events to build algorithms that can form the basis of a reasoning module for the machine. Currently, her work is focused on the context of understanding human manipulation of objects. By being able to file visual observations into categories, i.e. the subject or object, and recognise certain human hand motions, the machine can begin to interpret the action, and the rules for a more complex action response can be developed for the machine to behave autonomously.

SYMMETRY SHORTCUTS

The issue of visual-object recognition is a difficult one. One possibility to get a machine to recognise a cat is to train it on image sets of thousands of cats. However, the machine needs help to able to differentiate between the cat and the image background and if, for example, the machine has only been trained on images of black cats, it may not recognise a cat of a different colour as still being a cat. If the machine is shown a photograph of a cat from a different angle or partially obscured, it may also fail to recognise the object too.

Humans are very good at using object symmetry to speed up image processing times and to identify objects that are partially obscured. We are good at 'filling in the blanks' and interpolating images, which is why our brains can be easily fooled by optical illusions. Dr Fermüller has been applying this principle so that the machine too can make use of the symmetry properties of objects by using a fitting algorithm to reconstruct the missing parts of the object and identify items. This works particularly well if the robot needs to identify manmade objects, as they typically have high levels of symmetry.

THE ROBOT COLLABORATOR

One of Dr Fermüller's current research aims is to design a full cyber-physical system that is capable of helping humans with assembly tasks. This involves bringing together both the object recognition as part of her work in computer vision, alongside her developments in cognitive systems. Her work is heavily inspired by biological processes: she says, 'mathematics and computations can tell us what is possible, but we need inspiration from nature on what we should study'.

The intended device will be capable of recognising and predicting human actions in complex, noisy and cluttered environments and, by watching humans perform assembly tasks, identify mistakes and problems. Such a sophisticated, communicative device will also open the doors for such devices in other areas of robotics, such as search and rescue.

> **Cognitive systems are a type of translation kit that converts external stimuli into something a robot can process and act upon**

Q&A

What kinds of objects do computers typically struggle to recognise?

Computer Vision has made amazing progress in recognising objects in images. (The winners of a most recent academic challenge have shown 98% accuracy in recognising 1000 object categories). The approach involves training deep neural networks on large amounts of images. Currently, these approaches still struggle with cluttered scenes (many objects in the scene), and when objects are occluded. That is, when one object is partially behind another. However, in applications of Robotics, just recognising an object from an image is usually not sufficient. We also need to know about the object's geometry. We need to know how far away an object is, and its shape, so the robot can interact with it. Current approaches are based on detecting a combination of image and shape features on specific objects and can deal with about 100 specific objects. They work for textured objects, but do not work well when an object's surface does not have texture. Our current research seeks solutions by integrating image content with shape information and reasoning about objects' attributes to generalise object recognition to larger classes of objects and previously unseen examples.

How quickly can robots generally interpret complex visual information?

Using Computer Vision approaches based on machine learning, robots can interpret visual information very fast. If we use special hardware, called GPUs, fast means nearly real-time (at the speed images are recorded). However, at this point these approaches are not sufficiently reliable in complex scenarios for Robotics applications. While it is fine to detect an object or action in an image with 90 or even 98 percent accuracy if the goal is to summarise the image content, this is not sufficient for Robotics, where the robot has to act upon the visual information.

Approaches for object recognition are much more developed than approaches for recognising movements (actions). Current research focuses on generalisation mechanisms that combine visual learning with reasoning to address interpretation of complex actions and activities.

How accurately can robots predict human actions from small gestures?

We can recognise hand gestures very well, if the situation allows the use of specific hardware. For example, there has been great progress in recognising and tracking accurately hand poses using the Kinect sensors (e.g. hand-tracking software from Microsoft), or with a specialised hardware called Leap Motion, which uses infrared and works at a distance of one metre. These tools are well suited for applications of human computer interaction and virtual reality. However, if the robotic application is such that these sensors cannot be positioned so that they will always see the hand, solving the problem is still difficult. Furthermore, existing tools are designed for gestures when the hand does not touch objects. Recognition of hand gestures in object interactions currently does not work well.

What are some of the particular challenges for designing a cyber-physical system for assembly tasks?

When hands perform manipulations, they often cover the object in the observed images (video). As a result it is very difficult to create models of the hands and objects in order to make physical simulations of the fine motor actions. One creative approach we are using to address this issue, is to model not only the visual observations but also the forces (by recording data and using machine learning) and then use the sensory motor space for recognition.

The goal of this project is not only to demonstrate that we can interpret specific assembly actions, but to create a toolbox of software components for many assembly actions. When asked to monitor a new manipulation action, we will quickly combine our components and create a new system. We can deal with the components related to static quantities, i.e. the objects and tools, although there are still challenges with very small objects and occlusions. However, the creation of components related to the actions, the movements, grasps and force manipulations is still challenging, because of their complexity and the large variation with which people perform actions.

Detail

RESEARCH OBJECTIVES

Dr Fermüller works on Computational Vision. She aims to understand the process of how we interpret images, specifically of humans, to understand what they are doing. She uses this knowledge to develop programs that allow robots to process images and even predict behaviour.

FUNDING

NSF

COLLABORATORS

Prof Yiannis Aloimonos (UMD)
Prof John Baras (UMD)
Yezhou Yang (assistant professor at ASU)

BIO

Cornelia Fermüller received a PhD degree in Applied Mathematics from the Vienna University of Technology, Austria. She is a research scientist at UMIACS, University of Maryland, and her research is in the areas of Computer Vision, Human Vision, and Robotics. She develops biologically inspired computational solutions to problems of vision with the focus on problems of motion analysis for navigation and action recognition.

CONTACT

Dr Cornelia Fermüller
Associate Research Scientist
Computer Vision Laboratory
Center for Automation Research,
Institute for Advanced Computer Studies
University of Maryland at College Park
A.V. Williams Bldg., room 4459,
College Park, MD 20742, USA

E: fer@umiacs.umd.edu
T: +1 301 405 1768
W: http://www.cfar.umd.edu/~fer/
Robohub: http://robohub.org/teaching-a-robot-to-cook-by-showing-it-youtube-videos-of-cooking-shows/
Maryland Day 2017: http://www.wusa9.com/news/local/college-park/75000-expected-at-maryland-day-at-umd/435160895
Youtube channel: Videos illustrating the Vision Processes: (https://www.youtube.com/watch?v=Kf68Y-dwZxw)
How the robot learns from a human how to mix a specific drink: https://www.youtube.com/watch?v=pD8a4W9Y3Jg