Video recognition: Spot the difference?

By Markus Bernards

Computers can already recognise objects and faces quite well, and also that something is moving and in which direction. However, artificial intelligence still has difficulties in spotting the *type* of movement involved. This is what computers are now learning in Professor Hilde Kühne's laboratory at Goethe University.

The computer learns from thousands of cooking videos which movement is "cutting".

Photo: Uwe Dettmar

Following her 80th birthday, Lydia S., a lady living on her own, registered with a home emergency call service run by Caritas. They gave her a button on a lanyard. Pressing it would set off an alarm at the service's call centre, and a member of staff would call to ask if everything was all right. When she was at home, she hung the button round her neck and found it reassuring to be able to summon help, for example if she were to fall. When Lydia S. then suffered a mild stroke, fell and urgently needed the emergency service, the button was lying on the armchair where she had put it when watching television. She was lying on the floor, and it was out of her reach. It was a while before she managed to get up and drag herself to the telephone to call for help.

The disadvantage of emergency call systems for elderly people worn around the neck or on a wristband is that people can put them down and they are then out of reach at the crucial moment. However, video cameras in the home, for example, are hardly an acceptable alternative. "Nobody wants an emergency service monitoring their home via video," says Hilde Kühne, assistant professor for image recognition systems and machine learning at Goethe University. "And certainly not in the bedroom or the bathroom, although it's usually here that people quickly find themselves in a critical situation."

Protection of personal privacy

Perhaps it would be a different matter if video cameras were indeed mounted in the home, but only a computer could see the pictures and not a human? If the computer were to alert the emergency service in the event of a fall, but without transmitting the actual video data? This would guarantee privacy because "the computer isn't interested in the person moving around in the apartment," says Kühne, whose current research topic is automatic motion recognition. "For the computer, videos are simply columns of figures."

To report a fall, the computer would, however, first have to be capable of distinguishing it from other types of movement. However, this is more difficult than identifying faces and objects in photos since computer training with videos is more complex – simply because of the vast amounts of data to be processed. For a movement to become visible, 50 to 100 frames are needed – so 50 to 100 times the data volume of a photograph.

ABOUT HILDE KÜHNE

Professor Hilde Kühne, born in 1979, studied computer science (computational visualistics) at the University of Koblenz-Landau and earned her doctoral degree at Karlsruhe Institute of Technology. In the course of her academic career, she worked at the Fraunhofer Institute for Communication, Information Processing and Ergonomics (FKIE) and the University of Bonn before moving to the MIT-IBM Watson AI Lab of the Massachusetts Institute of Technology in the USA as a researcher. Since 2021, she has also been assistant professor for image recognition systems and machine learning at Goethe University. She is co-founder of ks-research and was recently awarded the ICCV Helmholtz Prize in acknowledgement of a paper which, ten years after its publication, has proven significant for research.

kuehne@em.uni-frankfurt.de

Many different words for the same movement

Moreover, computers are usually trained with texts that describe what can be seen in photos or video sequences. Such keywords, known as annotations, are formulated by humans who look at the pictures and describe them. In this way, the computer learns, for example, what a cup is when it sees lots of pictures with the annotation "cup". In the case of videos, compiling annotations and obtaining sufficient training material in the process is much more time-consuming by virtue of the large amounts of data and the greater length of time that needs to be calculated for videos.

There are two further problems on top: firstly, there are often different words for the same movement, which also depend on how long a movement is observed. Kühne: "If I watch someone for just three seconds, I can say, for example, 'he's running' or 'he's walking'. If I watch him for 20 seconds, I know 'he's sprinting' or 'he's jogging'. If I watch even more of the video and a dog appears, or I see a bus stop, I recognise that 'he's running away from the dog' or 'he's rushing to the bus stop'. That makes the task of recognising movement difficult to define, for humans as well as computers."

Solution: autonomous learning

The second problem lies in how humans process the flow of data received via their eyes and ears. We do not perceive movements as something continuous, but instead divide them into smaller segments in order to remember them. These segments are then pieced together again in the brain to form a continuous motion sequence. How many separate segments are perceived depends on the individual experiences and abilities of each observer. Hilde Kühne gives figure skating at the Olympic Games as an example: "The judges are trained and able to analyse the motion sequence in figure skating very precisely. A layperson sees the same sequence but can hardly distinguish between the individual elements."

So how is the computer supposed to learn? In Hilde Kühne's opinion: autonomously, by itself and no longer on the basis of annotations. Here, Kühne's team draws on a pool of 100 million YouTube videos. To learn, the computer is equipped with an artificial neural network. These are algorithms which in principle function like nerve cells in a brain. "But they are actually mathematical functions that convert columns of figures into other columns of figures," says Kühne.

Computer training

The computer is fed three pieces of information from each video clip: the actual video sequence showing a movement, the soundtrack and any subtitles that are perhaps superimposed on the video. An example would be a sequence from a cooking video in which the YouTube chef dices a pepper and says: "Now we're going to dice the pepper." "Dice the pepper" appears in the subtitles at the same time.

For the computer, the information – video, sound and subtitles – is three columns of figures, from which it calculates, with the help of a

IN A NUTSHELL

- Computers in the lab learn from 100 million YouTube videos how to recognise movements.
- As text descriptions for motion videos are relatively complex, the computers train by themselves. The goal: to use algorithms to link together the video images, sound and subtitles of a movement.
- Potential applications are found in assisted living or the recognition of dangerous situations in video surveillance.

mathematical function, three points in an "embedding space". This can be envisaged as a large, transparent cube. Kühne explains: "We want to find a mathematical function which translates the three columns of figures for the 'cutting' movement in such a way that they form three points close to each other in the embedding space. Video, sound and subtitle data of another movement, such as 'waving', should in turn generate three points at a different spot in the embedding space."

Training the computer consists now of analysing lots of videos and generating respective groups of points in the embedding space for different movements. In the next step, the computer scientists show the computer annotated videos so that it can link the groups of points to the corresponding words, such as "cutting" or "waving", and now "knows" what the respective movements are called.

Many applications

At some point in time, the computer should then be able to recognise a wide variety of movements, even if they are part of a longer video with lots of scenes, and assign the same movements to "cut", even if the person in the video says "chop", "slice", "debone", "dice" or "trim" instead of "cut". And it will also be able to distinguish *what* is being cut: vegetables, the garden hedge or a video.

And if it can differentiate between "fall" and "kneel", "bend down" or "sit down", it is perhaps ready to be a discreet helper in emergency care. Other applications could be autonomous driving, where it can contribute to preventing accidents, or science, where it can assist in the evaluation of behavioural studies.

So far, so good: a positive outlook for the future. But won't this technology also lead to us being kept under surveillance even more than we already are? "Surveillance in itself is not a bad thing," says Hilde Kühne. "In my opinion, surveillance is first of all neutral. Of course, it can be misused, like almost any technology, and we ought to keep a close eye on that. But motion recognition is precisely what could help protect personal privacy, for example in assisted living. If you want to spot a dangerous situation in an underground station, such as a fight, the computer can help prioritise certain activities over, for example, the picture of a platform with children playing catch. After all, nobody can watch all the surveillance videos the whole time. The idea is therefore not that the computer will become master of the Universe, but that by being able to process and filter vast amounts of data automatically, we will make decisions *easier* for people and in this way also enable them to make *better* decisions." ●

The author

Markus Bernards, born in 1968, holds a doctoral degree in molecular biology. He is a science journalist and editor of Forschung Frankfurt.

bernards@em.uni-frankfurt.de