Multi-modal interaction – How machines learn to understand pointing

December 2, 2019

As we learn more about the biological world around us, the list of things only humans can do has dwindled – and that's before computers started to play chess and Go. Counting? Birds can deal with Using tools? Dolphins in Shark Bay, Australia, are using sponges as a tool for hunting. Against this background, it may come as a surprise how specifically human pointing is: although it seems very natural and easy to us, not even chimpanzees, our closest living relatives, can muster more than the most trivial forms of pointing. So how could we expect machines to understand it?

Three forms of pointing
In 1934, the linguist and psychologist Karl Bühler distinguished three forms of pointing, all connected to language: The first is pointing "ad oculos," that is in the field of visibility centered around the speaker ("here") and also accessible to the listener. While we can point within this field with our fingers alone, languages offer a special set of pointing words to complement this ("here" vs. "there;" "this" vs. "that;" "left" and "right;" "before" and "behind" etc.). The second form of pointing operates in a remembered or imagined world, brought about by language ("When you leave the Metropolitan Museum, Central Park is behind you and the Guggenheim Museum is to your left. We will meet in front of that"). The third form is pointing within language: As speech is embedded in time, we often have the necessity to point back to something we said a little earlier or point forward to something we will say later. Such anaphoric use of pointing words ("How is the weather in Tokyo?" "Nice and sunny." "Are there any good hotels there?") is now a standard feature in many smart assistants (and distinguishes them from the not-so-smart). And the first mode of pointing at elements in the visible vicinity is now also available in today's smart assistants.

First automotive assistants to support "pointing"
At CES 2018 in Las Vegas , we demonstrated for the first time how drivers can point to buildings outside the car and ask questions like, "What are the opening hours of that shop?" and we have worked on this vision ever since. Watch out for it to appear in cars on the road very soon. And, the "pointing" doesn't need to be done with a finger. Exploiting the fact that Driver Monitoring Cameras are featured in more and more cars due to safety regulations, you can simply look at the object in question, something made possible by eye gaze detection available as a feature in these cameras. This technology is imitating human behavior, as humans are very good at guessing where somebody is looking just by observing his or her eyes.

Biologists suggest that the distinct shape and appearance of the human eye (a dark iris and a contrasting white surrounding) is no accident, but a product of evolution facilitating the ability of gaze detection. Artists have exploited that for many centuries: with just a few brush strokes of paint, they can make figures in their paintings look at other figures or even outside the picture – including at the viewer of the painting. Have a look at Raffael's Sistine Madonna, which is displayed in Dresden, and see how the figures' viewing directions make them point at each other and how that guides our view.

Multi-modal interaction: When speech, gesture, and handwriting work hand in hand
Machines can also do this based on image recognition and Deep Learning, capabilities which, originally coming out of our cooperation with DFKI, will bring us into the age of truly multi-modal assistants. It is important to remember that "multi-modal" does not just mean you have a choice between modalities (typing OR speaking OR handwriting on a pad to enter the destination into your navigation system), but that multiple modalities work together to accomplish one task. For example, when pointing to something in the vicinity (modality 1) and saying, "tell me more about this" (modality 2), both modalities are needed to explain what the person performing this wants to accomplish.

Multi-modal interaction – a key feature for Level 4 and 5 autonomous vehicles?
While it is obvious why such a capability is attractive to today's drivers, there are hints that it might become even more important as we enter the age of autonomous vehicles. Many people are wondering what drivers will do when they don't have to drive any more, something they would experience in Levels 4 and 5 of the autonomous driving scale. Some studies indicate that perhaps the answer is not that much, actually. For example, a 2016 German study asked people about the specific advantages they perceived in such vehicles, and "... that I can enjoy the landscape" came out as the top choice at all levels of autonomy. It's not too difficult to imagine a future with gaze and gesture detection, combined with a "just talk" mode of speech recognition – one where you can ask "what is that building?" without having to press a button or say a keyword first. And where we also have true multi-modal output combining screens, smart glass and voice. This future will give users of autonomous vehicles exactly what they want. And for today's users of truly multi-modal systems, machines just got a little more human-like again.